ISSN 1870-4069

Análisis automático de estilo de escritura en textos de longitud variable

Germán Ríos-Toledo

Tecnológico Nacional de México, Campus Tuxtla Gutiérrez, Chiapas, México
german_rios@ittg.edu.mx

Resumen. El análisis automático del estilo de escritura es una tarea que consiste en la creación de un modelo que represente el estilo de un autor. La representación del estilo de un autor, generalmente, se obtiene mediante la conteo de la frecuencia de uso de marcadores de estilo presentes en un texto. Para un análisis confiable, es importante considerar la cantidad de información de los textos en términos de palabras o en oraciones. En esta investigación se evaluaron novelas de diez diferentes autores utilizando un marcador de estilo conocido como n-grama de cuatro diferentes categorías. EL conjunto de novelas de cada autor se dividió en dos etapas. Cada novela se dividió en fragmentos de distintos tamaños en los que disminuye gradualmente el número de oraciones. Se entrenó un algoritmo de aprendizaje automático supervisado para predecir a cuál de las dos etapas de un autor pertenece un fragmento de texto. Los resultados de los experimentos mostraron por un lado, que la categoría de n-gramas basados en información sintáctica mostraron un porcentaje de clasificación superior a las categorías restantes y por el otro, que determinar la cantidad apropiada de texto depende en gran medida del estilo de escritura de cada autor.

Palabras clave: marcador de estilo, n-gramas, estilometría.

Automatic Analysis of Writing Style in Texts of Variable Length

Abstract. Automatic analysis of the writing style is a task that consists in creating a model that represents the style of an author. The representation of an author's style is generally obtained by counting the frequency of use of style markers present in a text. For a reliable analysis, it is important to consider the amount of information in the texts in terms of words or sentences. In this research, novels by ten different authors were evaluated using a style marker known as n-grams of four different categories. The set of novels of each author was divided into two stages. Each novel was divided into fragments of different sizes in which the number of sentences gradually decreases. A supervised machine learning algorithm was trained to predict to which of the two stages

of an author a piece of text belongs. The results of the experiments showed, on the one hand, that the category of n-grams based on syntactic information showed a higher classification percentage than the remaining categories, and on the other, that determining the appropriate amount of text depends largely on the style of writing of each author.

Keywords: style markers, n-grams, stylometry.

1. Introducción

La estilometría es una disciplina que se basa en la presunción de que cada persona tiene un estilo de escritura. El análisis automático del estilo de escritura tiene utilidad en áreas como el derecho penal o civil, pues coadyuva a la detección de plagio, la creación de perfiles de autor y a la protección del anonimato.

En el análisis de estilo comúnmente se utilizan las palabras como marcadores de estilo: palabras con contenido semántico (sustantivos, verbos, adjetivos, adverbios); palabras funcionales (preposiciones, adverbios, artículos, pronombres, adjetivos); longitud de palabras; categorías gramaticales; errores de escritura; lemas; entre otras.

El análisis de estilo se enfoca en la forma del texto y no en su contenido. Actualmente, el reto principal en este tipo de estilo es identificar marcadores de estilo robustos al contenido temático y al tipo de documento. En este sentido, el uso de la información sintáctica es una alternativa potencial. Los analizadores sintácticos modernos obtienen la información sintáctica de cada oración y la representan de forma estructurada en un árbol de dependencia. Dichos árboles muestran el orden no lineal entre las palabras, las categorías gramaticales y los nombres de las relaciones sintácticas existentes. Las características estilométricas basadas en información sintáctica posibilitan el desarrollo de análisis de estilo más completos.

El análisis automático de estilo puede abordarse con un enfoque de aprendizaje supervisado. Tanto los textos de entrenamiento como los de prueba se representan como vectores de características bidimensionales (matrices).Las celdas de la matriz contienen la frecuencia un marcador en un texto particular. Los algoritmos de aprendizaje reconocen patrones en las frecuencias y generan un modelo, dicho modelo recibe textos no vistos en la etapa de entrenamiento y genera predicciones. Un modelo ideal sería aquel que clasifica correctamente todos los textos.

