Obras de *Referencia*

Análisis de textos desde la estilometría

Francisco Javier Blasco Pascual
y Cristina Ruiz Urbón

Ediciones Universidad Salamanca

ANÁLISIS DE TEXTOS DESDE LA ESTILOMETRÍA

Francisco Javier Blasco Pascual
y Cristina Ruiz Urbón

ANÁLISIS DE TEXTOS DESDE LA ESTILOMETRÍA

Ediciones Universidad
Salamanca

OBRAS DE REFERENCIA
47

1ª edición: diciembre, 2022
ISBN (impreso) 978-84-1311-763-8 / Depósito Legal: S 525-2022
ISBN (PDF) 978-84-1311-764-5
ISBN (ePub) 978-84-1311-765-2

Ediciones Universidad de Salamanca
https://www.eusal.es

Maquetación (impreso y digital) e impresión
Gráficas Lope
https://graficaslope.es/

♠

La colección Obras de Referencia de Ediciones Universidad de Salamanca
está acreditada con el sello de calidad en ediciones académicas CEA-APQ,
sello promovido por la Unión de Editoriales Universitarias Españolas (UNE),
y avalado por ANECA y FECYT.

♠

BLASCO PASCUAL, Francisco J., autor
Análisis de textos desde la estilometría / Francisco Javier Blasco Pascual
y Cristina Ruiz Urbón.—1a. edición: diciembre, 2022.—Salamanca :
Ediciones Universidad de Salamanca, [2022]
216 páginas.—(Obras de referencia ; 47)
DL S 525-2022.—ISBN 978-84-1311-763-8 (impreso).—ISBN 978-84-1311-764-5
(PDF).—ISBN 978-84-1311-765-2 (ePub)
1. Análisis del discurso-Metodología. 2. Estilometría. I. Ruiz Urbón, Cristina, autor.
81’32

A Pilar Celma y a José Manuel Fradejas,
personas queridas y maestros generosos

ÍNDICE

NOTA DE LOS AUTORES

En un momento en el que los macrodatos forman parte fundamental de todo lo que nos rodea, desde la medicina hasta la política, conviene recordar que un texto (literario o no) es, entre otras cosas, una suma de datos no estructurados. Ignorar esta realidad es renunciar a uno de los ángulos posibles de análisis e interpretación. Estas páginas tienen el objeto de señalar, sobre todo a los jóvenes investigadores, las posibilidades de esta vía.

En línea con el título, el libro que te presentamos pretende aplicar a los análisis de textos (textos de variada naturaleza, de géneros y de épocas distintos) los recursos y técnicas de la estilometría, con la mediación de herramientas y recursos informáticos capaces de procesar grandes cantidades de datos.

Queremos dejar muy claro, desde esta nota inicial, que lo que aquí presentamos no es un tratado de estilometría, ni siquiera es una introducción a esa disciplina. En el mismo sentido, aunque recurrimos en distintos lugares del texto al tratamiento de datos estadísticos y, en otros, nos servimos de algunas líneas de código R, las páginas que siguen nunca podrían ser (porque tampoco lo pretenden) un manual de estadística ni una especie de tutorial del lenguaje de programación R aplicado al análisis de textos. Existen ya excelentes manuales de introducción a la estadística y a R.

Lo que ofrecemos en las páginas que siguen es un muestrario de maneras, hasta cierto punto nuevas, de afrontar el análisis de textos con herramientas y técnicas de la estilometría, limitándonos a presentar estrategias y protocolos de actuación ante el texto. Mostraremos qué se puede buscar (y encontrar) con algunas técnicas y herramientas estilométricas, sin detallar cómo aplicarlas de forma general (algo que no sería realista, pues cada problema requiere una solución diferente y un planteamiento también distinto). En principio, dado que todos los recursos que hemos utilizado cuentan con sus propios tutoriales o guías de uso, nos remitimos a ellos. Por otro lado, para

un catálogo de herramientas útiles para el análisis de textos más amplio del que se muestra en estas páginas, remitimos a la base de datos de la página de estilometría de Javier Blasco y Cristina Ruiz Urbón (2019). Todas las que se citan en este libro están reseñadas ahí.

Este libro es una invitación a la aventura, especialmente destinado a aquellos que todavía observan con escepticismo, cuando no con desconfianza, los nuevos horizontes que para la filología iluminan las humanidades digitales en general y la estilometría en particular.

