
Estilometria

Autor:

Data de publicació: 15-09-2023

L'estilometria és l'aplicació de l'estudi de l'estil lingüístic, generalment al llenguatge escrit. [1] També s'ha aplicat amb èxit a la música,[2] pintures,[3] i escacs. [4] Recentment, l'enginyeria de característiques s'utilitza per capturar automàticament el que és específic de l'estil d'un autor, basat en models d'espai vectorial. [5] [6]

L'estilometria s'utilitza sovint per atribuir l'autoria a documents anònims o en disputa. [7] Té aplicacions legals, acadèmiques i literàries, que van des de la qüestió de l'autoria de les obres de Shakespeare fins a la lingüística forense i té similituds metodològiques amb l'anàlisi de la llegibilitat del text.

Part d'una sèrie sobre

Lingüística

Esquema
de l'història
Índex

mostrar
Lingüística general

mostrar
Lingüística aplicada

mostrar
Marcs teòrics

mostrar
Temes

Portal

v
t
e

L'estilometria és l'aplicació de l'estudi de l'estil lingüístic, generalment al llenguatge escrit. [1] També s'ha aplicat amb èxit a la música,[2] pintures,[3] i escacs. [4] Recentment, l'enginyeria de característiques s'utilitza per capturar automàticament el que és específic de l'estil d'un autor, basat en models d'espai vectorial. [5] [6]

L'estilometria s'utilitza sovint per atribuir l'autoria a documents anònims o en disputa. [7] Té aplicacions legals, acadèmiques i literàries, que van des de la qüestió de l'autoria de les obres de Shakespeare fins a la lingüística forense i té similituds metodològiques amb l'anàlisi de la llegibilitat del text.

L'estilometria es pot utilitzar per desemmascarar autors pseudònims o anònims, o per revelar alguna informació sobre l'autor sense una identificació completa. Els autors poden utilitzar l'estilometria adversària per resistir-se a aquesta identificació eliminant les seves pròpies característiques estilístiques sense canviar el contingut significatiu de les seves comunicacions. Pot derrotar anàlisis que no tenen en compte la seva possibilitat, però l'eficàcia última de l'estilometria en un entorn adversari és incerta: la identificació estilomètrica pot no ser fiable, però tampoc es pot garantir la no identificació; La pròpia pràctica de l'estilometria adversària pot ser detectable.

Contingut

Història

L'estilometria va sorgir de tècniques anteriors d'anàlisi de textos per obtenir proves d'autenticitat, identitat d'autor i altres qüestions.

La pràctica moderna de la disciplina va rebre publicitat de l'estudi dels problemes d'autoria en el drama renaixentista

anglès. Els investigadors i els lectors van observar que alguns dramaturgs de l'època tenien patrons distintius de preferències lingüístiques, i van intentar utilitzar aquests patrons per identificar autors d'obres incertes o col·laboratives. Els primers esforços no sempre van tenir èxit: el 1901, un investigador va intentar utilitzar la preferència de John Fletcher per "em", la forma contractual, com a marcador per distingir entre Fletcher i Philip Massinger en les seves col·laboracions, però va emprar erròniament una edició de les obres de Massinger en la qual l'editor havia ampliat totes les instàncies de "em" a "ells". [8]

Els fonaments de l'estilometria van ser establerts pel filòsof polonès Wincenty Lutosławski a *Principes de stylométrie* (1890). Lutosławski va utilitzar aquest mètode per desenvolupar una cronologia dels Diàlegs de Plató. [9]

El desenvolupament dels ordinadors i les seves capacitats per analitzar grans quantitats de dades van potenciar aquest tipus d'esforços per ordres de magnitud. La gran capacitat dels ordinadors per a l'anàlisi de dades, però, no garantia una sortida de bona qualitat. Durant la dècada de 1960, el reverend A. Q. Morton va produir una anàlisi informàtica de les catorze Epístoles del Nou Testament atribuïdes a Sant Pau, que indicava que sis autors diferents havien escrit aquest cos d'obra. Una comprovació del seu mètode, aplicat a les obres de James Joyce, va donar com a resultat que *Ulysses*, la novel·la multiperspectiva i multiestil de Joyce, estava composta per cinc individus separats, cap dels quals aparentment va tenir cap paper en l'elaboració de la primera novel·la de Joyce, *Retrat de l'artista de jove*. [10]

