ISSN 1870-4069

Nuevo enfoque para la extracción de características en la clasificación de textos para la atribución de autoría

Omar González Brito[1], José Luis Tapia Fabela[1], Silvia Salas Hernández[2]

[1] Universidad Autónoma del Estado de México,
Unidad Académica Profesional Tianguistenco,
México
[2] Universidad Autónoma del Estado de México,
Centro Universitario Atlacomulco,
México

gonzalezbritoomar@gmail.com, joseluis.fabela@gmail.com,
salashernandezsilvia@gmail.com

Resumen. La tarea de atribución de autoría se ha realizado bajo dos enfoques: basado en perfil e instancias, a través del análisis de características textuales o rasgos lingüísticos que permiten encontrar el estilo de escritura del autor. El conjunto de datos es uno de los principales problemas que enfrentan los enfoques y métodos de atribución de autoría debido a que no siempre se cuenta con un conjunto de documentos representativos de cada autor. En los enfoques de perfil e instancias, al extraer las características a partir del conjunto de documentos de los autores, se genera una alta dimensionalidad de características. Por lo que se propone un nuevo enfoque que no dependa del conjunto de documentos para la extracción de características, la clasificación de textos se realizó con máquina de soporte vectorial como método de aprendizaje supervisado. Este determina si están contenidas todas las características de un autor en un solo documento que describan su estilo de escritura. Para la experimentación se trabajó con tres *corpus* (C10, C50 y PAN12), estos fueron seleccionados basándose en la revisión de la literatura. De acuerdo a los resultados obtenidos se concluyó que el enfoque muestra resultados superiores al estado del arte en muestras desbalanceadas, resultados consistentes cuando es evaluado en diferentes contextos y robusto al analizar 10 y 50 autores. A partir de este enfoque se puede determinar que con 500 palabras sin repetir se puede identificar el estilo de escritura de un autor, presentando una exactitud de clasificación del 79.68%.

Palabras clave: Atribución de autoría, clasificación de textos, máquina de soporte vectorial, extracción de características, enfoques de autoría.

New Approach for the Extraction of Characteristics in the Classification of Texts for Attribution of Authorship

Abstract. The task of attribution of authorship has been carried out under two approaches: based on profile and instances, through the analysis of textual features or linguistic features that allow finding the author's writing style. The

data set is one of the main problems faced by attribution approaches and methods of authorship because it does not always have a set of documents representative of each author. In profile and instance approaches, by extracting the features from the authors' set of documents, a high dimensionality of features is generated. Therefore, a new approach is proposed that does not depend on the set of documents for the extraction of features, the classification of texts was carried out with a vector support machine as a supervised learning method. This determines if all the features of an author are contained in a single document that describes her writing style. For experimentation, we worked with three corpora (C10, C50 and PAN12), these were selected based on the literature review. According to the results obtained, it was concluded that the approach shows results superior to the state of the art in unbalanced samples, consistent results when evaluated in different contexts and robust when analyzing 10 and 50 authors. From this approach, it is determined that an author's writing style is contained in 500 words without repeating, presenting a classification accuracy of 79.68%.

Keywords: Attribution of authorship, classification of texts, vector support machine, extraction of characteristics, approaches of authors.

1. Introducción

El problema computacional de identificación de autoría ha sido abordado bajo dos enfoques; basado en perfil e instancias, utilizando modelos de clasificación generativos o discriminativos [1]. En el enfoque basado en perfil se extrae el estilo de escritura de cada autor, este es creado a través de la concatenación de los documentos obteniendo un solo documento por autor, del cual son extraídas las características para el entrenamiento del modelo [1, 2]. A diferencia del enfoque basado en instancias, la extracción de características se realiza por cada documento del autor, una vez extraídas forman un solo conjunto con el que se realiza el entrenamiento del clasificador y se genera el modelo de atribución de autoría. El enfoque más utilizado es el basado en instancias mediante la clasificación de textos [1-5].

En [5] se implementa un enfoque de instancias mediante los modelos de aprendizaje: máquina de soporte vectorial (*Support Vector Machine*, por sus siglas en inglés SVM) y máquina de soporte tensorial implementado el método del término más frecuente para seleccionar los 2500 términos más frecuentes. En este trabajo analizan las muestras balanceadas y desbalanceadas volviéndose uno de los trabajos más referenciados con el *corpus* C10. El proceso de extracción de características se realiza con el conjunto de documentos de los autores. Sin embargo, no siempre existe el número de documentos para cada autor por lo que se requiere determinar si en un solo documento se puede encontrar el estilo de escritura de un autor.

