ISSN 1870-4069

Nuevo método para atribución de autoría en muestras de entrenamiento balanceadas y desbalanceadas del corpus C10

Cesar Alexis Estrada Palacios, José Luis Tapia Fabela

Universidad Autónoma del Estado de México,
México

{kirdrazler,joseluis.fabela}@gmail.com

Resumen. Atribución de Autoría es la tarea de identificar el autor de un texto anónimo. El uso de inteligencia artificial es muy común en esta tarea y uno de los principales problemas es que no se cuenta con suficientes textos de entrenamiento de uno o más candidatos autores, lo que genera un problema de desbalance, afectando en gran medida el rendimiento de los métodos propuestos en el estado del arte. La mayoría de estos métodos utilizan SVM como clasificador y una de las siguientes características como modelo de representación de texto: unigramas de palabra o trigramas de carácter. En este artículo se propone un método con un rendimiento superior a los métodos del estado del arte en muestras balanceadas y desbalanceadas, esto se logra mediante la combinación de distintos tamaños de n-gramas para formar la bolsa de n-gramas y generar un nuevo modelo de representación de texto. Los experimentos realizados muestran la importancia de entrenar al clasificador con una bolsa de n-gramas de diferentes tamaños y no solo con un tamaño fijo. Los mejores resultados se obtienen cuando se combinan unigramas y bigramas de palabra, además se muestra la importancia que tiene la estrategia de clasificación cuando se utiliza SVM.

Palabras clave: Atribución de autoría, representación de texto, estrategia de clasificación, SVM, bolsa de n-gramas.

New Method for Authorship Attribution in Balanced and Unbalanced Training Samples from the C10 Corpus

Abstract. Authorship Attribution is the task of identifying the author of an anonymous text. The use of artificial intelligence is very common in this task, and one of the main problems is that there are often not enough training texts from one or more candidate authors this generates an imbalance problem, greatly affecting the performance of the methods proposed in the state of the art. Most of these methods use SVM as a classifier and one of the following features as a

text representation model: word unigrams or character trigrams. This article proposes a method with superior performance to the methods of the state of the art in balanced and unbalanced samples, this is achieved by combining different sizes of n-grams to form the bag of n-grams and generate a new model. of text representation. The experiments carried out show the importance of training the classifier with a bag of n-grams of different sizes and not only with a fixed size. The best results are obtained when word unigrams and bigrams are combined, and the importance of the classification strategy when using SVM.

Keywords: Authorship attribution, text representation, classification strategy, SVM, bag of n-grams.

1. Introducción

La tarea de atribución de autoría consiste en identificar el autor de un texto del cual se desconoce su autoría a partir de un conjunto de posibles autores [1, 2]. Actualmente hay una gran cantidad de información alojada en la web en forma de texto, como e-mails, mensajes en foros, blogs, código fuente, entre otros [1, 3]; a partir de esta información surgen problemas como ciber bullying, plagio, spam, fraude, etc. En estos casos la atribución de autoría juega un papel importante [5]. Normalmente se utilizan métodos de inteligencia artificial para identificar el autor de dichos textos, estos métodos necesitan de textos de referencia para su entrenamiento, el problema es que a menudo se cuentan con pocos textos sobre un autor candidato, lo que genera un problema de desbalance, o en general se tiene pocos documentos de todos los autores lo cual afecta drásticamente el rendimiento de los métodos de aprendizaje supervisado [6].

Se han generado propuestas que funcionen en distintos escenarios de balance y desbalance, por ejemplo, en [1] se proponen varios métodos, donde se concluye que al utilizar 2500 características en promedio para la representación de texto se logran buenos resultados. En [2] se compara el rendimiento del clasificador SVM y STM, siendo superior SVM, teniendo buenos resultados con muestras balanceadas. Por otra parte [3] consigue buenos resultados en muestras desbalanceadas a través de una nueva representación de texto.

En ninguno de los casos anteriores hay un método que sea superior a los demás, ya que solo tienen un buen rendimiento en escenarios muy específicos, es por esto que el objetivo de este trabajo es desarrollar un método que tenga un rendimiento superior a los métodos del estado del arte tanto con muestras balanceadas como desbalanceadas.