Amb el temps, però, i amb la pràctica, investigadors i estudiosos han refinat els seus mètodes, per obtenir millors resultats. Un èxit primerenc notable va ser la resolució de l'autoria discutida de dotze dels *Federalist Papers* de Frederick Mosteller i David Wallace. [11] Tot i que encara hi ha preguntes sobre els supòsits i mètodes inicials (i, potser, sempre ho serà), pocs discuteixen ara la premissa bàsica que l'anàlisi lingüística dels textos escrits pot produir informació i una visió valuoses. (De fet, això era evident fins i tot abans de l'adveniment dels ordinadors: l'aplicació reeixida d'una anàlisi textual/lingüística al cànion de Fletcher per part de Cyrus Hoy i altres va donar resultats clars durant la dècada de 1950 i principis de la de 1960.)

Aplicacions

Les aplicacions de l'estilometria inclouen estudis literaris, estudis històrics, estudis socials, recuperació d'informació i molts casos i estudis forenses. [12][13] També es pot aplicar a codi informàtic [14] i detecció de plagi intrínsec, que és per detectar plagi en funció dels canvis d'estil d'escriptura dins del document. [15] L'estilometria també es pot utilitzar per predir si algú és parlant nadiu o no nadiu d'anglès per la seva velocitat d'escriptura. [16]

L'estilometria com a mètode és vulnerable a la distorsió del text durant la revisió. [17] També hi ha el cas que l'autor adoptés diferents estils al llarg de la seva carrera, com es va demostrar en el cas de Plató, que va optar per diferents polítiques estilístiques com les adoptades per als diàlegs inicial i mitjà que abordaven el problema socràtic. [18]

Trets

Els trets textuais d'interès per a l'atribució de l'autoria són, d'una banda, computar ocurrencies d'expressions o construccions idiosincràtiques (per exemple, comprovar com l'autor utilitza la interpuntuació o amb quina freqüència l'autor utilitza construccions passives sense agents) i, de l'altra, similars a les utilitzades per a l'anàlisi de llegibilitat, com ara mesures de variació lèxica i variació sintàctica. [19] Atès que els autors sovint tenen preferències per a certs temes, els experiments de recerca en l'atribució de l'autoria eliminen principalment paraules de contingut com substantius, adjectius i verbs del conjunt de característiques, conservant només elements estructurals del text per evitar ajustar excessivament els seus models al tema en lloc de les característiques de l'autor. [20] [21] Les característiques estilístiques sovint es calculen com a mitjanes sobre un text o sobre tota l'obra completa d'un autor, donant mesures com la longitud mitjana de les paraules o la longitud mitjana de la frase. Això permet identificar autors que tenen una preferència clara per les frases verbals o terses però que amaguen variacions: un autor amb una barreja de frases llargues i curtes tindrà la mateixa mitjana que un autor amb frases mitjanes consistents. Per capturar aquesta variació, alguns experiments utilitzen seqüències o patrons sobre observacions en lloc de freqüències observades mitjanes, assenyalant, per exemple, que un autor mostra una preferència per un determinat patró d'estrès o èmfasi, [22] [23] o que un autor tendeix a seguir una seqüència de frases llargues amb una de curta. [24] [25]

Una de les primeres aproximacions a la identificació de l'autoria, per Mendenhall, es pot dir que agrega les seves observacions sense fer-ne una mitjana. [26]

Els models d'atribució d'autoria més recents utilitzen models d'espai vectorial per capturar automàticament el que és específic de l'estil d'un autor, però també es basen en l'enginyeria de característiques assenyades per les mateixes raons que els models més tradicionals. [27] [28]

Estilometria adversària

L'estilometria adversària és la pràctica d'alterar l'estil d'escriptura per reduir el potencial de l'estilometria per descobrir la identitat de l'autor o les seves característiques. [29] Aquesta tasca també es coneix com a ofuscació de l'autoria o anonimització de l'autoria. L'estilometria planteja un desafiament significatiu a la privacitat en la seva capacitat per desemmascarar autors anònims o vincular pseudònims amb altres identitats d'un autor,[30] la qual cosa, per exemple, crea dificultats per als denunciants,[31] activistes,[32] i els enganyadors i estafadors. [33] S'espera que el risc per a la privadesa creixi a mesura que es desenvolupin tècniques d'aprenentatge automàtic i corpus de text. [34]