Las secciones restantes que conforman el artículo son: Trabajos relacionados, este apartado describe los estudios previos que abordan el problema de la cantidad de texto requerida para tareas del Procesamiento de Lenguaje Natural como Atribución de Autoría. Metodología, aquí se detalla el proceso típico que se sigue para el análisis automático de textos. La sección de Organización de los experimentos describe la forma en que se realizaron las pruebas de clasificación con el algoritmo de aprendizaje automático supervisado y por último, las Conclusiones derivadas de los resultados obtenidos.

2. Trabajos relacionados

En la tarea del Atribución de Autoría, Eder [4] utilizó pruebas controladas con textos de diferente longitud, idiomas y géneros. utilizó la métrica Delta propuesta por John F. Burrows [2], la cual mide la diferencia entre dos textos basándose en las palabras frecuentes. Extrajo muestras de textos formadas por 200, 400, 600, 800, ..., 20000 palabras. Las palabras se obtuvieron en forma aleatoria y en secuencia. Eder concluyó que el uso de ejemplos de 2500 palabras difícilmente proporcionará un resultado confiable.

En [14] aplicaron la técnica de los núcleos de secuencia de palabras en tarea de Atribución de Autoría. Evaluaron un conjunto de textos relativamente cortos de 50 periodistas que cubrían más de un tema aplicando el enfoque de cadenas de Markov. Crearon textos de 312, 625, 1250, 2500 y 5000 palabras y de ellos obtuvieron 1750, 3500, 7000, 14000 y 28000 caracteres. Los investigadores indicaron que la cantidad de texto para el entrenamiento de los algoritmos tiene más influencia que la cantidad de textos de prueba. Además concluyeron que se requieren aproximadamente entre 1250 y 5000 palabras en los textos de entrenamiento para obtener un rendimiento relativamente bueno.

Corney et al. [3] realizaron experimentos para identificar la autoría de correos electrónicos utilizando marcadores apropiados para este tipo de mensajes. Dichos mensajes contenían hasta 964 palabras, con una longitud promedio de 92 palabras. Utilizaron el algoritmo de aprendizaje automático supervisado SVM para discriminar entre las clases de autoría. Descubrieron que aproximadamente 20 mensajes con aproximadamente 100 palabras cada uno, deberían ser suficientes para discriminar la autoría en la mayoría de los casos. Mencionaron que el rendimiento del clasificador mejoró cuando agregaron un conjunto de características específicas de correo electrónico.

Luyckx y Daelemans [9] analizaron ensayos de un mismo tópico. Los ensayos contenían aproximadamente 1400 y provenían de 145 estudiantes. Utilizaron palabras y n-gramas de etiquetas POS con el algoritmo de aprendizaje automático SVM. En sus conclusiones, argumentaron que su propuesta mostró solidez al tratar con datos limitados, ya que de los 145 autores, casi el 50 % de los textos fueron clasificados correctamente.

En el ámbito de la Atribución de Autoría, Luyckx y Daelemans [10] evaluaron el efecto que tienen el número de autores y la cantidad de textos de entrenamiento en correos electrónicos, validaron su propuesta por medio de Aprendizaje Automático. Algunos de marcadores que utilizaron fueron longitud de sentencia, longitud de palabra, palabras función y de contenido y n-gramas de caracteres. En sus conclusiones indicaron que, como se esperaba, la precisión del clasificador se deterioró al aumentar el número de autores y disminuir el tamaño de los textos de entrenamiento. Además, indicaron que los n-gramas de caracteres son robustos a los cambios en el tamaño del conjunto de autores y el tamaño de los textos.

Feiguina y Hirst [6] evaluaron bigramas sintácticos como marcadores de estilo para la Atribución de Autoría en textos cortos. Utilizaron un analizador sintáctico y el algoritmo de aprendizaje automático SVM. Evaluaron novelas del

corpus Brontë Sister, mismas que dividieron en fragmentos de 200, 500 y 1000 palabras. Encontraron que los bigramas sintácticos fueron útiles al discriminar textos cortos del Corpus Brontë pero no para datos forenses simulados donde las distinciones sintácticas parecían menos necesarias. Argumentan que esto podría atribuirse a un desbalance de datos y al pequeño tamaño del conjunto de datos forense.