Consecuencia de esa invitación es el estilo divulgativo en que están escritas las páginas que siguen. Porque nuestro papel no es el del informático que trabaja en el desarrollo de nuevos algoritmos o herramientas; ni es el del experto en estadística. Nuestro papel es el del filólogo que, interpretando el trabajo de los expertos, examina sus posibilidades en el análisis de textos.

Esperamos que, para seguir nuestros planteamientos, el lector no precise de otro equipamiento que un poco de sentido común, salpimentado con unos conocimientos filológicos elementales. Para repetir algunos de nuestros planteamientos y aplicar las técnicas de análisis que proponemos, sí que el lector necesitará unos mínimos conocimientos del funcionamiento de R en la consola de RStudio, para lo que cuenta con un muy excelente manual de nuestro colega José Manuel Fradejas Rueda (2021), que explica paso a paso el código base de importantes recursos para el cosechado de textos de la web o de grandes repositorios, para el cálculo de frecuencias léxicas, para el etiquetado automático en clases de palabras, para la evaluación de la emotividad y de los tópicos de los textos, etc. A él remitiremos puntualmente en el momento que proceda.

Los autores de esta obra llevamos más de diez años realizando peritajes lingüísticos forenses que nos han colocado frente a casos de autoría, plagio, falsificación documental, suplantación de identidad, extorsión, acoso laboral, etc., y nuestra idea es trabajar desde ahora mismo en un segundo volumen de estilometría aplicada a casos prácticos, donde se explique con detenimiento tanto las herramientas más adecuadas a cada caso en particular, como los enfoque que puedan resultar más productivos.

Finalmente, queremos aprovechar esta nota para agradecer muy sinceramente el excelente trabajo de los anónimos informantes que leyeron, para su aprobación, nuestro original, y que tuvieron la generosidad y oportunidad de señalar puntualmente deficiencias y sugerir soluciones. Todas sus recomendaciones han sido rigurosamente incorporadas a la versión que ahora el lector tiene en sus manos.

I
¿QUÉ ES LA ESTILOMETRÍA?

La estilometría es una disciplina dedicada al análisis estadístico de textos escritos, basada en la cuantificación de diversos aspectos del discurso, tales como la longitud de las palabras o de las frases, la frecuencia de determinadas clases de palabras o partes de la oración (POS), el estudio de las palabras de función, las palabras de contenido, las colocaciones o las correlaciones…, sirviéndose para ello del concurso de otras disciplinas auxiliares, como la informática, la estadística y la filología[1]. En términos muy generales, podríamos decir que se trata de una disciplina que nos proporciona herramientas, métodos y protocolos para el análisis cuantitativo de los textos, sabiendo que podemos cuantificar, al menos, el léxico, las clases de palabras y las asociaciones y distribución de estas palabras en el texto (Bock, 1986 y Bock y Loebell, 1990). La historia de la disciplina ha sido trazada de una forma sintética, pero suficiente, por David Holmes y Judit Kardos (2003), lo que, remitiendo a su trabajo, nos libera ahora de extendernos en este punto.

Para una descripción básica de la estilometría, basta recordar que con ayuda del ordenador, basándonos en la cuantificación de determinados fenómenos lingüísticos y contando con las herramientas adecuadas a cada caso, podemos identificar, con un porcentaje de fiabilidad muy alto (si se dan unas condiciones mínimas), al autor de un texto dado; detectar fenómenos de plagio o de reutilización de textos previos; definir los usos de escritura de la persona que escribió un determinado documento; determinar el estado emocional del autor en el momento de la escritura; distinguir, en caso de que

1. Remitimos a la página www.estilometría.com, donde se da cuenta más detallada de herramientas y de procedimientos de la estilometría.

se trate de un texto escrito en colaboración, la porción de texto correspondiente a cada uno de los autores que participaron en su redacción; estudiar conflictos de marcas y casos de falsificación; pronosticar las posibilidades de éxito de una novela; y, en fin, extraer en cuestión de pocos minutos los grandes temas que vertebran el sentido tanto de un texto concreto como de un corpus muy extenso. En las páginas que siguen mostraremos, con ejemplos prácticos, varias de estas potenciales aplicaciones de la estilometría, útiles para el análisis de textos literarios, pero también en otros campos como la lingüística forense, la psicología, la política, la publicidad y, en general, para cualquier realidad fundada en la palabra.