En la presente investigación se propone un nuevo enfoque para atribución de autoría, este propone utilizar un método de clasificación de textos extrayendo las características de un solo documento del conjunto de entrenamiento sin la implementación de métodos de selección de características, analizando este nuevo enfoque con características léxicas (*n-gramas* y bolsa de palabras), con una representación booleana y máquina de soporte vectorial como método de aprendizaje supervisado.

Para la experimentación se utilizaron tres diferentes *corpus:* C50, C10 y PAN12, analizando diferentes contextos, tamaños de muestras y diferentes números de autores.

2. Trabajos relacionados

Actualmente los métodos del estado del arte analizan ambos enfoques: perfil e instancias como se muestra en los trabajos de [2, 6]. En [2] realizaron una clasificación mediante los enfoques de perfil e instancias con el modelo LDA (*Latent Dirichlet Allocation,* por sus siglas en inglés LDA) para el análisis de datos dispersos y de alta dimensionalidad, utilizando la medida de similitud de coseno para determinar la autoría. Este modelo utilizó el conjunto de documentos de entrenamiento para la extracción de las características, posteriormente implementa el modelo LDA para realizar la selección de características.

En el trabajo presentado en [6] se realizó la construcción de grafos sintácticos bajo los enfoques de instancias y perfil debido que al implementar las técnicas de bolsa de palabras o *n-gramas* se pierde información sintáctica y semántica, caso que en los grafos sintácticos se conserva, para determinar la autoría usaron la métrica de similitud de coseno, esta ha sido implementado en las dos tareas de autoría; verificación y atribución. En los enfoques de perfil e instancias, al extraer las características a partir del conjunto de documentos de los autores, se genera una alta dimensionalidad de características, debido a esto surge la necesidad de implementar métodos de selección de características para extraer las más representativas, como lo realizaron [5, 7, 8].

El conjunto de datos es uno de los principales problemas que enfrentan los métodos de atribución de autoría [6]. En el análisis de las muestras desbalanceadas los resultados no son los mismos que en muestras balanceadas [9], esto se debe a que dependen del número de documentos para extraer las características.

3. Artefactos propuestos

El enfoque propuesto se analiza a partir de la clasificación de textos; que se compone de seis etapas: adquisición de datos, análisis y etiquetado de datos, construcción y ponderación de características, selección o proyección de características, entrenamiento del modelo y evaluación de la solución [10]. Los métodos de clasificación han sido utilizados en los enfoques de instancias y perfil de autor. En la presente investigación la etapa del análisis y etiquetado de datos se realizó considerando la extracción de características de un solo documento, la etapa de selección o proyección fue omitida. Para la etapa de experimentación el conjunto de datos con el que se trabajó fueron los *corpus* C50, C10 y PAN12. Los *corpus* C50 y C10 están compuestos por documentos del *corpus* REUTERS volumen uno, que se encuentra dentro de la categoría CCAT formado de noticias acerca de la industria, por cada autor tiene 50 documentos de entrenamiento y 50 de validación, la diferencia entre el *corpus* C50 y el C10 es el número de autores, en el C50 se consideran 50 autores y en el C10 se consideran 10 autores. El *corpus* PAN12 está compuesto por 14 autores cada uno con tres documentos en idioma ingles del contexto novelas [1, 11].

3.1. Adquisición de datos

En 2009 surge la primera competición internacional de detección de plagio en el marco del taller PAN (CLEF, *Conference and Labs of the Evaluación Forum*). Su objetivo ha sido fomentar el desarrollo de herramientas automáticas para la detección de plagio y actualmente también la identificación de autoría y otros usos abusivos del *software* social [1]. En la página web de PAN (https://pan.webis.de/) se encuentran los *corpus* C50, C10 y PAN12, estos corpus han sido utilizados para la tarea de atribución de autoría en diferentes trabajos de la literatura, el acceso a estos es de dominio público.