Como primer características del método, se propone utilizar una bolsa de n-gramas conformada por distintos tamaños de n-gramas de palabra, con la finalidad de que el clasificador no solo aprenda de un tipo de característica, esto se basa en la observación de los autores [4] donde se utilizan los 100 n-gramas de palabra más frecuentes de tamaño 1, 2 y 3 (300 en total), obteniendo resultados competentes en comparación con los 2500 utilizados en los trabajos de [1–3]. Con base en los trabajos anteriormente mencionados se propone formar una bolsa de n-gramas a partir de los 2500 más frecuentes probando con distintos tamaños.

Como segunda característica del método y debido a los buenos resultados mostrados por el clasificador SVM en trabajos del estado del arte sobre autoría, se pone a prueba distintas variantes de SVM y a diferencia de los trabajos mencionados anteriormente, comparamos la estrategia de clasificación utilizada por el clasificador; ya que los resultados muestran que es uno de los factores que más influye en el proceso de clasificación.

Para evaluar esta idea, se presentan una serie de experimentos con muestras de entrenamiento balanceadas y desbalanceadas propuestas por [2] basadas en el corpus C10.

Los resultados muestran que utilizar una bolsa de n-gramas con diferentes tamaños de n-grama mejora los resultados obtenidos por el clasificador. El mejor resultado se obtuvo cuando se combinan palabras y bi-gramas de palabra. Además, la estrategia de clasificación uno contra todos obtiene los mejores resultados. Las distintas variantes de SVM juegan un papel mínimo en el resultado de la clasificación; sin embargo, los resultados muestran que SVM Dual es la mejor opción.

2. Corpus y método

2.1. Corpus

Para evaluar el método propuesto, se utiliza el conjunto de datos C10 [2], el cual es un subconjunto de Reuters Corpus Volumen 1 [5] contempla a los 10 autores con mayor cantidad de documentos relacionados al tema de noticias corporativas e industriales; de acuerdo con el autor, con la finalidad de que el tema no sea un factor para distinguir entre los autores. Cada autor cuenta con 50 documentos de *train* y 50 de *test*.

A partir del corpus C10, se experimenta en 6 posibles escenarios de entrenamiento, 3 escenarios balanceados donde se utilizan 50, 10 y 5 textos de entrenamiento por autor y 3 escenarios desbalanceados en donde el número de textos de entrenamiento por autor varia en un rango de 2:10, 5:10 y 10:20; considerando el caso 2:10 se pueden utilizar entre 2 a 10 documentos por autor.

Estos escenarios fueron propuestos anteriormente por [2] para evaluar el comportamiento de su método simulando distintos escenarios realistas donde puede existir un desbalance o se cuenta con muy pocos documentos por autor.

La Figura 1 y Figura 2 son ejemplos de una distribución de muestra balanceada y desbalanceada respectivamente.

2.2. Método

El método propuesto sigue las etapas que normalmente se siguen en la clasificación de textos, ya que atribución de autoría puede ser vista como una tarea de clasificación de textos [2].

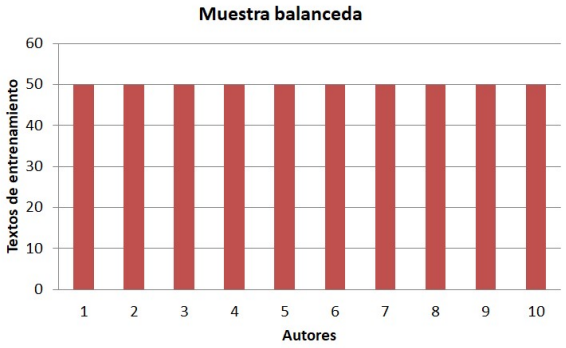


Fig. 1. Distribución de la muestra balanceada con 50 datos de entrenamiento.

Fig. 2. Distribución de la muestra desbalanceada para el caso 2:10.

En general, el proceso de clasificación de textos según [1] consiste en las siguientes etapas.

1. Adquisición de datos: El método propuesto necesita de un train y test, debido a que el aprendizaje utilizado es del tipo supervisado; porque utiliza SVM como algoritmo de clasificación. Con fines comparativos se utilizó el corpus C10, descrito en la sección 2.1.

2. Análisis y etiquetado de los datos: El corpus C10 originalmente está dividido en train y test, el etiquetado de datos dependerá del enfoque elegido, ya sea basado en instancias o basado en perfil [7], la elección del enfoque dependerá: del tipo de desbalance que se tenga, del clasificador a utilizar y del costo de entrenamiento, entre otros factores. El método propuesto utiliza el enfoque basado en instancias; porque es el recomendado cuando el desbalance es producido por la cantidad de textos de entrenamiento disponibles y cuando se utiliza un algoritmo de aprendizaje, en este caso SVM.

