ISSN 2007-9737

Verificación de autoría, clasificación por vecindad

Daniel Castro[1], Yaritza Adame[1], María Pelaez[1], Rafael Muñoz[2]

[1] Desarrollo de Aplicaciones, Tecnología y Sistemas (DATYS), Santiago de Cuba, Cuba

[2] Universidad de Alicante, España, Departamento de Lenguajes y Sistemas Informáticos, España

daniel.castro@cerpamid.co.cu, {yaritza.adame, maria.pelaez}@datys.cu, rafael@dlsi.ua.es

Resumen. El análisis de autoría se ha convertido en una herramienta determinante para el análisis de documentos digitales en las ciencias forenses. Proponemos un método de Verificación de Autoría mediante el análisis de las semejanzas entre documentos de un autor por vecindad, sin estimar umbrales a partir de un entrenamiento, implementamos dos estrategias de representación de los documentos de un autor, una basada en instancias y otra en el cálculo del centroide. Evaluamos colecciones según el número de muestras, los géneros textuales y el tema abordado. Realizamos un análisis del aporte de cada función de comparación y de cada rasgo empleado así como una combinación por mayoría de los votos de cada par función-rasgo empleado en la semejanza entre documentos. Las pruebas se realizaron usando las colecciones públicas de las competencias PAN 2014 y 2015. Los resultados obtenidos son prometedores y nos permiten evaluar nuestra propuesta y la identificación del trabajo futuro a desarrollar.

Palabras clave. Análisis de autoría, verificación de autoría, funciones de comparación, rasgos lingüísticos.

Authorship Verification, Neighborhood-based Classification

Abstract. The Authorship Analysis task has become a determining tool for the analysis of digital documents in forensic sciences. We propose a neighborhood classification method of Authorship Verification analyzing the similarities of a document of unknown authorship between samples documents of one author, without estimating parameters values from a training data, we implemented two strategies of representation of the documents of an author, an instance based and a profile based one. We will evaluate the methods in different data collections according to the number of samples, the textual genres and the topic addressed. We perform an analysis of the contribution of each function of comparison and each feature used to take as final decision a combination by majority of the votes of each function-feature pair used in the similarity between documents. The tests were carried out using the public data sets of the Authorship Verification PAN 2014 and 2015 competitions. The results obtained are promising and allow us to evaluate our proposal and the identification of future work to be developed.

Keywords. Authorship detection, author identification, similarity measures, linguistic features.

1. Análisis de autoría

El mundo actual está matizado por grandes avances tecnológicos que abarcan casi todas las esferas de la sociedad. Un ejemplo de esto, es el desarrollo de las tecnologías de la información, donde desempeña un papel importante internet, el cual rápidamente se ha convertido en la principal forma de intercambio de información, permitiendo la comunicación casi en tiempo real, sin tener en cuenta la distancia. La mayor parte de esta información se encuentra almacenada en forma textual no estructurada y escrita en diferentes idiomas, posibilitando que muchos documentos digitales puedan servir de fuentes de consulta. Esta disponibilidad de información conlleva a que muchas veces las personas para un bienestar propio incurran en abusos, como es el caso de la apropiación del conocimiento. Estos “abusos” de la información constituyen un robo de material intelectual [11,14].

En las ciencias forenses, cada día aumenta la necesidad del empleo de métodos computacionales que humanicen y aligeren el trabajo desarrollado por los peritos. El análisis documental es una de las disciplinas que tradicionalmente presenta, entre sus esferas de investigación, la construcción e identificación de perfiles de autores y, más en detalle, la identificación de autoría de documentos sospechosos. Desde sus inicios y aún en la actualidad, se analizan los rasgos caligráficos en los textos manuscritos.

A partir del auge de la digitalización de la sociedad, se comienzan a presentar investigaciones en las que es necesario identificar los rasgos de autores de documentos digitales, aprovechando para esto el creciente desarrollo de métodos de Inteligencia Artificial (IA), que involucran algoritmos de áreas del Procesamiento del Lenguaje Natural (PLN), la Minería de Textos (MT), el Reconocimiento de Patrones (RP), entre otros.

La comunidad científica, fundamentalmente a partir de la década de los 90, dedica esfuerzos crecientes a la investigación y desarrollo de métodos y algoritmos en la tarea de Análisis de Autoría (AA), profundizando en diferentes subtareas como por ejemplo: el Agrupamiento de muestras de autores, la Detección de Plagio, la Detección y Verificación de Autoría, entre otros [8, 9, 14, 16].

Un impulso importante en las investigaciones y en el desarrollo de algoritmos de AA se logra a partir de la plataforma de experimentación y colaboración PAN[1], principalmente en las ediciones que han tenido lugar desde el 2012 hasta la actualidad [2, 3, 12].