3.2. Análisis y etiquetado de datos

Para la extracción de características únicamente se consideraron las características de un solo documento por autor. El análisis del método se realizó con dos modelos de representación que han sido utilizados en trabajos de la literatura [3, 5]: bolsa de palabras y trigramas de carácter. Para la representación con bolsa de palabras se eliminaron los siguientes caracteres especiales; signos de puntuación, signos de admiración e interrogación, y con el modelo de trigramas no se realiza ningún pre procesamiento.

3.3. Construcción y ponderación de características

La ponderación booleana o binaria consiste en asignar un valor a un término dentro de un documento, el valor asignado al término refleja la importancia del mismo, se asigna un valor de uno cuando el término se encuentra presente en caso contrario se asigna un valor de cero. A través de esta ponderación se puede conocer la importancia de cada uno de los términos. Lo anterior se representa a través de la ecuación uno, Donde: t_j representa la frecuencia de término j que tiene la oración p_i [12, 13]:

$$p_i\,(t_j) = \begin{cases} 0, si\ aparece, \\ 1, en\ otro\ caso. \end{cases} \quad (1)$$

La Tabla 1 muestra la representación vectorial utilizada en la presente investigación, donde en una columna se coloca el documento de cada autor y en la fila el termino, posteriormente se va llenando la tabla con la ponderación, donde 1 significa que el término está presente dentro del documento del autor, caso contrario se coloca un cero, este proceso se realiza hasta terminar con todos los documentos de los autores.

Tabla 1. Representación vectorial con ponderación booleana.

Documento/autor	Término 1	Término2	Término n-1	Término n
Documento1 autor1	0	0	1	1
Documento n-1 autor1	0	1	0	0
Documento n autor1	0	0	1	1
Documento n-1 autor n-1	1	1	0	1
Documento n autor n	1	0	0	1

3.4. Entrenamiento del modelo

El modelo se construyó a partir de aprendizaje supervisado implementando máquina de soporte vectorial, los parámetros fueron un *kernel* lineal, el parámetro C igual a uno, utilizando una clasificación de uno contra todos, la máquina de soporte vectorial fue entrenada con las características de cada uno de los autores, una vez generado el modelo se realizó la evaluación.

3.5. Evaluación de la solución

La métrica utilizada para la evaluación fue la exactitud. Esta consiste en el porcentaje de instancias que se clasifican correctamente se define en términos de verdaderos positivos (*True positives*, por sus siglas en inglés TP), Falsos positivos (*False positives*, por sus siglas en inglés FP), Verdaderos negativos (*True Negatives*, por sus siglas en inglés TN) y falsos negativos (*False Negatives*, por sus siglas en inglés FN) como se muestra en la ecuación 2 [14]:

$$exactitud = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

3.6. Experimentación y resultados

Para la comprobación del método propuesto se realizaron 4 experimentos, estos fueron validados mediante la técnica de validación cruzada a diez pliegues, utilizando una máquina de soporte vectorial para la clasificación, estos experimentos muestran la robustez del método, ya que se analiza en primera instancia el método con respecto a los del estado del arte con el *corpus* C10.

En el segundo experimento se observa la robustez del método con cincuenta autores con el *corpus* C50.

En el tercer experimento se analiza la exactitud para la clasificación con diferentes *corpus*, para este experimento se utilizaron los *corpus* C10 y PAN12, en el cuarto experimento se analizan las muestras balanceadas y desbalanceadas con el *corpus* C50.

El primer experimento se realizó con el *corpus* C10, el objetivo de este experimento fue observar la robustez del método analizando una muestra balanceada con diez autores y por cada autor 50 documentos de entrenamiento y 50 de validación.

El método propuesto se compara con uno de los más referenciados del *corpus* C10 el modelo de espacio de tensores [5], y algunos de los métodos más recientes como Doc2vec presentado en [14] y el método de aprendizaje profundo presentado en [15] como se muestra en la figura 1.