Fig. 3. Proceso general de un método de clasificación de textos basado en (Mirończuk & Protasiewicz, 2018).

3. Construcción y pesado de características: Para poder construir las características deseadas, en este caso palabras y bigramas de palabra, es necesario realizar primero un preprocesamiento; el cual consiste en dar un formato adecuado al texto, eliminando la información que no se necesita. Existen diversas formas de llevar a cabo el preprocesamiento, específicamente el método propuesto utiliza las siguientes: eliminación de saltos de línea, números y signos de puntuación; dejando solo los caracteres que son letras. A partir de este punto surge la posibilidad de utilizar stemming o no, en la etapa de experimentación se prueba su efectividad. De cada texto preprocesado, se obtienen los n-gramas de palabra, los cuales conforman la bolsa de n-gramas. Se proponen tres opciones a explorar, unigramas + bigramas, unigramas + trigramas y unigramas + bigramas + trigramas; lo anterior, con el fin de averiguar si al combinar n-gramas mejora los resultados y cuál es la mejor combinación. Como siguiente etapa se obtiene el pesado de características, también conocido como pesado de términos, como técnica se utilizó pesado binario, esto quiere decir que se asigna el valor de 1 si el término aparece en el texto y 0 si no aparece.

4. Selección de características: Consiste en seleccionar las características que utiliza el clasificador. De acuerdo con [1–3] en promedio 2500 características son necesarias para obtener buenos resultados; por lo tanto, para entrenar al clasificador se propone utilizar las 2500 características más frecuentes.

5. Entrenamiento del modelo de clasificación: El modelo utilizado en la mayoría de los trabajos de autoría es SVM; debido a sus buenos resultados [2, 3]. Existen diferentes variantes de SVM, uno de los más conocidos; además de utilizado por los autores

Tabla 1. Comparación de rendimiento entre diferentes tamaños de n-gramas de palabra.

Característica	**5**	**10**	**50**	**02:10**	**05:10**	**10:20**
Unigramas	**67.74**	**72.07**	79.60	**60.87**	**68.94**	**73.58**
Bigramas	62.41	69.34	**80.40**	56.56	65.87	71.38
Trigramas	49.15	56.53	68.40	47.78	52.97	60.02

Tabla 2. Comparación de la combinación de las bolsas n-gramas.

Característica	**5**	**10**	**50**	**02:10**	**05:10**	**10:20**
Unigramas+bigramas	**68.05**	73.58	**83.60**	60.89	**69.79**	**74.97**
Unigramas+trigramas	67.30	72.49	80.40	**61.48**	69.25	73.93
Bigramas+trigramas	60.49	69.27	77.20	56.20	63.57	70.78
Unigramas+bigramas +trigramas	67.73	**73.60**	82.20	61.00	69.48	74.56

Tabla 3. Aplicación de stemming en el modelo de n-gramas combinado.

Característica	**5**	**10**	**50**	**02:10**	**05:10**	**10:20**
Unigramas+bigramas	**68.05**	**73.58**	**83.60**	**60.89**	**69.79**	**74.97**
Unigramas+bigramas+stemming	67.63	72.67	81.60	60.72	69.05	74.16

Tabla 4. Comparación del rendimiento de las variantes de SVM.

Clasificador	**5**	**10**	**50**	**02:10**	**05:10**	**10:20**
SVM Linear	66.89	73.33	83.40	60.06	68.98	74.60
SVM Primal	**68.38**	73.00	82.20	60.79	69.67	74.49
SVM Dual	68.05	**73.58**	**83.60**	**60.89**	**69.79**	**74.97**

mencionados anteriormente es SVM Lineal. Se experimentó con la versión LibSVM Linear; SVM L2 Primal y L2 SVM Dual logrando los mejores resultados con SVM Dual. Cuando se presentan problemas multiclase SVM necesita de una estrategia de clasificación. En nuestro caso de estudio, el corpus utilizado consta de 10 autores, por lo tanto, de 10 clases; por esto es necesario utilizar una estrategia de clasificación, las más comunes son la estrategia uno contra uno y uno contra todos.