Hoy, la conjunción de ciertos recursos informáticos y estadísticos nos permite enfrentarnos a los textos desde ángulos diferentes, y siempre complementarios, a los tradicionales, que son esencialmente de carácter filológico. En cuestión de muy pocos minutos, es posible convertir grandes cantidades de textos en bases de datos que, tratadas con diferentes recursos informáticos y sometidas a algoritmos de inteligencia artificial, nos pueden permitir localizar en un corpus (de millones de palabras) patrones verbales caracterizadores. De este modo, la lingüística forense (trazado de perfiles verbales; estudio de casos de plagio, de anónimos o amenazas; detección de la mentira; localización de ambigüedades y formulaciones equívocas en contratos; interpretación de artículos de la ley; descubrimiento de abusos en interrogatorios policiales; estudio de las estrategias de persuasión comunes a los diferentes tipos de estafas, etc.), la psicología (análisis de los textos para evaluación de estados emocionales de los pacientes), los estudios de género (análisis de las diferencias expresivas de hombres y mujeres), la historia literaria (cambios de estilo entre autores o épocas y detección de autoría), los estudios de opinión política o comercial sobre una producto o servicio o, en fin, la lingüística de corpus (creación de grandes archivos de textos digitales), son disciplinas que se han beneficiado de los recursos que ofrece la estilometría. Y, a la vez, todas estas disciplinas han aportado planteamientos y necesidades que han orientado el desarrollo de la estilometría en los últimos años y han contribuido positivamente a mostrar su eficiencia en el análisis de textos escritos.

Puesto que nuestra especialidad es la literatura, partiremos de un ejemplo de este campo. Los recursos informáticos y estadísticos con los que hoy cuenta el analista nos permiten enriquecer la lectura próxima (esto es, la lectura meditada o placentera de un texto) con la lectura distante (asistida por la informática y aplicada a grandes corpus de textos). Si hasta hace bien poco trabajar con un centenar de novelas era ya empresa titánica, hoy podemos

trabajar con miles de novelas. Y, de esta posibilidad, surge una pregunta: *¿lo digital modifica nuestro conocimiento de la literatura?*

La respuesta a la pregunta anterior es claramente afirmativa. Contamos con herramientas informáticas preparadas para extraer información de los textos a muy diferentes niveles (fonético, léxico, morfológico, sintáctico, semántico y pragmático) y se han elaborado algoritmos muy precisos relacionados con la inteligencia artificial. Hablamos de aprendizaje automático supervisado (la máquina aprende de datos introducidos por el usuario) y no supervisado (la máquina trabaja con datos no etiquetados previamente, buscando por sí misma patrones o relaciones entre ellos) y de algoritmos capaces de llevar a cabo operaciones múltiples con un volumen ingente de información, algo que hace apenas unos años hubiera sido imposible de imaginar. El modelo de aprendizaje supervisado utiliza datos de entrenamiento etiquetados para inferir una función de clasificación: en un caso de atribución de autoría, por ejemplo, creamos un archivo de entrenamiento con textos de los distintos candidatos para que la máquina identifique los patrones distintivos de cada uno de ellos y, contra ese archivo de entrenamiento, confrontamos el texto dubitado. En el análisis no supervisado los algoritmos son concebidos para manejar billones de datos y predecir conclusiones sobre dichos datos.

Hablamos de algoritmos que ni siquiera requieren suposiciones previas sobre las relaciones subyacentes entre las variables que someten a cálculo: en los casos que competen a la estilometría, el analista introduce los textos y el ordenador los procesa, encontrando patrones sobre los que hacer predicciones. El algoritmo acierta a comprender, por ejemplo, que un texto dado pertenece a una novela histórica, distinguiendo de manera inequívoca el estilo de este género del que presenta una novela policíaca (Fradejas Rueda, 2016); de la misma manera el ordenador, basándose en datos estilométricos, puede identificar, con un porcentaje muy cercano al 100% (si se dan las condiciones adecuadas), quién es el autor de un texto; extraer los temas dominantes de un escrito dado, así como el porcentaje con que un tema se impone sobre otro; determinar la legibilidad de un texto (importantísimo para la educación, para la comunicación en general y para la conformación de una administración más «amigable»); establecer cuantitativamente el porcentaje de emociones que componen una página o un capítulo de un texto dado; pronosticar (con notable fiabilidad) si una novela va a tener éxito de ventas o no; definir el perfil lingüístico o idiolecto de un autor (edad, sexo, procedencia geográfica, educación, personalidad, usos de escritura dominantes); distinguir en un texto escrito en colaboración la parte que corresponde a

cada una de las manos que en él han intervenido; discriminar entre lenguajes semejantes; y en fin, un cúmulo considerable de otras operaciones.