3. Metodología

La Figura 1 muestra el esquema general propuesto para el análisis de cambio de estilo de escritura con enfoque de aprendizaje automático supervisado. Las novelas evaluadas se descargaron del proyecto *Gutenberg*[1], éstas incluyen información de la editorial, datos biográficos del autor, semblanzas entre otras cosas. En el preprocesamiento se eliminó toda esa información así como las oraciones de uno y dos palabras, ya que los 3-gramas de palabras, etiquetas POS y relaciones sintácticas requieren al menos tres palabras.

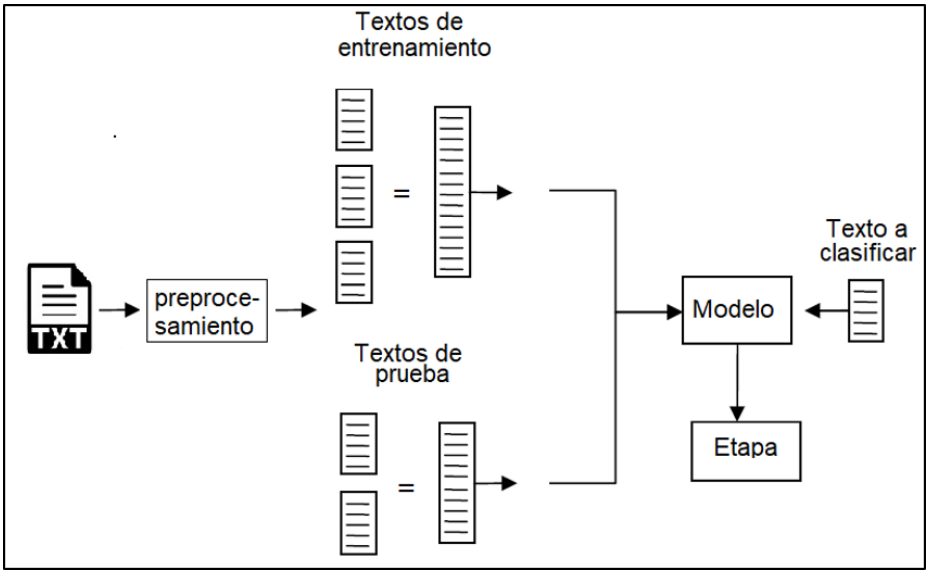

Fig. 1. Esquema de detección de cambio de estilo de escritura

La parte medular de esta investigación fue dividir cada novela en textos de diferente tamaños. Esta acción tiene dos consecuencias: primero, aumenta el número de textos disponibles y segundo, disminuye la cantidad de información en cada uno de ellos. Cada novela se dividió en 10 fragmentos con el mismo

[1] `https://www.gutenberg.org/`

número de sentencias. La notación para referir a cada tamaño de texto fue la siguiente: **1** fragmento que contiene la novela completa, **2** fragmento que contiene la mitad de una novela, **3** fragmento que contiene un tercio de una novela, así sucesivamente. La Tabla 1 muestra el número de sentencias del autor Booth Tarkington. Para el resto de los autores se aplicó el mismo procedimiento.

Tabla 1. Distribución de sentencias en novelas de Booth Tarkington.

Novela	**Tamaño del texto**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Canaan	4,598	2,299	1,532	1,149	919	766	656	574	510	459
Gentleman	5,350	2,675	1,783	1,337	1,070	891	764	668	594	535
Penrod	3,841	1,740	1,160	870	768	640	548	480	426	384
Seventeen	3,917	1,958	1,305	979	783	652	559	489	435	391
Turmoil	5,892	2,946	1,964	1,473	1,178	982	841	736	654	589
Vanrevels	2,802	1,401	934	700	560	467	400	350	311	280

Después de preprocesamiento, se realizó el etiquetado y el análisis sintáctico de cada texto. Las etiquetas POS se generaron con el etiquetador de la herramienta NLTK[2] y el análisis sintáctico se llevó a cabo con Stanford Parser[3]. Cuando se utilizan n-gramas se debe definir el valor de n, dicho valor indica el número de términos del n-grama. Valores comunes de n son 1, 2, 3, 4 y 5. Estudios previos reportaron que el más apropiado para n es 3: detección de plagio [1], atribución de autoría [5,11,15], categorización de textos [13] e identificación de autores[8].