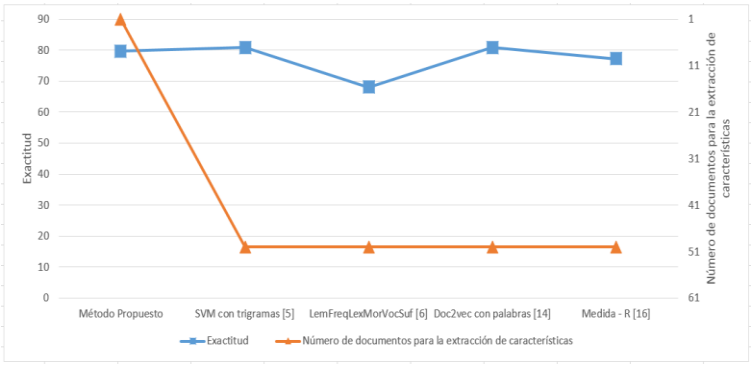

Fig. 1. Exactitud en relación al número de documentos para la extracción de características en el *corpus* C10.

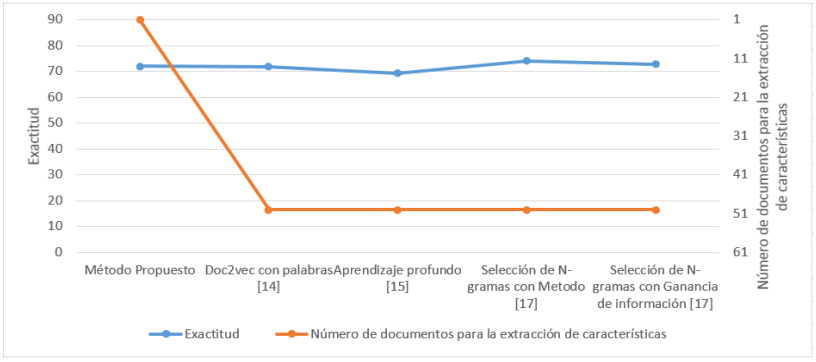

Fig. 2. Exactitud en relación al número de documentos para la extracción de características en el *corpus* C50.

El método propuesto alcanza una exactitud de 79.68 % con una diferencia de 1.12 % del modelo de espacio de tensores presentado en [5], obteniendo mejores resultados que algunas propuestas del estado del arte como se puede observar en la Fig. 1.

El segundo experimento se realiza de la misma manera que el primero, pero con el *corpus* C50, que contiene 50 autores cada uno de ellos con 50 documentos de entrenamiento y 50 de validación a diferencia del C10 que contiene 10 autores, el objetivo de este experimento es conocer si el método obtiene una buena exactitud en la clasificación con cincuenta autores, la exactitud obtenida es de 72.04, un resultado competitivo con respecto al estado del arte como se muestra en la Fig. 2.

En el tercer experimento se analizó el método bajo dos diferentes contextos: correos electrónicos y obras literarias. Se analizó el método propuesto considerando como características a los trigramas y bolsa de palabras. Utilizando el trabajo de [16] para comparar los resultados, en este trabajo reproducen los métodos más referenciados del estado del arte para atribución de autoría realizando la experimentación con los *corpus* C10 y PAN12.

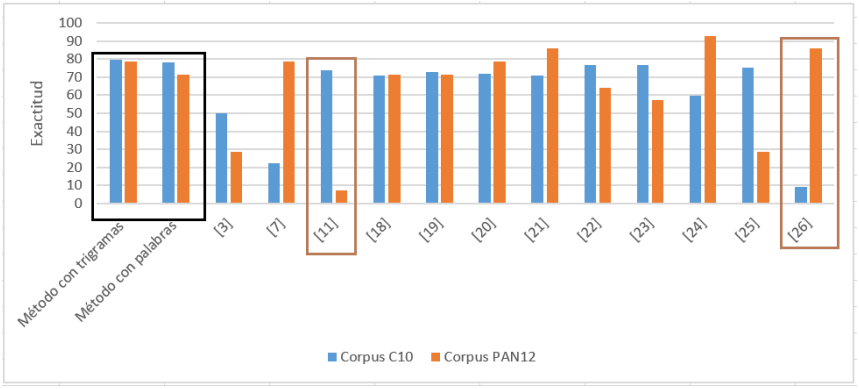

Fig. 3. Evaluación del método bajo diferentes contextos.