Junto a los algoritmos de *machine learning*, el análisis cuantitativo hace uso de técnicas y procedimientos paralelos, como aquellos en los que el texto analizado se confronta con *corpora* previamente seleccionados o con diccionarios (léxicos, más bien) de términos previamente etiquetados. Un gráfico, que tomamos de Medhat y otros (2014: 1095), resume visualmente el panorama en el que se sitúa cualquiera de las técnicas conocidas para el análisis cuantitativo de textos, por ejemplo para sus aspectos emocionales o afectivos:

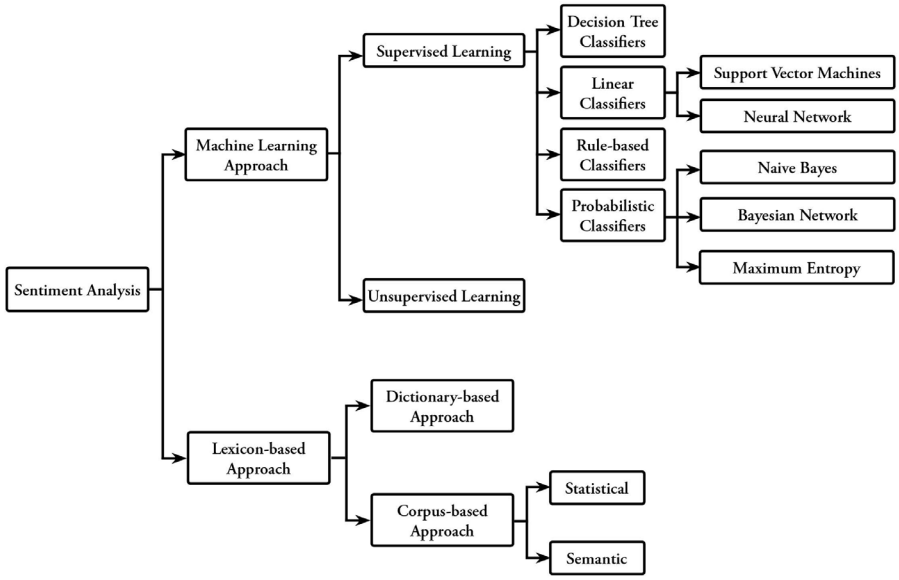

Figura 1. Técnicas para clasificar sentimientos.
Fuente: Medhat, Hassan y Korashy (2014: 1095).

Basta lo dicho para responder a la pregunta arriba planteada, sin ningún género de atenuación, con un rotundo sí: la estilometría está llamada a cambiar sustancialmente los estudios literarios.

Podríamos, dando un paso más, plantear una nueva pregunta, complementaria de la anterior: *¿Se pueden abordar cuestiones de complejidad cultural cuantitativamente?* Y la respuesta, igualmente, es afirmativa. La conjunción de lingüística, informática y estadística en el territorio de la estilometría modifica sustancialmente nuestra manera de relacionarnos con la literatura y, en general, con los textos escritos, sean o no de naturaleza literaria. Y dicha modificación es muy profunda. La estilometría no es una nueva teoría literaria,

alternativa, por ejemplo, al positivismo, al formalismo o al estructuralismo. La estilometría, en todo caso, da nombre a un enfoque complementario a todos los que tradicionalmente se han venido utilizando en los estudios de textos (literarios o no). Además de permitir el trabajo con volúmenes enormes de textos, lo que hace es abrir la posibilidad de plantearles a esos textos preguntas que hasta hace apenas unos años hubieran sido inimaginables.