Las principales etapas del desarrollo de un sistema computacional [15] se basan en la siguiente metodología:

- **Formulación** del problema no matemático, es decir, el problema que se quiere resolver.
- **Formalización** del problema, es decir, creación del problema matemático.
- **Selección** de la forma de solución del problema.
- **Solución** del problema matemático.
- **Análisis e interpretación** de los resultados, respecto al problema no matemático original que se quiere resolver.

La mayoría de los trabajos consultados sobre la tarea de Análisis de Autoría, dedican los esfuerzos a las etapas de **Formalización**, **Selección** y **Solución**. Sin embargo, pocos parten, o no lo publican, de un análisis de las características en situaciones reales y la solución dada. Por supuesto, una de las complejidades radica en la obtención y luego publicación de colecciones reales de problemas a resolver. En este sentido cabe destacar nuevamente la plataforma de prueba e intercambio para las investigaciones en este tema, que se brinda en las competencias PAN. Los organizadores se esfuerzan por proporcionar colecciones variadas tanto en género textual y longitud de los textos como en temas abordados; y esto constituye un recurso y una oportunidad de incalculable valor para experimentar y desarrollar aproximaciones.

Los principales esfuerzos en las investigaciones de análisis de autoría se han centrado en las etapas de [4,14]:

1. Selección de rasgos y características de la redacción: captura el estilo y los patrones de redacción que lo identifican y diferencian del resto de los autores. Si solo se cuenta con muestras del autor en análisis (más desafiante y complejo), pues no se obtendrían características que lo diferencien.

2. Representación computacional del estilo de redacción: elemento este de suma importancia, pues impone o canaliza la riqueza de información y rasgos que se almacenan.

3. Método de aprendizaje para la clasificación e identificación de autor, empleando un clasificador binario: es la etapa en la que se toma la decisión sobre la autoría de un documento sospechoso o anónimo, se respondería la pregunta ¿Es o no redactado por el autor?

A modo de resumen, las características principales de los trabajos presentados en las ediciones de PAN del 2012 al 2015, y recogidas en los resúmenes de los organizadores, son:

[1] http://pan.webis.de/, PAN is a series of scientific events and shared tasks on digital text forensics.

1. Rasgos y características de redacción

La mayoría de las aproximaciones utiliza algunos de los rasgos expuestos por [14], donde se plantean agrupados en diferentes capas o niveles de análisis del contenido escrito. Niveles de análisis de Caracteres, Léxicos, Sintácticos, Semánticos y específicos de un Dominio de Aplicación. De estos niveles, se analizan y usan con mayor sistematicidad los rasgos léxicos y de caracteres, debido a la facilidad y disponibilidad de herramientas de Procesamiento de Lenguaje para varios idiomas; de ahí, la generalidad de las soluciones. Además, según los resultados experimentales y las consideraciones de los autores de los trabajos, se han obtenido buenos resultados con estos, sin embargo la incorporación de otros rasgos sintácticos y semánticos no aporta significativos aumentos de precisión.

2. Representaciones computacionales

La propuesta más abordada y utilizada se corresponde con la Bolsa de Palabras (del inglés Bag of Words), de manera general es un n-uplo de rasgos lingüísticos y de contenido [14]. Se han presentado aproximaciones haciendo uso de representaciones con grafos, pero estas son las más escazas [5].

Otro elemento a considerar es el espacio de representación de las muestras de cada autor, en este escenario se han presentado trabajos orientados al análisis de cada una de las instancias (instance based) o documentos o a la construcción de representantes de autores (profile based) [14].

3. Métodos de clasificación y decisión

Los enfoques han sido, de manera general, distribuidos en dos grupos, aquellos considerados perezosos (lazy) o de poco esfuerzo y los del grupo de algoritmos con esfuerzo (eager), siendo los primeros los que menos parámetros ajustan o que basan su análisis considerando únicamente los datos que se ofrecen a clasificar sin entrenamiento, y los segundos los que necesitan de muestras recogidas con anterioridad o entrenamiento para el ajuste de los modelos [2, 3, 14, 16].

Los trabajos presentados utilizan en gran medida métodos de clasificación basados en máquinas de vectores soporte (SVM), árboles de decisión, métodos probabilísticos, estrategias de vecindad y una buena parte emplean métodos de combinación de varios clasificadores homogéneos o heterogéneos.

Los clasificadores basados en instancias responden sorpresivamente bien en dominios de clasificación de documentos [14] y el AA puede considerarse una sub-tarea de la clasificación de documentos, en la que se debe hacer especial énfasis en las etapas de la representación de los documentos y la identificación de los rasgos.

Resaltamos entre otros, los trabajos [8, 9, 12], los que presentan estrategias de clasificación basadas en instancias a partir de la vecindad de los objetos de la clase. Nosotros presentamos una implementación basada en instancias y vecindad con similitudes a las expuestas en los trabajos mencionados, con la diferencia que evaluamos tres variantes de decisión de pertenencia a partir de umbrales calculados por la semejanza intra clase de los documentos. Proponemos estudiar una estrategia basada en instancias, como una basada en prototipo. Una diferencia importante que evaluamos es la combinación de varios clasificadores, utilizando más de una función de comparación, permitiendo que la combinación no sea sensible a la eficacia de una sola función de comparación cuando se reciben documentos de diferente tema, género textual o tamaño.