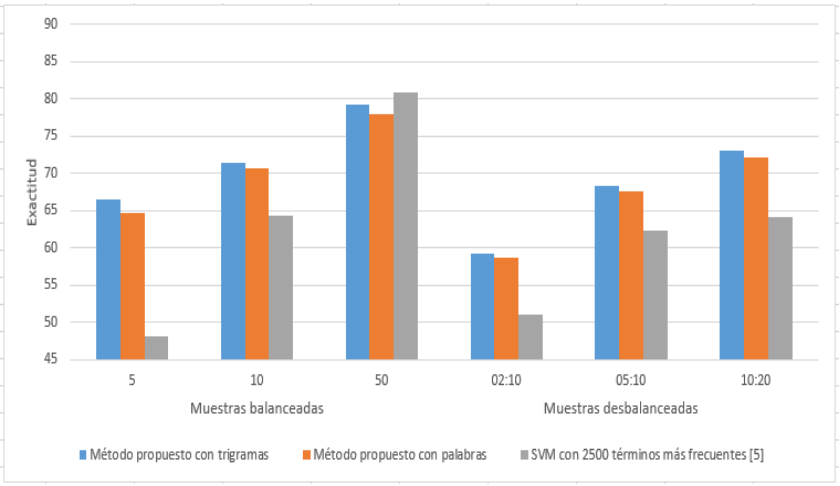

Fig. 4. Análisis de muestras balanceadas y desbalanceadas en el *Corpus* C10.

Se puede observar en la Fig. 3 que el método propuesto se mantiene constante en los dos *corpus* obteniendo resultados superiores al 70%, en comparación a los evaluados por [16] donde los métodos con un *corpus* obtienen buenos resultados y con el otro bajo resultado.

El método propuesto obtiene mejores resultados que algunos de los métodos evaluados por [16]. El análisis de muestras balanceadas y desbalanceadas se aborda principalmente en el estado del arte con el *corpus* C10. Se analiza el método propuesto considerando como características a los trigramas y bolsa de palabras.

Como se puede observar en la Fig. 4, los métodos propuestos obtienen mejores resultados en las muestras desbalanceadas que el propuesto por [5] y en las muestras balanceadas se obtienen resultados competitivos e incluso en las muestras balanceadas de tamaño cinco y diez es mejor, donde se puede determinar que el método propuesto es también robusto con muestras balanceadas y desbalanceadas.

4. Conclusiones y trabajo futuro

El número promedio de palabras por documento en el *corpus* C10 es de 507 con esto se obtiene una exactitud de 79.68%, para el C50 el promedio es de 500 de acuerdo a la experimentación, con estas se obtiene una exactitud de 72.004%, lo anterior puede observarse en la Fig. 1 y Fig. 2 donde el método presenta competitividad con los trabajos revisados en la literatura. A través de esta experimentación se concluye que en un solo documento se encuentra contenido el estilo de escritura de un autor.

La comparativa de resultados entre el *corpus* C10 y PAN12 en diferentes contextos muestra la robustez del método, esto se puede observar en la Fig. 3, donde el método se mantiene consistente en los resultados esto comparado con los métodos analizados en [16]. El análisis de métodos presentado por [16] muestra que algunos métodos presentan variabilidad en sus resultados cuando trabajan con *corpus* de diferentes contextos, un ejemplo de este es el método de [11] donde el resultado con el *corpus* C10 es de 78% y 7.1% de exactitud con PAN12 o el método [26] donde el resultado con el *corpus* C10 es de 9% y 85.7% de exactitud con PAN12.

El análisis de muestras desbalanceadas puede trasladarse a la vida real donde en algunos casos no existe el mismo número de documentos para cada autor, los resultados de la experimentación muestran la superación de los trabajos revisados en la literatura, donde el porcentaje de superación de la muestra 2:10 es de 8.3%, para la muestra 5:10 es de 6% y para 10:20 es de 8.92%, con este tipo de muestras permite asemejar la situaciones de la vida real donde no se cuenta con el mismo número de documentos por cada uno de los autores como se observa en la Fig. 4.

Los resultados presentados muestran que la presente investigación supero a los trabajos revisados en la literatura, esto se puede observar en la Fig. 4. Para las muestras balanceadas el método supera a dos de las tres muestras de tamaño cinco y diez con una diferencia de 18.32% y 7.04%. Sin embargo, en la muestra de tamaño 50 el método propuesto por [5] supera al método con 1.12% como se observa en la Fig. 4. Sin embargo, hay que considerar que utilizamos en promedio 1820 características y el método de [5] utiliza 2500.