Por otro lado, debe establecerse la frecuencia mínima con la que debe aparecer un n-grama. Las frecuencias 1 y 2 generan una gran cantidad de marcadores y contribuyen muy poco o nada al estilo de escritura del autor. En esta investigación se trabajó con n-gramas de frecuencia mayor o igual a 3. Para obtener los n-gramas de caracteres, palabras y etiquetas POS se utilizó el programa *text2ngram*[4], un software libre bajo licencia GPL. Dicho programa requiere como parámetros el tipo de n-grama, la longitud, la frecuencia y el texto de entrada. Los 3-gramas de relaciones sintácticas se generaron con un programa desarrollado en Python [12], este programa genera tres tipos de n-gramas sintácticos: de palabras, etiquetas POS y relaciones de dependencia; requiere como parámetros el valor de n y el archivo que contiene la información sintáctica de cada las oraciones. Además, se programó una rutina en Python para realizar el conteo de 3-gramas de cada sentencia, la frecuencia se establece dentro del propio código. La Tabla 2 muestra la información generada en esta etapa. Información similar se generó para los 3-gramas de caracteres, etiquetas POS y relaciones de dependencia.

[2] https://www.nltk.org/

[3] `https://nlp.stanford.edu/software/lex-parser.shtml`

[4] `https://homepages.inf.ed.ac.uk/lzhang10/ngram.html`

Tabla 2. Total de 3-gramas de palabras por autor.

Autor	Tamaño de texto									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BT	2,762	1,679	1,215	1,020	878	771	714	660	593	572
CD	19,896	13,068	10,138	8,503	7,420	6,703	6,176	5,750	5,400	5,100
ER	3,693	2,423	1,847	1,511	1,284	1,138	1,042	965	881	824
FM	7,488	4,742	3,609	2,998	2,565	2,327	2,082	1,911	1,805	1,646
GM	5,285	3,232	2,416	1,951	1,690	1,495	1,321	1,191	1,108	1,060
GV	2,841	1,617	1,131	914	725	637	559	478	429	402
IM	8,761	5,554	4,176	3,534	2,965	2,631	2,401	2,248	2,052	1,979
JB	3,982	2,393	1,794	1,444	1,235	1,077	952	871	804	740
LT	2,601	1,483	1,056	780	669	566	471	396	358	339
MT	5,785	3,821	2,961	2,656	2,326	2,072	1,918	1,864	1,706	1,636

Posteriormente se crearon matrices *término-documento.* para grupo de 3-gramas. En estas matrices, las filas representan documentos, las columnas representan 3-gramas y las celdas contienen la frecuencia. La Tabla 3 muestra una vista parcial de la matriz de de caracteres en las novelas del autor Booth Tarkington.

Tabla 3. Matriz término-documento de Booth Tarkington.

Novela	3-gramas de caracteres									
	the	**and**	**ing**	**her**	**you**	**his**	**ere**	**tha**	**was**	**she**
Gentleman	8,977	4,307	3,018	2,465	1,525	1,652	1,820	1,421	1,393	1,095
Vanrevels	5,270	2,358	1,905	1,630	1,024	1,040	924	1,036	934	755
Canaan	6,370	2,913	2,420	1,873	1,479	1,487	1,287	1,250	1,117	933
Ramsey	2,929	1,577	1,210	1,027	789	646	642	698	606	447
AliceAdams	5,147	2,511	2,803	2,761	2,412	1,176	1,068	1,232	880	1,551
Julia	4,729	2,438	2,081	2,316	1,320	1,155	996	1,073	1,009	1,156

Algunos 3-gramas pueden tener frecuencias muy altas y en consecuencia predominan sobre las frecuencias bajas. Previendo este fenómeno, se realizó una normalización de las frecuencias. Existen diversos métodos de estandarización. Aquí se aplicó el método que consiste en extraer la media del grupo del valor de cada variable y dividir el valor resultante por la desviación estándar. Existen otros métodos para normalizar como el valor *tf-idf* entre otros.