1.1. Motivación de nuestro trabajo

Un caso de estudio práctico en las ciencias forenses se manifiesta cuando el perito debe evaluar la autoría de un documento desconocido y solo cuenta con muestras certificadas de un autor. Ante esto deberá: responder si fue redactado o no por el consiguiente autor, abstenerse o definir en qué grado pudo ser redactado, entre otros elementos, atendiendo a la semejanza con las muestras conocidas. Este caso de estudio se corresponde con las investigaciones realizadas en la Verificación de Autoría (VA).

En nuestra propuesta evaluaremos los siguientes aspectos:

1. Utilizar un método de clasificación basado en el promedio de semejanza entre objetos de un grupo, sin necesidad de ajustar parámetros para la comparación y decisión de la clasificación de un documento de autoría desconocida. Estudiar la semejanza del documento desconocido con respecto a las muestras del autor y determinar el mecanismo de pertenencia al grupo.
2. Método de clasificación calculando el centroide entre objetos de un grupo. Estudiar la semejanza del documento desconocido con respecto al centroide y a las muestras del autor y determinar mecanismo de pertenencia al grupo.
3. Estudiar la efectividad de la clasificación para las diferentes respuestas que se esperan, siendo estas: Sí redactado, No redactado o Abstención.
4. Evaluar en colecciones escritas en idioma español variando el número de muestras de un autor.
5. Evaluar con colecciones cuando varían la homogeneidad en cuanto a los géneros textuales.
6. Evaluar el impacto del uso de cada función de comparación y rasgo empleado.

2. Métodos implementados

El problema que nos proponemos evaluar se corresponde con una tarea de Verificación de Autoría, donde implementamos un método que determina la autoría de un documento desconocido usando una estrategia Intrínseca (donde solo se cuenta con muestras de un autor), con rasgos de los presentados en la literatura a partir de un análisis de caracteres, léxico y sintaxis; emplearemos una aproximación basada en Instancias y otra basada en Representantes, que no dependa de realizar la construcción de un modelo entrenado o la calibración de umbrales con colecciones de entrenamiento.

Proponemos para esto dos algoritmos apoyados en el cálculo de la semejanza entre pares de objetos, definiendo una función de comparación y estableciendo una representación vectorial de los documentos a partir de un tipo de rasgo escogido [10].

Específicamente, restringimos el dominio de aplicación a un entorno donde solo se cuenta con documentos de muestra de un autor (una clase) y dado un documento desconocido, debemos responder si fue redactado por este autor, no redactado o abstenerse. Nos queda definir bajo qué criterios un objeto nuevo pertenece o no la clase, ya sea usando un algoritmo por promedio o uno por el centroide.

De manera formal definimos los siguientes elementos:

Autor o Grupo: conjunto de documentos redactados por una sola persona (documentos conocidos) y lo representamos con la notación $A = \{D_1, D_2, \dots, D_n\}$, donde los D_i se corresponden con cada uno de los documentos redactados por el autor.

Un documento será representado por un conjunto de Rasgos Lingüísticos extraídos a partir de un procesamiento realizado bien a nivel de caracteres, léxico o sintáctico, utilizando para cada caso herramientas de PLN. En nuestro trabajo vamos a considerar un total de 10 Clases de Rasgos (F), los que se describirán en secciones siguientes, y denotaremos con la siguiente expresión $F = \{F_1, F_2, \dots, F_{10}\}$. Para un F_i, cada documento se representa como $F_i(D) = (x_1(D), x_2(D), \dots, x_n(D))$, donde n denota el total de rasgos en el espacio de representación de los documentos para un F_i, siendo $F_i(D)$ la descripción del documento D y cada $x_i(D)$ el valor del rasgo x_i.

Semejanza entre un par de documentos $\beta(D_i, D_j)$ $i \neq j$: utilizamos tres funciones de comparación, Jaccard, Coseno y Minmax. Estas funciones han sido usadas en diversos trabajos presentados en las competencias PAN [5, 10, 13].

2.1. Arquitectura propuesta

Las colecciones de verificación de autoría ofrecidas en la competencia PAN [2, 3, 12], se estructuran por un conjunto de autores (problemas) y por cada autor se brinda una lista de documentos redactados por este y un documento de autoría desconocida. La tarea consiste en responder si el texto desconocido es redactado por el autor en análisis, no redactado por él o en abstenerse de responder.

Este escenario es similar al problema práctico que nos enfrentamos y queremos resolver, por lo que la base de nuestra propuesta radica en la implementación de un clasificador que sea capaz

de dar una respuesta de la autoría de un documento desconocido partiendo, únicamente, de las muestras conocidas de un autor (Verificación de Autoría Intrínseca VAI).