Un área de oportunidad que surge a partir de esta investigación es el desarrollo de métodos que puedan determinar la autoría a partir de un solo documento mediante la implementación de métricas de similitud, con una exactitud superior a la de los trabajos de la literatura.

Referencias

1. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3), pp. 538–556 (2009)
2. Anwar, W., Bajwa, I., Ramzan, S.: Design and implementation of a machine learning based authorship identification model. Scientific Programming, pp. 1–14 (2019)
3. De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. In: Association for Computing Machinery Special Interest Group on Management of Data (ACM SIGMOG) 30(4), 55 (2001)
4. Solorio, T., Montes, M., Pillay, S.: Authorship identification with modality specific meta features. In: Conference CLEF Labs and Workshop, Notebook Papers (2011)

5. Plakias, S., Stamatatos, E.: Tensor space models for authorship attribution. In: Proceedings of the Conference Hellenic Conference on Artificial Intelligence. Lecture Notes in computer Science, 5138, pp. 239–249, Springer, Heidelberg (2008)
6. Gómez-Adorno, H., Sidorov, G., Pinto, D., Vilariño, D., Gelbukh, A.: Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. Sensors 16(9), 1374 (2016)
7. Koppel, M., Shler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research 8, pp. 1261–1276 (2007)
8. Pavlyshenko, B.: Genetic optimization of keywords subset in the classification analysis of texts authorship. Journal of Quantitative Linguistics 21, pp. 341–349 (2014)
9. Pastor, A., Montes-y-Gómez, M., Villaseñor-Pineda, L., Ariel, J., Martínez-Trinidad, J.: A new document author representation for authorship attribution. In: Mexican Conference on Pattern Recognition, pp. 283–292 (2012)
10. Mirończuk, M., Protasiewicz, J.: A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106, pp. 36–54 (2018)
11. Stamatatos, E.: Authorship attribution based on feature set subspacing ensembles. International Journal on Artificial Intelligence Tools 15 (05), pp.823–838 (2006)
12. Ledeneva, Y., García, R.: Generación automática de resúmenes: retos, propuestas y experimentos (2017)
13. Zhang, Z., Li, X., Tian, X.: Research on feature weights of Liheci word sense disambiguation. In: 8th International Symposium on Computational Intelligence and Design (ISCID), 2, pp. 7–10 (2015)
14. Posadas-Durán, J., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D., Chanona-Hernández, L.: Application of the distributed document representation in the authorship attribution task for small corpora. Soft Computing, 21, pp. 627–639 (2017)
15. Qian, C., He, T., Zhang, R.: Deep learning based authorship identification (2017)
16. Potthast, M., Braun, S., Buz, T., Duffhauss: Who wrote the web? Revisiting influential author identification research applicable to information retrieval. In: European Conference on Information Retrieval. Lecture Notes in Computer Science, 9626, pp. 393–407, Springer, Cham (2016)
17. Houvardas, J., Stamatatos, E.: N-Gram feature selection for authorship identification. In: Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, Lecture Notes in Computer Science, 4183, pp. 77–86. Springer, Varna (2006)
18. Keselj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING, pp. 255–264 (2003)
19. Benedetto, D., Caglioti, E., Loreto, V.: Language trees and zipping. Physical Review Letters 88(4), 048702 (2002)
20. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Language resources and evaluation 45(1), pp. 83–94 (2010)
21. Stamatatos, E.: Author identification using imbalanced and limited training texts. In: Proceedings International Workshop on Database and Expert Systems, DEXA, pp. 237–241, Regensburg (2007)
22. Khmelev, D.V., Teahan, W.J.: A repetition-based measure for verification of text collections and for text categorization. In: Proceedings Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.104–110 (2003)
23. Peng, F., Schuurmans, D., Wang, S.: Augmenting naive Bayes classifiers with statistical language models. Information Retrieval, 7, pp. 317–345 (2004)

24. Burrows, J.D.: A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing 17(3), pp. 267–287 (2002)
25. 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, pp. 288–298 (2011)
26. Arun, R., Suresh, V., Veni, C.: Stopword graphs and authorship attribution in text corpora. In: International Conference on Semantic Computing, pp. 192–196 (2009)