4. Entorno experimental

La colección de novelas de cada autor se ordenó de forma cronológica con base en el año de publicación, de la más antigua a la mas reciente. Tomando una novela como punto medio en el tiempo, las novelas que la preceden pertenecen a la etapa 1 y las que la suceden pertenecen a la etapa 2. El conjunto de autores evaluados se muestran en la Tabla 4. De aquí en adelante, los autores se refieren a través de la leyenda que se muestra en el Cuadro.

Tabla 4. Conjunto de autores evaluados

Autor	Leyenda
Booth Tarkington	BT
Charles Dickens	CD
Edgar Rice	ER
Frederick McDonald	FM
George McDonald	GM
George Vaizey	GV
Iris Murdoch	IM
John Buchan	JB
Louis Tracy	LT
Mark Twain	MT

En Aprendizaje Automático los datos se dividen en conjuntos de entrenamiento y prueba, en estos experimentos la proporción de cada conjunto fue 66 % y 33 % respectivamente. La regresión logística es un algoritmo de clasificación supervisada que se utiliza para predecir una variable dependiente categórica. En los experimentos se utilizó la implementación de *scikit-learn*[5]. En cada una de las dos etapas el número textos es el mismo, por lo que la métrica *exactitud* resulta apropiada para medir la eficiencia del clasificador [7]. La exactitud indica la proporción de textos clasificados correctamente. Dado que la clasificación es binaria, la probabilidad de asignar al azar un texto a la etapa correcta es de 50 %, este valor se consideró la línea de base.

5. Resultados

5.1. Promedios generales de exactitud por autor

En la categoría de 3-gramas de caracteres, la Figura 2 muestra que los autores mejor clasificados fueron BT y ER al superar el 90 % de exactitud. Por otro lado, el autor GV muestra el promedio más bajo al no superar la línea base. El resto de los autores muestran valores que oscilan entre el 70 % y el 80 %. Aparentemente para algunos autores, disminuir la cantidad de información en los textos no fue un factor relevante.

La Figura 3 muestra la exactitud de los autores utilizando 3-gramas de palabras. La mayor parte de los autores superan el 70 % de exactitud. Al comparar con 3-gramas de caracteres, se observa que autores como GB obtuvieron mejor resultado mientras que otros como ER y BT disminuyeron sus promedios de exactitud.

La Figura 4 muestra los resultados en la categoría de etiquetas POS. La tendencia en los porcentajes de clasificación correcta se mantiene. Sin embargo, contrario a lo obtenido en 3-gramas de caracter y palabras, el autor GM presenta una disminución significativa ya que no supera el 60 %. En este punto es posible