El objetivo que nos trazamos se corresponde con utilizar una combinación de respuestas de cada clasificador implementado y dar una respuesta final usando un voto por mayoría. Dividimos el total de respuestas en que se dice Sí sobre el total de respuestas. Obtenemos un valor entre 0 y 1, si la respuesta es menor a 0.5 entonces la decisión final es que no fue redactado por el autor, si es igual a 0.5 lo consideramos una abstención y el resto de los casos, o sea, cuando es mayor que 0.5 entonces se considera redactado por el autor.

Cada clasificador debe tomar una decisión a partir de las muestras que se tienen en el instante de la clasificación, sin contar para esto con fases de entrenamiento donde se puedan ajustar parámetros o realizar selección de rasgos o identificación de objetos no representativos.

2.2. Clasificador

En cada clasificador construido definimos 3 etapas necesarias, una primera etapa para la representación de los documentos; una segunda donde se comparan estas representaciones de cada documento y se analiza el grado de semejanza entre cada par de documento; y una tercera etapa en la que se determina si el documento desconocido ha sido redactado por el autor del que se dispone de muestras, utilizando una regla de decisión propia para este clasificador, ver Figura 1.

La etapa de representación es el paso inicial y una de las etapas más importantes en toda tarea de Análisis de Autoría. Para nuestro trabajo se propone emplear diferentes familias de rasgos a partir de analizar el contenido y la redacción de los documentos. Debemos aclarar que en un clasificador, se define un Tipo de rasgo de una de las familias de rasgos del contenido. Se emplean 3 familias de rasgos, basados en Caracteres, Léxico y Gramatical y en cada una diferentes Tipos de rasgos. Para la ejecución de un clasificador se debe contar con los documentos de muestra del autor y un documento de autoría desconocida. Las representaciones escogidas se explican con detalles en el epígrafe “Representaciones de los objetos”.

Luego, procedemos al cálculo de la semejanza entre cada par de documentos, con el propósito de conocer en qué medida son similares dos documentos a partir de la coincidencia de rasgos y a la frecuencia de uso de los mismos. Cobra vital importancia la identificación e implementación de las funciones de comparación entre documentos, aspecto este explicado con detalles en el epígrafe “Cálculo de la semejanza entre objetos, funciones de comparación”.

Proponemos dos estrategias de clasificación para el análisis de la semejanza de los objetos; una orientada a considerar cada documento como una instancia del problema y la segunda a partir de la construcción de un representante o prototipo de las muestras conocidas. Para cada una de estas estrategias definimos 3 reglas de decisión que nos permiten evaluar la pertenencia del documento desconocido como un documento redactado por el autor del que tenemos muestras conocidas. Los aspectos relacionados con la estrategia de clasificación basada en instancias y las reglas de decisión adoptadas en esta, se exponen en el epígrafe “Regla de decisión utilizando el promedio de semejanza entre objetos de una clase”; y en el epígrafe: “Regla de decisión utilizando la semejanza con centroide de una clase” se exponen detalles de la estrategia basada en prototipos.

El clasificador debe dar como respuesta: documento desconocido es redactado por el autor de las muestras conocidas (valor mayor a 0.5), se abstiene en determinar si fue redactado por este autor (valor 0.5) o determina que el documento de autoría desconocida no fue redactado por el autor de las muestras (valor menor de 0.5). Estos datos numéricos son los valores que permiten obtener un voto por mayoría en la combinación final de los clasificadores.

2.3. Representaciones de los objetos

Los rasgos lingüísticos son el núcleo de la tarea de análisis de autoría (independientemente de la subtarea de las mencionadas en la que se trabaje), ellos permiten codificar los documentos con algún modelo matemático, siendo tradicionalmente el más usado el modelo de bolsa de palabras (Bag of