Tota estilometria adversària comparteix la idea central de parafrasejar fidelment el text original de manera que el significat no canviï però els senyals estilístics quedin enfosquits. [35][36] Una parafrasi tan fidel és un exemple adversari per a un classificador estilomètric. [37] Hi ha diversos enfocaments amplis sobre això, amb algunes superposicions: la imitació, substituint l'estil propi de l'autor pel d'un altre; traducció, aplicant la traducció automàtica amb l'esperança que elimini l'estil característic del text original; i ofuscació, modificar deliberadament l'estil d'un text perquè no s'assembli al propi de l'autor. [35]

L'estil ocult manualment és possible, però laboriós; [38] En algunes circumstàncies, és preferible o necessari. [39] Les eines automatitzades, ja siguin semi o totalment automàtiques, podrien ajudar un autor. [38] La millor manera de dur a terme la tasca i el disseny d'aquestes eines és una pregunta de recerca oberta. [40][34] Tot i que s'ha demostrat que alguns enfocaments són capaços de derrotar anàlisis estilomètriques particulars,[41] particularment aquells que no tenen en compte el potencial de l'adversarialitat,[42] establir seguretat davant anàlisis desconegudes és un problema. [43] Garantir la fidelitat de la parafrasi és un repte crític per a les eines automatitzades. [34]

És incert si la pràctica de l'estilometria adversària és detectable en si mateixa. Alguns estudis han trobat que determinats mètodes produeixen senyals en el text de sortida, però un estilometrista que no està segur de quins mètodes es poden haver utilitzat pot no ser capaç de detectar-los de manera fiable. [34]

Recerca actual

L'estilometria moderna utilitza ordinadors per a l'anàlisi estadística, i la intel·ligència artificial i l'accés al creixent corpus de textos disponibles a través d'Internet. [44] Sistemes de programari com Signature[45] (programari gratuït produït pel Dr. Peter Millican de la Universitat d'Oxford), JGAAP[46] (Java Graphical Authorship Attribution Program—programari gratuït produït pel Dr. Patrick Juola de la Universitat de Duquesne), stylo[47][48] (un paquet de codi obert R per a una varietat d'anàlisis estilomètriques, inclosa l'atribució d'autoria, desenvolupat per Maciej Eder, Jan Rybicki i Mike Kestemont) i Stylenet[49] per al neerlandès (programari gratuït en línia del professor Walter Daelemans de la Universitat d'Anvers i el doctor Véronique Hoste de la Universitat de Gant) fan que el seu ús sigui cada vegada més factible, fins i tot per als no experts.

Espais acadèmics i esdeveniments

Els mètodes estilomètrics s'utilitzen per a diversos temes acadèmics, com a aplicació de la lingüística, la lexicografia o l'estudi literari,[1] juntament amb el processament del llenguatge natural i l'aprenentatge automàtic, i s'apliquen a la detecció de plagi, l'anàlisi de l'autoria o la recuperació d'informació. [44]

Lingüística forense

L'Associació Internacional de Lingüistes Forenses (IAFL) organitza la Conferència Biennal de l'Associació Internacional de Lingüistes Forenses (13a edició el 2016 a Porto) i publica The International Journal of Speech, Language and the Law amb l'estilística forense com un dels seus temes centrals.

AAAI

L'Associació per a l'Avanç de la Intel·ligència Artificial (AAAI) ha acollit diversos esdeveniments sobre anàlisi subjectiva i estilística del text. [50][51][52]

PAN

Tallers PAN (originalment, anàlisi de plagi, identificació d'autoria i detecció gairebé duplicada, més tard més generalment taller sobre descobriment de plagi, autoria i mal ús de programari social) organitzats des de 2007 principalment juntament amb conferències d'accés a la informació com ACM SIGIR, FIRE i CLEF. PAN formula tasques de desafiament compartit per a la detecció de plagi,[53] identificació d'autoria,[54] identificació de gènere de l'autor,[55] perfils d'autor,[56] detecció de vandalisme,[57] i altres tasques d'anàlisi de text relacionades, moltes de les quals depenen de l'estilometria.

Casos d'interès

El 1439, Lorenzo Valla va demostrar que la Donació de Constantí era una falsificació, un argument basat en part en una comparació del llatí amb l'utilitzat en documents autèntics del segle IV.