[5] `https://scikit-learn.org/stable/`

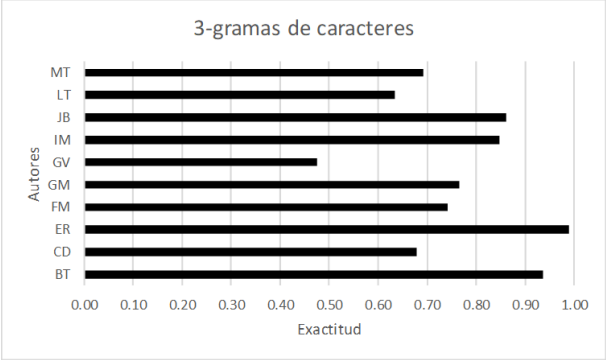

Fig. 2. Exactitud promedio en 3-gramas de caracter

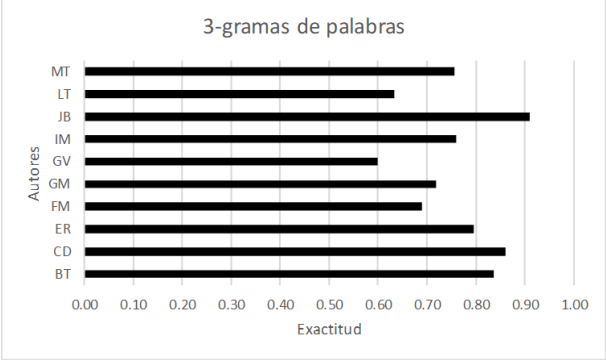

Fig. 3. Exactitud promedio en en 3-gramas de palabras

observar que los autores GV y LT son los que mantienen resultados muy cercanos a la línea base.

La Figura 5 muestra los resultados que se obtuvieron en los 3-gramas de relaciones de dependencia. Destaca que con excepción de LT, todos los autores superan holgadamente el 70 % de exactitud. Contrario a la tendencia mostrada en los otros 3-gramas, GV mostró un incremento significativo con exactitud cercana al 80 %. Algo similar ocurrió con IM, logrando casi el 100 % de exactitud. En general, en esta categoría de 3-gramas todos los autores mejores sus porcentajes independientemente del tamaño del texto.

5.2. Exactitud por tamaño de texto

La Tabla 5 muestra los resultados en 3-gramas de caracteres. Los autores mejor clasificados son BT y ER, BT obtiene 96 % en bloques de tamaño 4 y en el resto supera el 90 %, mientras que ER logra 100 % en los primeros cuatro bloques. Los casos atípicos son CD, GM, GV y LT, quienes logran sus mejores

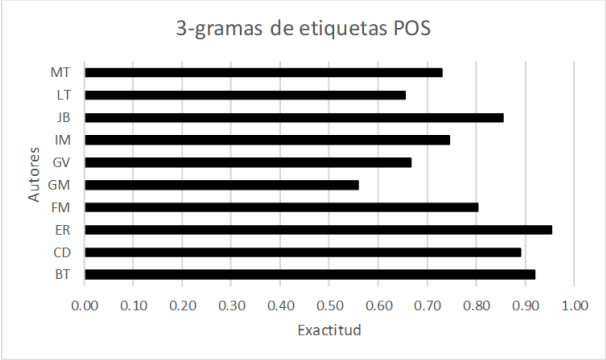

Fig. 4. Exactitud promedio en 3-gramas de etiquetas POS

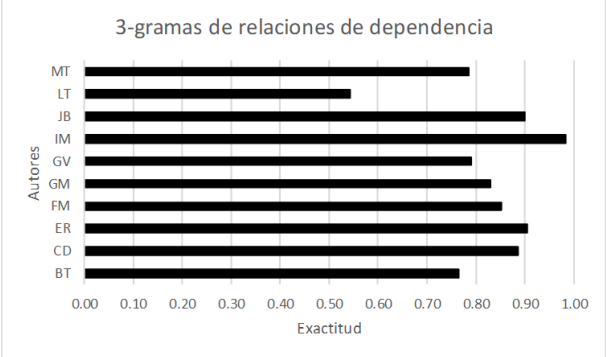

Fig. 5. Exactitud promedio en 3-gramas de relaciones de dependencia

resultados a partir de bloques de tamaño 7. También se observa que GV no superó el 50 % de exactitud en ninguno de los distintos bloques. El resto de los autores muestran mejores resultados en los bloques de tamaño 1 y 2, siendo éstos los que poseen mayor cantidad de información.

En la Tabla 6 se observa que en los 3-gramas de palabras todos los autores logran porcentajes más altos en textos de tamaño 1 y 2. Por otro lado, CD, IM y JB tienen resultados constantes a traves de los distintos tamaños. En el resto de los autores la exactitud disminuyó de forma gradual al reducir la cantidad de información.

La Tabla 7 muestran los resultados en 3-gramas de etiquetas POS. Los autores BT, ER y JB son los que muestran los porcentajes de clasificación más altos. El autor FM logra máxima exactitud en bloque de tamaño 4, mientras que el resto de los autores lo hace en bloques de texto de tamaño 1 y 2. Un caso peculiar se observó con CD. En el bloque 1 logra 72 % de exactitud y en los bloques 2, 5 y 8 un 92 %. Además, en el resto de los bloques supera el 90 %.