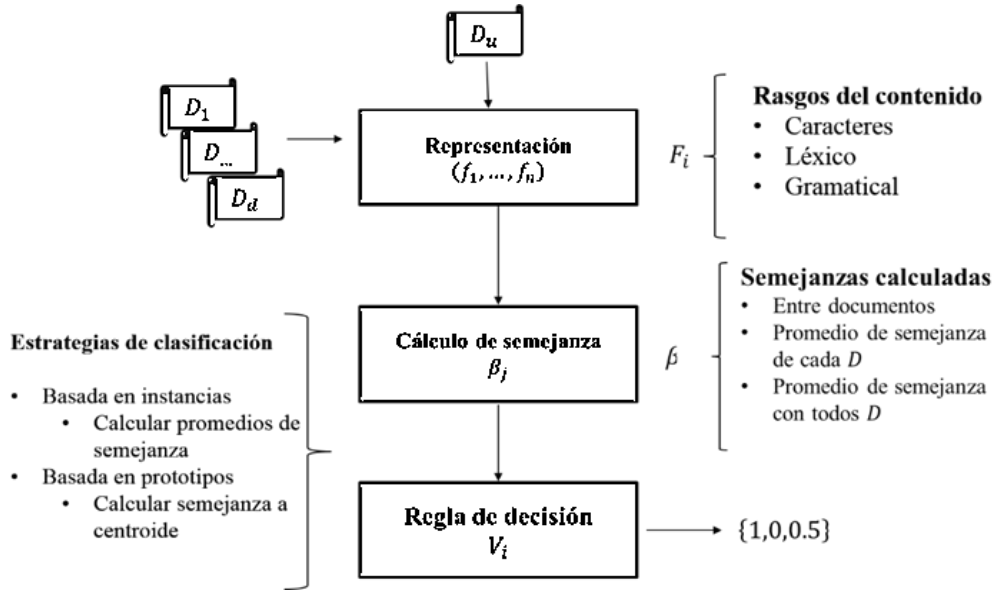

Fig. 1. Etapas del clasificador de Verificación de Autoría Intrínseco

Word, BoW), empleando como representación un n-uplo de rasgos. El propósito radica en intentar identificar un estilo propio de redacción para cada autor que lo diferencie del resto, en nuestro enfoque solo que lo caracterice a él, puesto que no dispondremos de muestras de otros autores.

Existe una gran cantidad de rasgos que han sido tomados en cuenta para la tarea de análisis de autoría por los investigadores, en la generalidad o mayoría, se usa una distribución o identificación de rasgos por capas lingüísticas (podemos llamarlos además, rasgos obtenidos a partir del contenido de la redacción).

Para nuestra propuesta escogimos 10 Tipos de Rasgos lingüísticos de los reportados en la literatura [2, 10, 14] agrupados en las siguientes Familias o Capas de rasgos de análisis lingüístico.

– Capa de caracteres:
  — **N-gramas de caracteres**: se obtienen como rasgos, todas las secuencias de n caracteres, sin eliminación de elementos en el texto. Para los experimentos se probó con varios valores de N y los mejores resultados se aprecian para N 3 y 4. Se construye una BoW con N = 3 y otra con N = 4. Para mostrar los resultados en los experimentos, relacionamos 3-grama de caracteres con (*F2*) y 4-grama de caracteres con (*F3*).
  — **N-gramas de Prefijos de tamaño k**: se construye una representación BoW tomando solo las N secuencias de caracteres de tamaño k a inicio de palabras. Para mostrar los resultados en los experimentos, relacionamos 2-grama-prefijo-tamaño-2 (*F5*).
  — **N-gramas de Sufijos de tamaño n**: esta es similar a la representación anterior, pero tomando las N secuencias de caracteres de tamaño k al final de cada palabra. Para mostrar los resultados en los experimentos, relacionamos 2-grama-sufijo-tamaño-2 (*F6*).

Los rasgos de esta capa son sencillos de calcular y nos permiten emplear herramientas no dependientes de un idioma. Para su cálculo se utilizan herramientas sencillas como los segmentadores de texto, que son usados para buscar patrones de redacción a través del uso de sufijos, prefijos, signos de puntuación, secuencias consecutivas de caracteres, entre otros.

– Capa léxica:
  — **N-gramas de palabras**: secuencias de N términos consecutivos luego de segmentado un texto. Construimos dos representaciones, una

con N = 1 y otra con N = 3. Para mostrar los resultados en los experimentos, relacionamos 1-grama de palabras (*F1*) y 3-grama de palabras (*F4*). Se toma N con 1 y 3 luego de probar con N de 1 a 5 y obtener los mejores resultados con 1 y 3.

Al igual que los rasgos de la capa de caracteres, los rasgos léxicos se pueden obtener empleando herramientas sencillas como los segmentadores de texto y son usados para buscar patrones de redacción a través del uso de palabras, secuencias consecutivas de palabras, entre otros.

- Capa gramatical:
  - **N-gramas de lemas**: secuencias de N lemas consecutivos luego de lematizado un texto. Construimos dos representaciones, una con N = 1 y otra con N = 3. Para mostrar los resultados en los experimentos, relacionamos 1-grama de lemas (*F7*) y 3-grama de lemas (*F9*).
  - **N-gramas de Etiquetas Gramaticales (PoS)**: secuencias de N etiquetas gramaticales consecutivas luego de etiquetado un texto. Construimos dos representaciones, una con N = 1 y otra con N = 3. Para mostrar los resultados en los experimentos, relacionamos 1-grama de PoS (*F8*) y 3-grama de PoS (*F10*).

Los rasgos de esta capa son un poco más complejos, dependiendo de herramientas de etiquetado y lematización de textos, son dependientes del idioma, requieren más tiempo para ser calculados y son usados para determinar patrones de redacción a través del uso de las categorías gramaticales y lematización de las palabras.

Para ilustrar el proceso de representación interna de los documentos usando cada uno de los rasgos lingüísticos supongamos que disponemos de un documento.