6. Con fines comparativos se utilizó la métrica *accuracy*, debido a que es la métrica utilizada en el estado del arte por diversos autores para evaluar el corpus C10 [4, 8, 9].

3. Experimentos

En esta sección se evalúa el rendimiento del método propuesto, se presenta detalle en que consistió cada experimento, además de analizar los resultados obtenidos. En los siguientes experimentos se ponen a prueba diferentes tamaños de n-gramas, combinación de n-gramas, stemming, variantes de SVM y estrategias de clasificación. Cada experimento se evalúa en 6 escenarios diferentes, 3 con muestras de entrenamiento balanceadas y 3 desbalanceadas, estas muestras se detallan en la sección de corpus.

Con fines estadísticos y de réplica, cada resultado mostrado en cada experimento es el promedio de 100 corridas y no solo 10 corridas como se realizó en los experimentos de [3] permitiéndonos mostrar resultados más confiables y precisos.

3.1. Modelo de n-gramas simple

En este experimento se compara el rendimiento del método propuesto cuando se utiliza n-gramas de palabra como modelo de representación de texto. En la tabla 1 se compara el rendimiento entre unigramas, bigramas y trigramas de palabra.

El mejor resultado se obtiene al utilizar unigramas de palabras, que son básicamente palabras. Se puede observar que entre más grande es el n-grama peores resultados se obtienen tanto en muestras balanceadas como desbalanceadas, por lo que no es necesario experimentar con un tamaño de n-grama mayor.

3.2. Modelo bolsa de n-gramas compuesta

En este experimento se crea una bolsa de n-gramas a partir de n-gramas de palabra de diferentes tamaños, en la tabla 2 se muestran las diferentes combinaciones posibles entre unigramas, bigramas y trigramas.

Como se muestra en la tabla 2, los mejores resultados se obtienen a partir de la combinación de unigramas y bigramas de palabra. Además, ocurre un comportamiento semejante al experimento anterior, entre más pequeños sean los n-gramas a combinar mejores resultados se obtienen. Se observa también que la estrategia de clasificación influye en gran medida en el resultado obtenido.

Los resultados muestran que la representación de texto del método propuesto mejora los resultados en todas las pruebas realizadas, especialmente en la muestra 50, donde se obtiene casi 4% de diferencia respecto a las demás formas de representación de texto.

3.3. Stemming

Con el fin de saber si aplicar stemming mejora los resultados, en este experimento se evalúa el efecto de aplicar stemming al modelo de bolsa de n-gramas combinado unigramas + bigramas de palabra ya que fue la combinación que obtuvo mejores resultados en el experimento anterior. Los resultados muestran que utilizar stemming tiene un efecto negativo con todas las muestras de entrenamiento evaluadas y en ningún caso se logró una mejora. También se observa que entre más documentos de

entrenamiento se tengan para el entrenamiento peores resultados se obtendrán si se aplica stemming, es por ello que existe una mayor diferencia en los resultados de la muestra balanceada 50.

3.4. Clasificador

El clasificador SVM ha sido utilizado en cada uno de los trabajos con los cuales comparamos el método propuesto [2, 3, 10]; sin embargo, ninguno de esos trabajos ha evaluado las variantes de este clasificador, por lo tanto, los siguientes dos experimentos se realizaron con la intención de conocer cual variante de SVM y estrategia de clasificación logra obtener los mejores resultados.

Variantes de SVM. En este experimento, se compara el rendimiento de las variantes de SVM (Linear, l2 Primal y l2 Dual) para conocer cual tiene un mejor rendimiento en la clasificación. Los parámetros utilizados fueron c= 1 mismo que fue utilizado por [2] y el valor de épsilon = 0.001 estos dos parámetros aplican para las 3 variantes de SVM ya mencionadas. Como modelo de representación texto, se utiliza el modelo combinado de unigramas + bigramas de palabra sin aplicar stemming.

La variante del clasificador con mejor rendimiento fue SVM Dual, logrando los mejores resultados en 5 de 6 muestras; solo fue superado en la muestra balanceada con 5 documento por una diferencia mínima. SVM Primal obtuvo un rendimiento inferior en la mayoría de las pruebas en comparación de SVM Dual; sin embargo, fue superior a SVM Linear en la mayoría de las pruebas, con excepción de la muestra 50. SVM Linear no logró obtener el mejor resultado en ninguna de las muestras.