Tabla 5. Exactitud por tamaño de texto en 3-gramas de caracteres.

Autor	Tamaño de texto									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BT	94	92	94	**96**	94	94	93	94	91	94
CD	56	67	65	65	67	69	72	**74**	72	72
ER	**100**	**100**	**100**	**100**	97	**100**	99	99	96	98
FM	61	69	76	76	74	76	**79**	75	78	77
GM	72	78	76	78	77	77	**79**	76	77	76
GV	**50**	**50**	48	46	46	49	47	44	48	47
IM	83	**92**	85	86	83	83	86	84	85	81
JB	**89**	86	87	82	84	85	87	87	88	87
LT	**72**	69	63	64	63	57	65	64	61	57
MT	61	58	72	69	72	71	69	72	73	**76**

Tabla 6. Exactitud por tamaño de texto en 3-gramas de palabras.

Autor	Tamaño de texto									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BT	**100**	97	91	86	82	79	76	81	72	73
CD	**89**	**89**	87	86	87	85	88	85	84	80
ER	**89**	83	**89**	82	81	81	78	78	67	68
FM	**78**	72	65	65	69	69	72	64	67	69
GM	**78**	72	61	76	77	73	**78**	67	70	67
GV	61	**67**	61	61	46	59	60	63	56	66
IM	72	**81**	78	75	73	75	78	74	77	76
JB	89	**92**	91	90	93	90	92	92	90	91
LT	72	**78**	65	68	61	68	58	59	51	53
MT	83	83	**85**	76	78	74	73	69	69	67

Tabla 7. Exactitud por tamaño de texto en 3-gramas de etiquetas POS.

Autor	Tamaño de texto									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BT	**100**	**100**	96	97	97	88	87	85	85	84
CD	72	92	91	89	**92**	91	90	**92**	90	91
ER	**100**	**100**	94	96	97	97	95	92	92	91
FM	78	81	81	**83**	82	79	82	79	78	80
GM	**61**	47	56	53	56	56	60	57	56	57
GV	**72**	67	61	**72**	67	67	63	65	66	67
IM	**78**	69	76	71	76	74	75	74	75	77
JB	**100**	83	89	85	83	82	83	84	83	83
LT	67	**72**	63	68	69	63	62	61	64	65
MT	**78**	75	74	71	72	74	73	69	72	71

La Tabla 8 muestra que en la categoría de 3-gramas de relaciones de dependencia. La máxima exactitud de cada autor ocurrió en textos de tamaño 1 y 2. IM mostró una clasificación casi perfecto sin importar el tamaño de texto. LT

apenas superó la línea base, sin embargo logra su mejor resultado en bloques de texto de tamaño 9.

Tabla 8. Exactitud por tamaño de texto en 3-gramas de relaciones de dependencia.

Autor	**Tamaño de texto**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BT	**83**	**83**	80	78	78	70	73	71	76	73
CD	89	**92**	91	86	90	86	87	90	88	87
ER	**100**	94	93	92	88	89	87	88	88	85
FM	89	**92**	85	89	83	81	85	85	81	82
GM	**89**	86	85	85	84	84	81	81	78	77
GV	**83**	78	80	79	78	79	79	80	78	76
IM	**100**	**100**	**100**	96	97	**100**	98	98	96	98
JB	**100**	92	94	92	87	90	83	87	89	86
LT	50	50	56	53	53	56	55	56	**61**	52
MT	78	**83**	78	81	78	82	79	76	77	74