D1: *Me gusta pescar y navegar en las profundas aguas del mar Caribe.*

F1: *{(me, 1); (gusta, 1); (pescar, 1); (y, 1); (navegar, 1); (en, 1); (las, 1); (profundas, 1); (aguas, 1); (del, 1); (mar, 1); (caribe, 1); (.,1)}.*

F7: *{(me, 1); (gustar, 1); (pescar, 1); (y, 1); (navegar, 1); (en, 1); (el, 1); (profundo, 1); (agua, 1); (del, 1); (mar, 1); (caribe, 1); (.,1)}.*

F8: *{(PP1CS00, 1); (VMIP3S0, 1); (VMN0000, 1); (CC 1); (VMN0000, 1); (SP, 1); (DA0FP0, 1); (AQ0FP00, 1); (NCFP000, 1); (SP, 1); (NCCS000, 1); (NP00000, 1); (FP,1)}.*

En cada clasificador se determina como un parámetro de configuración, el Tipo de Rasgo con el que se representarán los documentos. Se construye un n-uplo de rasgos (términos) binario o pesado por la Frecuencia de su uso en el documento en análisis (term frequency, TF), dependiendo de la función de comparación que se empleará para el cálculo de la semejanza entre los objetos.

2.4. Cálculo de la semejanza entre objetos, funciones de comparación

Debido a que en nuestro problema práctico podemos encontrar documentos en las muestras de un autor con características muy variables, como, el tamaño, el género literario, la temática que abordan, entre otras. Además de la necesidad de un método general para ser usado en cualquier entorno de aplicación en la tarea de análisis de autoría decidimos escoger 3 funciones de comparación reportadas en la literatura, con el objetivo de tener un marco flexible capaz de adaptarse a cualquier entorno de aplicación.

Las funciones de comparación pueden dividirse en funciones de semejanzas y funciones de distancia. Las primeras evalúan la similitud entre dos objetos otorgando un valor cercano a 1 mientras más semejantes sean; en contraposición las de distancia determinan que dos objetos son semejantes a medida que el cálculo se acerca a 0. Para el desarrollo de los experimentos se implementaron funciones de semejanza para n-uplos binarios y para n-uplos pesados y una función de distancia.

El índice de Jaccard (1), mide la proporción existente entre la cantidad de elementos de la intersección de dos conjuntos sobre el total de elementos de la unión.

Siempre toma valores entre 0 y 1, correspondiente este último a la igualdad total entre ambos conjuntos. En informática se utiliza para medir la distancia entre vectores principalmente definidos sobre un espacio

vectorial booleano (las componentes del vector sólo pueden ser 0 o 1):

$$\frac{A \cap B}{|A \cup B|}. \quad (1)$$

La medida de similitud coseno (2) es usada para medir el valor del coseno ángulo comprendido entre dos vectores en un espacio, mientras menor sea el ángulo mayor es el coseno y en consecuencia mayor es la similitud entre los dos vectores. Es una medida ampliamente usada en la literatura. En comparación con el índice de Jaccard es una medida más exigente ya que no mide solamente la presencia de una determinada característica sino el nivel de importancia de esa característica en ambos vectores:

$$\frac{\sum_{i=0}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{|X|}(x_i)^2} + \sqrt{\sum_{i=1}^{|Y|}(y_i)^2}}. \quad (2)$$

En las funciones de distancia mientras más pequeño es el valor más cercano están los dos vectores y viceversa, mientras mayores sean los valores más alejados se encuentran. Las funciones de distancia pueden ser fácilmente convertidas en funciones de semejanza mediante la resta del valor 1 con el valor de la función de distancia.

A pesar de que en la literatura, la distancia euclidea es una de las más usadas, no la empleamos debido a que obtiene valores semejantes a la función coseno cuando los n-uplos están normalizados como en nuestro problema [1].

La distancia MinMax (3) determina la proporción existente entre los valores mínimos y los valores máximos pero tiene el inconveniente que solo toman en cuenta aquellas características que se encuentran en ambos documentos; ha sido utilizada en el algoritmo de [13], el cual se ubicó entre los primeros trabajos de la edición PAN 2013:

$$\frac{\sum_{i=1}^{r} \mathrm{Min}(x_i, y_i)}{\sum_{i=1}^{r} \mathrm{Max}(x_i, y_i)}. \quad (3)$$

2.5. Regla de decisión utilizando el promedio de semejanza entre objetos de una clase

La etapa final del clasificador se corresponde con las reglas de decisión que nos permiten obtener la respuesta, en nuestra tarea responder si el documento desconocido fue redactado por el autor de las muestras (responder 1), abstenerse de dar una respuesta (responder 0.5) o determinar que no fue redactado por él (responder 0). A continuación, expondremos los detalles de la clasificación basada en instancias y las tres reglas de decisión propuestas a partir de esta estrategia. A modo de resumen, se puede observar la figura 2.