Estrategia de clasificación. Usando el resultado del experimento anterior, donde se comprobó que usando SVM Dual se obtiene los mejores resultados en la clasificación, se evalúa ahora la estrategia de clasificación: uno contra uno OvO y uno contra todos OvA, para saber cuál estrategia ayuda más al clasificador.

La estrategia de clasificación uno contra todos es superior en todos los casos a la estrategia uno contra uno, en todas las muestras se tiene una mejora considerable, en especial en las muestras donde existe un mayor desbalance y se cuenta con el menor número de documentos. Por lo anterior podemos concluir que la estrategia de clasificación es el factor que más influye en el clasificador cuando existe desbalance.

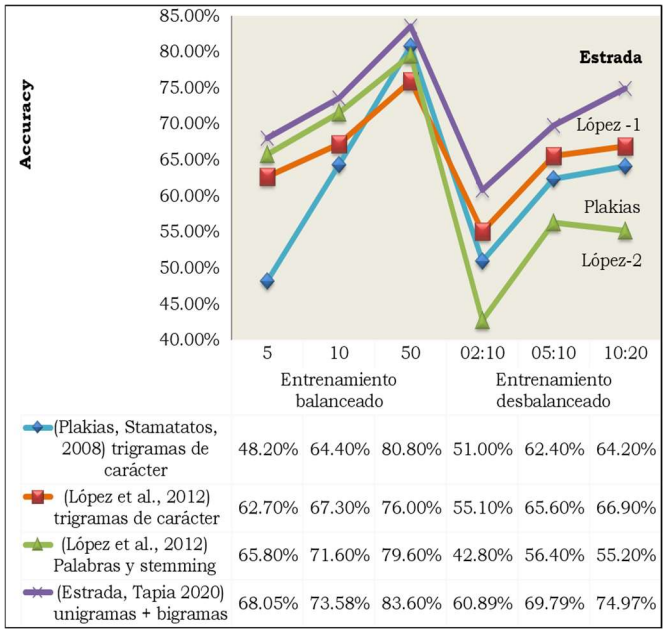
4. Discusión

Con base en los resultados obtenidos en la etapa de experimentación, se observa que la estrategia de clasificación es el factor que más influye en el resultado cuando hay pocos documentos disponibles y existe un desbalance. Combinar n-gramas mejora ligeramente los resultados en comparación con no combinar n-gramas. Entre más documentos se tengan mayor ventaja tiene el combinar n-gramas.

A pesar de que SVM Dual es superior en todas las pruebas a SVM Linear y SVM Primal, la diferencia es mínima, por lo que la variante del clasificador no tiene un gran impacto sobre el resultado obtenido. En cuanto al stemming, este produce un impacto negativo en los resultados obtenidos mediante el método propuesto.

Tabla 5. Comparación de las estrategias de clasificación uno contra uno y uno contra todos.

Característica	**Tamaño**	**OvO**	**OvA**	**Diferencia**
Unigramas + bigramas	5	65.47	*68.05*	2.58
	10	71.16	*73.58*	2.42
	50	79.60	*83.60*	4.00
	2:10	52.06	*60.89*	8.83
	5:10	64.53	*69.79*	5.26
	10:20	71.53	*74.97*	3.44
Unigramas + trigramas	5	64.11	*67.30*	3.19
	10	71.32	*72.49*	1.17
	50	77.00	*80.40*	3.40
	2:10	52.33	*61.48*	9.15
	5:10	65.05	*69.25*	4.20
	10:20	71.84	*73.93*	2.09
Bigramas + trigramas	5	57.18	*60.49*	3.31
	10	66.20	*69.27*	3.07
	50	74.60	*77.20*	2.60
	2:10	49.58	*56.20*	6.62
	5:10	58.56	*63.57*	5.01
	10:20	66.92	*70.78*	3.86
Unigramas + bigramas + trigramas	5	64.59	*67.73*	3.14
	10	70.90	*73.60*	2.70
	50	78.40	*82.20*	3.80
	2:10	52.25	*61.00*	8.75
	5:10	64.34	*69.48*	5.14
	10:20	71.75	*74.56*	2.81



	5	10	50	02:10	05:10	10:20
(Plakias, Stamatatos, 2008) trigramas de carácter	48.20%	64.40%	80.80%	51.00%	62.40%	64.20%
(López et al., 2012) trigramas de carácter	62.70%	67.30%	76.00%	55.10%	65.60%	66.90%
(López et al., 2012) Palabras y stemming	65.80%	71.60%	79.60%	42.80%	56.40%	55.20%
(Estrada, Tapia 2020) unigramas + bigramas	68.05%	73.58%	83.60%	60.89%	69.79%	74.97%

Fig. 4. Comparación del método propuesto por Estrada con el estado del arte.