6. Conclusiones

El análisis automático de estilo de escritura se refiere a la forma de un texto y no a su contenido. Por ello, se deben tomar en cuenta dos aspectos. Primero, el marcador estilo utilizado. Es recomendable recurrir a marcadores que en la medida de lo posible sean inmunes al tipo de documento y al tópico que estos tratan. En cierta forma, la frecuencia de uso de las palabras dan cuenta del tópico de documento y de alguna forma influirán en la frecuencia de los n-gramas de caracteres. Para superar esta barrera, se ha explorado el uso n-gramas de etiquetas POS y relaciones de dependencia. Las etiquetas POS o categorías gramaticales indican el tipo de palabra que se esta usando en el texto, como adjetivos, verbos o sustantivos. Así, el análisis de estilo mostrará con que frecuencia una persona utiliza categorías de palabras. Estos marcadores revelan patrones muy distintos a los que ocurren cuando el texto se analiza en su forma lineal. La información sintáctica de una oración se muestra en forma de árboles, que muestran como incluso palabras que aparecen distantes entre sí están relacionadas, lo que permite descubrir nuevos patrones. Los resultados de estos experimentos mostraron que el uso de la información sintáctica es una alternativa viable para crear marcadores de estilo.

El segundo aspecto a considerar es la cantidad de información que hay en los textos en términos de palabras. Los trabajos relacionados dan cifras exactas para la tarea de Atribución de Autoría, las cuales varían en función del tipo de marcador y el tipo de documento (correos, ensayos, novelas). Esta investigación se enfoca en el cambio de estilo de escritura a través del tiempo por medio de aprendizaje automático supervisado. El objetivo principal fue observar el efecto que tiene sobre el clasificador disminuir la cantidad de oraciones en cada uno de los textos. Cada autor tiene un estilo de escritura propio por lo que

una cifra que para un autor resultó favorable en algún tipo particular de n-grama puede no serlo para otro. Por otro lado, en algunos autores disminuir la cantidad de información no influyó de forma significativa en sus resultados. Es necesario realizar investigaciones más exhaustivas para garantizar que ciertos valores resultan apropiados para un análisis de cambio de estilo de escritura independientemente del autor.

Referencias

1. Barrón-Cedeño, A., Rosso, P.: On automatic plagiarism detection based on n-grams comparison. In: European conference on information retrieval. pp. 696–700. Springer (2009)
2. Burrows, J.: 'delta': a measure of stylistic difference and a guide to likely authorship. Literary and linguistic computing 17(3), 267–287 (2002)
3. Corney, M.W., Anderson, A.M., Mohay, G.M., de Vel, O.: Identifying the authors of suspect email (2001)
4. Eder, M.: Does size matter? authorship attribution, small samples, big problem. Digital Scholarship in the Humanities 30(2), 167–182 (2015)
5. Escalante, H.J., Solorio, T., Montes-y Gómez, M.: Local histograms of character n-grams for authorship attribution. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 288–298. Association for Computational Linguistics (2011)
6. Feiguina, O., Hirst, G.: Authorship attribution for small texts: Literary and forensic experiments (2007)
7. García, V., Mollineda, R.A., Sánchez, J.S.: Index of balanced accuracy: A performance measure for skewed class distributions. In: Iberian conference on pattern recognition and image analysis. pp. 441–448. Springer (2009)
8. Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: International conference on artificial intelligence: Methodology, systems, and applications. pp. 77–86. Springer (2006)
9. Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). pp. 513–520 (2008)
10. Luyckx, K., Daelemans, W.: The effect of author set size and data size in authorship attribution. Literary and linguistic Computing 26(1), 35–55 (2011)
11. Posadas-Duran, J.P., Sidorov, G., Batyrshin, I.: Complete syntactic n-grams as style markers for authorship attribution. In: Mexican International Conference on Artificial Intelligence. pp. 9–17. Springer (2014)
12. Posadas-Durán, J.P., Sidorov, G., Batyrshin, I., Mirasol-Meléndez, E.: Author verification using syntactic n-grams. Working notes papers of the CLEF (2015)
13. Rahmoun, A., Elberrichi, Z.: Experimenting n-grams in text categorization. Int. Arab J. Inf. Technol. 4(4), 377–385 (2007)
14. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 482–491 (2006)
15. Sapkota, U., Solorio, T., Montes, M., Bethard, S., Rosso, P.: Cross-topic authorship attribution: Will out-of-topic data help? In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1228–1237 (2014)