El 1952, el sacerdot suec Dick Helander va ser elegit bisbe de Strängnäs. La campanya va ser competitiva i Helander va ser acusat d'escriure una sèrie de cent cartes anònimes calumnioses sobre altres candidats a l'electorat del bisbat de Strängnäs. Helander va ser condemnat primer per escriure les cartes i va perdre la seva posició com a bisbe, però més tard va ser parcialment exonerat. Les lletres van ser estudiades utilitzant una sèrie de mesures estilomètriques (i també característiques de la màquina d'escriure) i els diversos casos judicials i exàmens posteriors, molts contrets pel mateix Helander durant els anys fins a la seva mort el 1978, van discutir el mètode estilomètric i el seu valor com a prova amb cert detall. [58][59]

El 1975, després que Ronald Reagan hagués servit com a governador de Califòrnia, va començar a donar comentaris setmanals de ràdio sindicats a centenars d'emissores. Després que les seves notes personals es fessin públiques en el seu 90è aniversari el 2001, un estudi va utilitzar mètodes estiloestadístics per determinar quines d'aquestes xerrades van ser escrites per ell i quines van ser escrites per diversos ajudants. [60]

El 1996, l'anàlisi estilomètrica del controvertit llibre d'autor pseudònim Primary Colors, realitzat pel professor del Vassar College Donald Foster[61] va cridar l'atenció del tema a un públic més ampli després d'identificar correctament l'autor com Joe Klein. (Aquest cas només es va resoldre després que una anàlisi manuscrita confirmés l'autoria.)

El 1996, es van utilitzar mètodes estilomètrics per comparar el manifest d'Unabomber amb cartes escrites per un dels sospitosos, Theodore Kaczynski, que van resultar en la captura i posterior condemna de Kaczynski. [62]

A l'abril de 2015, investigadors utilitzant tècniques d'estilometria van identificar una obra de teatre, Double Falsehood, com l'obra de William Shakespeare. [63][64] Els investigadors van analitzar 54 obres de Shakespeare i John Fletcher, i van comparar la longitud mitjana de la frase, van estudiar l'ús de paraules inusuals i van quantificar la complexitat i la valència psicològica del seu llenguatge.

El 2016, MacDonald P. Jackson, professor emèrit d'anglès a la Universitat d'Auckland, Nova Zelanda i membre de la Royal Society of New Zealand, que havia passat tota la seva carrera acadèmica analitzant l'atribució de l'autoria, va escriure un llibre titulat Who Write "The Night Before Christmas"?: Analyzing the Clement Clarke Moore vs. Henry Livingston Question.[65] en què avalua els arguments oposats i, per primera vegada, utilitza les tècniques d'atribució d'autor de l'estilística computacional moderna per examinar la controvèrsia de llarga durada. Jackson empra una sèrie de proves i n'introdueix una de nova, l'anàlisi estadística dels fonemes; arriba a la conclusió que Livingston és el veritable autor de l'obra clàssica.

El 2017, Simon Fuller i James O'Sullivan van publicar un estudi que afirmava que l'autor supervendes James Patterson no escriu en les seves novel·les aparentment coautors. [66][67][68] Segons O'Sullivan, la seva col·laboració amb l'expresident dels Estats Units Bill Clinton, The President is Missing, és una excepció a aquesta regla. [69]

El 2017, un grup de lingüistes, informàtics i estudiosos van analitzar l'autoria d'Elena Ferrante. A partir d'un corpus creat a la Universitat de Pàdua amb 150 novel·les escrites per 40 autors, van analitzar l'estil de Ferrante a partir de set de les seves novel·les. Van poder comparar el seu estil d'escriptura amb altres 39 novel·listes utilitzant, per exemple, estil. [47] La conclusió va ser la mateixa per a tots ells: Domenico Starnone és l'autor secret d'Elena Ferrante. [70]

El 2018, Mark Glickman, professor titular d'estadística a la Universitat de Harvard, va treballar amb Ryan Song, un antic estudiant d'estadística a Harvard, i Jason Brown, professor de la Universitat Dalhousie de Nova Escòcia, aplicant l'estilometria per descobrir que, molt probablement, la cançó "In My Life" de The Beatles va ser composta per John Lennon, però amb un 50% de probabilitats que Paul McCartney escrivís els vuit mitjans. [71][72]