En la figura 4 se compara el mejor resultado obtenido a partir del método propuesto (Estrada) respecto al estado del arte en las muestras balanceadas y desbalanceadas del corpus C10.

Los resultados muestran que el método propuesto, supera a los métodos presentados en el estado del arte en todas las muestras, balanceadas y desbalanceadas, principalmente existe una mayor diferencia cuando se entrena con muestras desbalanceadas, siendo estas las que mayor se presentan en un escenario realista, además el método propuesto es robusto, debido a que mantiene un mejor rendimiento cuando se tiene menos documentos de entrenamiento o existe un desbalance.

5. Conclusiones

Se puede hacer las siguientes conclusiones:

1. Combinar distintos tamaños de n-gramas de palabras como modelo de representación de texto beneficia al clasificador logrando una mejor accuracy tanto en muestras balanceadas como desbalanceadas.
2. Utilizar la combinación unigramas+bigramas de palabras como modelo de representación de texto produce los mejores resultados. Estos resultados son superiores a los expuestos por [2, 3] en los 6 diferentes escenarios.
3. El factor que más influye en el accuracy del método es la estrategia de clasificación.

4. En términos de accuracy la estrategia de clasificación uno contra todos es muy superior a la estrategia uno contra uno, pues se obtienen resultados superiores con todos los modelos de representación de texto evaluados en la etapa de experimentación, con excepción de trigramas de carácter en la muestra de 50.
5. Al igual que el estado del arte, el método propuesto disminuye su rendimiento cuando se tienen menos documentos o es mayor el desbalance.
6. Usar palabras es más robusto que utilizar n-gramas de carácter.
7. Utilizar stemming empeora los resultados con el modelo de representación de texto propuesto.

Referencias

1. Stamatatos, E.: Author identification using imbalanced and limited training texts. In: 18th International Conference on Database and Expert Systems Applications (DEXA 2007). IEEE, Regensburg, Germany, pp. 237–241 (2007) doi: 10.1109/DEXA.2007.5
2. Plakias, S., Stamatatos, E.: Tensor space models for authorship identification. Darzentas, J., Vouros, G. A., Vosinakis, S., and Arnellos, A. (eds.). Artificial intelligence: theories, models and applications. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 239–249 (2008) doi: 10.1007/978-3-540-87881-0_22
3. López-Monroy, A. P., Montes-y-Gómez, M., Villaseñor-Pineda, L., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F.: A New Document author representation for authorship attribution. Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., Olvera López, J. A., Boyer, K. L. (eds.) Pattern Recognition. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 283–292 (2012) doi: 10.1007/978-3-642-31149-9_29
4. Sari, Y., Stevenson, M., Vlachos, A.: Topic or style? exploring the most useful features for authorship attribution. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA pp. 343–353 (2018)
5. Lewis, D. D., Yang, Y., Rose, T. G., Li, F.: RCV1: A new benchmark collection for text categorization research. 37 (2004)
6. Mirończuk, M. M., Protasiewicz, J.: A recent overview of the state-of-the-art elements of text classification. Expert Syst. Appl., vol. 106, pp. 36–54 (2018) doi: 10.1016/j.eswa. 2018.03.058
7. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol., vol. 60, pp. 538–556 (2009) doi: 10.1002/asi.21001
8. Popescu, M., Grozea, C.: Kernel methods and string Kernels for authorship analysis. CLEF (2012)
9. Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J. M., Köhler, J., Lötzsch, W., Müller, F., Müller, M. E., Paßmann, et al: Who wrote the web? Revisiting influential author identification research applicable to information retrieval. Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Di Nunzio, G. M., Hauff, C., Silvello, G.

(eds.) Advances in Information Retrieval, pp. 393–407, Springer International Publishing, Cham (2016) doi: 10.1007/978-3-319-30671-1_29

10. Escalante, H. J., Solorio, T., Montes y Gomez, M.: Local histograms of character n-grams for authorship attribution. 11 (2011)