Los algoritmos con que trabajan nuestras herramientas han cambiado la forma de acercarnos a los textos y el objeto de nuestras búsquedas en ellos. Porque nuestro análisis en «lectura distante» poco tiene que ver con la vivencia del hecho literario que experimentamos en una «lectura cercana». Hemos de tener muy claro que los textos literarios son ante todo actos comunicativos, objetos verbales resultado y desencadenantes de experiencias estéticas, en tanto que, convertidos en corpus de análisis estilométricos, dan acceso a la localización de patrones que nos permiten descubrir, en el envés de la textualidad, usos de escritura con enorme potencial para el análisis. Sin anular, ni mucho menos negar, el placer de la «lectura cercana» (que siempre será prioritaria en el aercamiento a un texto escrito con voluntad literaria), la «lectura distante» por la que apuesta la estilometría nos permite, por ejemplo, someter a un análisis libre de subjetividad varios miles de novelas del periodo de transición de los siglos XIX y XX (novelas canónicas y populares de todo género) y, desde ahí, corregir los problemáticos panoramas que las historias de la literatura ofrecen, por ejemplo, para lo que en dichas historias se etiqueta como novela realista, novela naturalista o novela modernista, rescatando matices y diferencias que se escapan a la lectura próxima del más avezado historiador (cuyo conocimiento se halla limitado siempre a un corpus mucho más reducido). Frente a la lectura tradicional, basada en la intuición, la estilometría se basa en el dato. Y muchos se asustan ante esta realidad, puesto que para la mayoría de filólogos (y de los lectores en general) es el «placer» del texto el que nos guía, al menos en un primer momento, y el que provoca las preguntas que nos hacemos ante el hecho literario: *¿qué nos dice este texto?, ¿cómo formula su mensaje?, ¿a quién se dirige?* Y otras muchas, que no es preciso recordar ahora.

Pero no se trata de prescindir de una forma tradicional de lectura, sino de enriquecerla añadiendo nuevos horizontes. Nunca podremos prescindir de la lectura cercana del texto, si no queremos suprimir el disfrute y la vivencia del mismo. La estilometría no pretende sustituir a la historia ni a la crítica literaria tradicionales, sino que busca complementar su trabajo proporcionando un recurso alternativo. La estilometría no nos aleja del texto. Al contrario:

nos permite encarar el análisis de corpus mucho más ricos y extraer de ellos información, a la que de otro modo nunca tendríamos acceso, para afrontar la lectura cercana con otros pertrechos.

El análisis cuantitativo en el que se basa la «lectura distante» –decíamos– permite plantearse preguntas que hubieran sido casi inimaginables hace unos pocos años. En efecto, una pregunta como *¿El suspense usa los mismos mecanismos en todos los períodos y géneros, y para todos los tipos de lectores, o sus técnicas varían según el tiempo, el lugar y el destinatario?,* sólo hubiera podido recibir respuestas provisionales y parciales derivadas de corpus limitados. El investigador, si era honesto, estaba condenado a confesar que sus juicios reducían su alcance al corpus abarcado por las lecturas que él había podido hacer. Un análisis comparativo del «suspense» en textos de más de 200 años supera con mucho las posibilidades de cualquier análisis tradicional. Por eso, una de las recomendaciones básicas para los doctorandos en la elección del tema de su tesis, clave fundamental de su futuro éxito o fracaso, tuvo siempre que ver con la acotación del corpus de trabajo. Y lo mismo cabe predicar de preguntas como las siguientes: *¿Qué variedad de discursos (lo que Bakhtin denomina «heteroglosia») concurren en un texto narrativo y cómo esta concurrencia varía según las épocas? ¿Qué términos se asocian con diversos sociolectos a lo largo del tiempo? ¿Qué conceptos académicos, políticos y sociales se asocian en torno a una palabra, por ejemplo «seguridad», a lo largo de la historia?*

No obstante, el salto de la tradicional estilística a la estilometría todavía provoca vértigo para muchos de los formados en los estudios de lengua y de literatura tradicionales. Es lo que suele ocurrir –y quizás es bueno que así sea– con las transformaciones en el mundo académico, si los naturales cambios (naturales, porque términos como *ciencia* y *progreso* o *avance* en el conocimiento son sinónimos) no se acompañan de las pertinentes garantías. En este sentido, los métodos de la estilometría, cuando menos, deberían cumplir con el *Marco de la razón de verosimilitud*, establecido por el Tribunal Supremo de EE.UU., después del fallo judicial de 1993, en el caso «Daubert v. Merrell Dow Pharmaceuticals, Inc.» Este «marco» recoge las «reglas de evidencia» que los jueces deberían contemplar para determinar si un método científico es fiable y, por tanto, merece ser admitido en una prueba pericial:

a) la metodología debe haber sido probada y debe ser replicable;
b) debe existir una tasa de error real o probable sobre la técnica aplicada;
c) deben existir y se deben mantener unos estándares para el control de la aplicación de la técnica;
d) la metodología debe haber sido sometida a revisión y publicación.