Dados dos grupos de documentos A_1 y A_2, donde A_1 contiene el conjunto de muestras de un autor y A_2 el documento desconocido D_u, construimos un grupo nuevo $A = A_1 \cup A_2$ formado por la unión de todos los documentos de A_1 y A_2 y calculamos el promedio de semejanza del grupo A (PS_A) y el promedio de semejanza de cada documento D_i con el resto del grupo A ($PS_{D_i}^A$):

$$PS_{D_i}^A = \frac{\sum_{j=1, j\neq i}^{|A|} \beta(D_i, D_j)}{|A| - 1}, \quad (4)$$

$$PS_A = \frac{\sum_{i=1}^{|A|} PS_{D_i}^A}{|A|}. \quad (5)$$

Evaluamos las siguientes tres estrategias:

1. Se calcula el promedio de semejanza de todos los objetos del grupo, considerando al desconocido; si el promedio del documento desconocido es mayor que la media de los promedios del grupo, entonces, es bien semejante a la mayoría de los objetos conocidos y se considera redactado por este autor. Ver figura 2. Esta aproximación debe permitir que si el documento no fue redactado por el autor, entonces, aunque la semejanza del desconocido afecte a las semejanzas del resto de los objetos del grupo, este tendrá el menor promedio de semejanza, o su promedio no será mayor que la media del grupo. El punto débil se debe presentar en que se equivoque en aquellos objetos que se deben considerar redactados por el

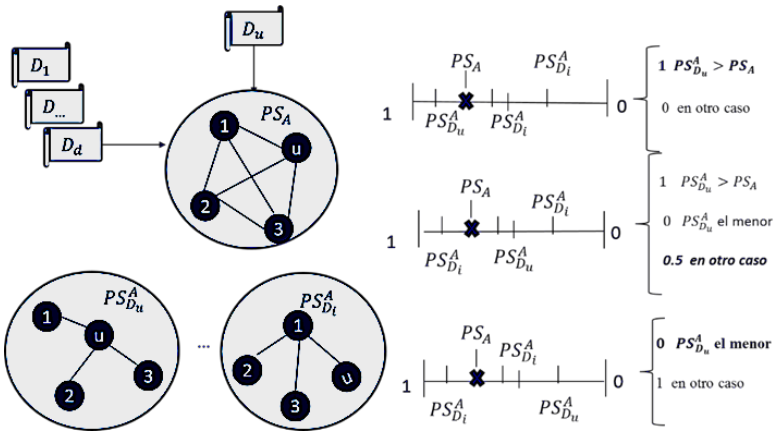

Fig. 2. Clasificador basado en instancias, calculando el promedio de semejanzas entre todos los documentos de muestra y el desconocido. Reglas de decisión a partir del promedio de semejanza del desconocido con las muestras del autor

autor y que su semejanza de promedio no es mayor que la media. Puede ser la menor semejanza promedio o no ser la menor, pero tampoco mayor que la promedio. Estos casos se darían como errores.

2. Una segunda estrategia implementada, considera como respuesta de redactado por el autor, los casos en que el promedio de semejanza del desconocido, sea mayor que el promedio del grupo, la respuesta de no redactado sería cuando el promedio de semejanza sea el menor y daría una respuesta de Abstención (Abs) si no es el menor promedio, pero no supera el promedio del grupo.

   En este caso, se busca que aquellos documentos desconocidos que son redactados por el autor y que no son semejantes a la mayoría, pero más semejantes que el menos semejante, no los dé como una respuesta de no redactado y evaluar en qué grado, los documentos no redactados tendrían el menor promedio de semejanza o se incluirían entre estos de Abs.

   Si se consideran entre los Abs es una señal de que la representación no está diferenciando los no redactados por algunas muestras. Este sería un indicio de que se pudiera trabajar en evaluar los rasgos obtenidos por cada tipo de rasgo, emplear selección de rasgos y evaluar técnicas de análisis de objetos no representativos.

3. La tercera idea reside en considerar como redactado por el autor a la muestra desconocida que no tenga el menor promedio de semejanza, por el criterio de que el documento no redactado debe tener el menor promedio de semejanza, lo que no quita que documentos si redactados tengan el menor promedio de semejanza.

2.6. Regla de decisión utilizando la semejanza con centroide de una clase

La segunda propuesta de clasificación empleada está basada en la construcción de un representante o prototipo de un grupo de documentos de muestra. La idea es que este representante permitiría agrupar todas las características presentes en las redacciones de los documentos de muestra.

La decisión de la pertenencia de un documento desconocido a las muestras conocidas parte entonces de evaluar la semejanza de este documento desconocido con el representante construido. A continuación, se explican los detalles de la clasificación basada en prototipo y las tres reglas de decisión para obtener la respuesta de un clasificador. Ver figura 3.