El 2019, el projecte ETSO: Estilometria aplicada al teatre espanyol del Segle d'Or,[73] dirigit per Álvaro Cuéllar González i Germán Vega García-Luengos (Universitat de Valladolid) va aconseguir reunir 3000 obres del Segle d'Or espanyol. Després d'aplicar l'anàlisi estilomètrica, es va ratificar l'atribució de Mujeres y criados a Lope de Vega[74][75], i es va detectar un problema d'autoria a La monja alférez, obra atribuïda a Pérez de Montalbán que, gràcies a aquestes anàlisis i a través de la investigació històrica i filològica, finalment es va atribuir a Juan Ruiz de Alarcón. [76][77][78][79]

El 2023, el mateix projecte va trobar Lope de Vega com a autor de La francesa Laura, malgrat que el manuscrit va ser escrit anys després de la seva mort. [80] La comèdia va ser classificada com una obra tardana de Lope de Vega i datada de 1628 a 1630, ja que el seu tractament afalagador de França es podria atribuir a la bona relació momentània entre Espanya i França durant la Guerra dels Trenta Anys, tenint Anglaterra com un enemic comú. [81] En aquesta anàlisi, es comparen les 500 paraules més freqüents del text investigat amb les 500 de la resta d'obres. En el cas de La francesa Laura, la troballa va detectar que les 100 obres amb què estava més a prop eren gairebé totes de Lope de Vega. Els mètodes d'aprenentatge automàtic, com l'anàlisi automàtica de vectors de suport, també es van dur a terme amb una àmplia gamma de paràmetres. L'anàlisi filològica tradicional sobre l'autoria d'obres ha confirmat les investigacions d'estilometria i intel·ligència artificial. [82]

El 2020, Rachel McCarthy i James O'Sullivan van argumentar que Emily Brontë és la veritable autora de Cims borrascosos, posant fi a les especulacions d'alguns crítics que la novel·la podria haver estat escrita per un dels seus germans, concretament Branwell o Charlotte. [83]

El 2020, Hartmut Ilseman va utilitzar Rolling Delta i Rolling Classify de la suite de programes R Stylo per demostrar que el corpus de Marlowe és estilísticament inhomogeni i que l'autor de les dues Tamburlaines amb prou feines estava present en el corpus oficial restant de Marlowe. [84][85][86]

El 2022, els estudiosos italians Simone Rebora i Massimo Salgaro van demostrar, utilitzant el mètode de la "distància Delta" de John F. Burrows, que Felix Salten és l'autor més probable de la novel·la anònima Josefine Mutzenbacher del 1906, les pàgines finals excloses. [87]

Dades i mètodes

Atès que l'estilometria té tant casos d'ús descriptius, utilitzats per caracteritzar el contingut d'una col·lecció, com casos d'ús identificatius, per exemple, la identificació d'autors o categories de textos, els mètodes utilitzats per analitzar les dades i característiques anteriors van des dels construïts per classificar elements en conjunts o per distribuir elements en un espai de variació de característiques. La majoria dels mètodes són de naturalesa estadística, com l'anàlisi de clústers i l'anàlisi discriminant, normalment es basen en dades i característiques filològiques, i són dominis d'aplicació fructífers per als mètodes moderns d'aprenentatge automàtic.

Mentre que en el passat, l'estilometria emfatitzava els elements més rars o cridaners d'un text, les tècniques contemporànies poden aïllar patrons d'identificació fins i tot en parts comunes del discurs. La majoria de sistemes es basen en l'estadística lèxica, és a dir, utilitzen les freqüències de paraules i termes del text per caracteritzar el text (o el seu autor). En aquest context, a diferència de la recuperació d'informació, els patrons d'ocurrència observats de les paraules més comunes són més interessants que els termes tòpics, que són menys freqüents. [88][89]

El mètode estilomètric principal és l'invariant de l'escriptor: una propietat comuna per tots els textos, o almenys tots els textos prou llargs com per admetre anàlisis que donen resultats estadísticament significatius, escrits per un autor determinat. Un exemple d'invariant d'escriptor és la freqüència de les paraules de funció utilitzades per l'escriptor.

En un d'aquests mètodes, s'analitza el text per trobar les 50 paraules més comunes. A continuació, el text es divideix en 5.000 trossos de paraules i s'analitza cadascun dels trossos per trobar la freqüència d'aquestes 50 paraules en aquest tros. Això genera un identificador únic de 50 números per a cada tros. Aquests números situen cada tros de text en un punt d'un espai de 50 dimensions. Aquest espai de 50 dimensions s'aplana en un pla mitjançant l'anàlisi de components principals (PCA). Això es tradueix en una visualització de punts que corresponen a l'estil d'un autor. Si dues obres literàries se situen en el mateix pla, el patró resultant pot mostrar si ambdues obres són del mateix autor o d'autors diferents.

Estadístiques gaussianes

Les dades estilomètriques es distribueixen segons la llei de Zipf-Mandelbrot. La distribució és extremadament punxeguda i leptokurtica, la raó per la qual els investigadors no van poder utilitzar estadístiques per resoldre, per exemple, problemes d'atribució d'autoria. No obstant això, l'ús de l'estadística gaussiana és perfectament possible aplicant la transformació de dades. [90]

Xarxes neuronals

Les xarxes neuronals, un cas especial de mètodes estadístics d'aprenentatge automàtic, s'han utilitzat per analitzar l'autoria dels textos. Els textos d'autoria indiscutible s'utilitzen per entrenar una xarxa neuronal mitjançant processos com la retropropagació, de manera que es calcula l'error d'entrenament i s'utilitza per actualitzar el procés per augmentar la precisió. A través d'un procés semblant a la regressió no lineal, la xarxa guanya la capacitat de generalitzar la seva capacitat de reconeixement a nous textos als quals encara no ha estat exposada, classificant-los a un grau de confiança declarat. Aquestes tècniques es van aplicar a les reivindicacions de llarga data de col·laboració de Shakespeare amb els seus contemporanis John Fletcher i Christopher Marlowe,[91][92] i van confirmar l'opinió, basada en estudis més convencionals, que aquesta col·laboració s'havia produït.

Un estudi de 1999 va mostrar que un programa de xarxes neuronals va assolir un 70% de precisió en la determinació de l'autoria de poemes que encara no havia analitzat. Aquest estudi de la Vrije Universiteit va examinar la identificació de poemes de tres autors holandesos utilitzant només seqüències de lletres com "den". [93]

Un estudi va utilitzar xarxes de creences profundes (DBN) per al model de verificació d'autoria aplicable per a l'autenticació contínua (CA). [94]

Un problema d'aquest mètode d'anàlisi és que la xarxa pot arribar a ser esbiaixada en funció del seu conjunt d'entrenament, possiblement seleccionant autors que la xarxa ha analitzat amb més freqüència. [93]

Algorismes genètics

L'algoritme genètic és una altra tècnica d'aprenentatge automàtic utilitzada per a l'estilometria. Això implica un mètode que comença amb un conjunt de regles. Una regla d'exemple podria ser: "Si però apareix més d'1,7 vegades en cada mil paraules, llavors el text és l'autor X". El programa es presenta amb text i utilitza les regles per determinar-ne

l'autoria. Les regles es proven amb un conjunt de textos coneguts i a cada regla se li atorga una puntuació d'aptitud. No s'utilitzen les 50 regles amb les puntuacions més baixes. Les 50 regles restants reben petits canvis i s'introdueixen 50 regles noves. Això es repeteix fins que les regles evolucionades atribueixen correctament els textos.

Parelles rares

Un mètode per identificar l'estil s'anomena "parells rars", i es basa en hàbits individuals de col·locació. L'ús de certes paraules pot, per a un autor en particular, associar-se idiosincràticament amb l'ús d'altres paraules previsibles.

Atribució d'autoria en missatgeria instantània

La difusió d'internet ha desplaçat l'atenció de l'atribució d'autoria cap a textos en línia (pàgines web, blogs, etc.), missatges electrònics (correus electrònics, tuits, posts, etc.), i altres tipus d'informació escrita molt més curta que un llibre mitjà, molt menys formal i més diversa pel que fa a elements expressius com colors, maquetació, tipografies, gràfics, emoticones etc. Es van informar dels esforços per tenir en compte aquests aspectes tant a nivell d'estructura com de sintaxi. [95] A més, es van introduir indicacions idiosincràtiques i específiques del contingut (per exemple, models de temes i eines de verificació gramatical) per revelar opcions estilístiques deliberades. [96]

S'han emprat característiques estilomètriques estàndard per categoritzar el contingut d'un xat per missatgeria instantània,[97] o el comportament dels participants,[98] però els intents d'identificar els participants del xat són encara pocs i primerencs. A més, la similitud entre les converses orals i les interaccions de xat s'ha descuidat, alhora que és una diferència important entre les dades de xat i qualsevol altre tipus d'informació escrita.