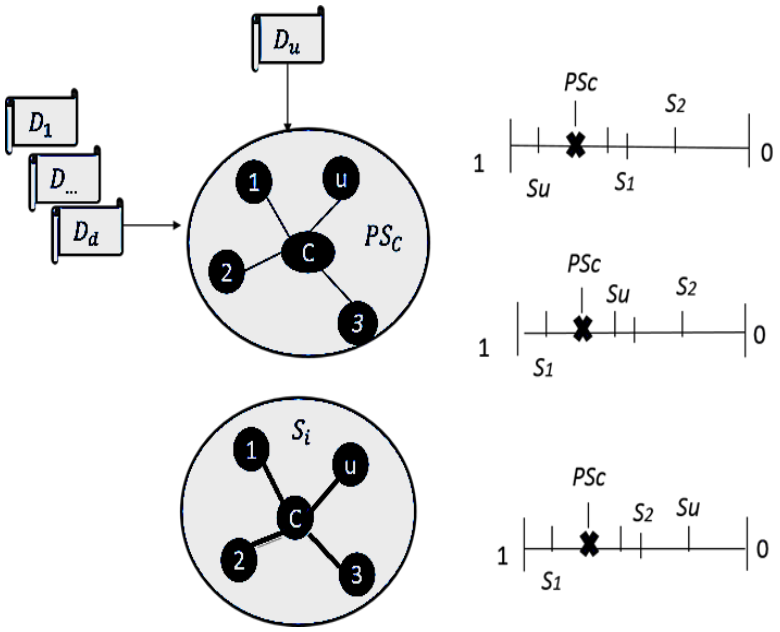

Fig. 3. Clasificador basado en prototipo, calculando la semejanza de los documentos con respecto al prototipo. Reglas de decisión a partir de la semejanza del desconocido con el representante del autor

Dados dos grupos de documentos A_1 y A_2, donde A_1 contiene el conjunto de muestras de un autor y A_2 el documento desconocido D_u, construimos un grupo nuevo $A = A_1 \cup A_2$ formado por la unión de todos los documentos de A_1 y A_2 y calculamos el centroide suma [11] del grupo A (C_A) y la semejanza de cada documento D_i del grupo A con C_A (6). Por último obtenemos el promedio de las semejanzas con el centroide PS_c(7):

$$S_{D_i}^{C} = \beta(D_i, C), \tag{6}$$

$$PS_c = \frac{\sum_{i=1}^{|A|} S_{D_i}^{C}}{|A|}. \tag{7}$$

Evaluamos las siguientes tres estrategias:

1. Se construye un centroide del grupo integrado por los documentos redactados y el desconocido, luego se calcula la semejanza de cada documento con el centroide. Para la comparación calculamos el promedio de las semejanzas con el centroide. Si la semejanza calculada del desconocido con el centroide es mayor que este promedio de semejanzas calculado, entonces se considera redactado por el autor. Se responde No redactado en caso contrario.
2. Si la semejanza del desconocido con el centroide es mayor que el promedio de las semejanzas al centroide, entonces es redactado. Si no es superior al promedio pero no es la menor semejanza entre el resto, entonces se considera una abstención. Es No redactado cuando presenta la menor semejanza con el promedio.
3. Si la semejanza del desconocido al centroide no es la menor semejanza, entonces consideramos al documento desconocido como redactado por el autor. No redactado en caso contrario.

3. Resultados experimentales

Para los experimentos se consideraron las colecciones de la competencia PAN de los años 2014 y 2015 para el español. En la edición del 2014 [3], la colección de autores para español presentaba como característica que las muestras

Tabla 1. Estructura y distribución de documentos y problemas de verificación de autoría en PAN 2014

	Idioma	Género	#Autores	#Docs	Promedio de docs conocidos por autor	Promedio de palabras por docs
Entrenamiento	Holandés	Ensayos	96	268	1.8	412.4
	Holandés	Comentarios	100	202	1.0	112.3
	Inglés	Ensayos	200	729	2.6	848.0
	Inglés	Novelas	100	200	1.0	3137.8
	Griego	Artículos	100	385	2.9	1404.0
	Español	Artículos	100	600	5.0	1135.6
	Total		696	2,384	2.4	1091.0
Prueba	Holandés	Ensayos	96	287	2.0	398.1
	Holandés	Comentarios	100	202	1.0	116.3
	Inglés	Ensayos	200	718	2.6	833.2
	Inglés	Novelas	200	400	1.0	6104.0
	Griego	Artículos	100	368	2.7	1536.6
	Español	Artículos	100	600	5.0	1121.4
	Total		796	2,575	2.2	1714.9

Tabla 2. Estructura y distribución de documentos y problemas de verificación de autoría en PAN 2015

	Idioma	Tipo	# Autores	#Docs	Promedio de docs Conocidos por problemas	Promedio de palabras por docs
Entrenamiento	Holandés	multi-género	100	276	1.76	354
	Inglés	multi-tópico	100	200	1.0	366
	Griego	multi-tópico	100	393	2.93	678
	Español	mixto	100	500	4.0	954
	Total		400	1369	2.42	588
Prueba	Holandés	multi-género	165	452	1.74	360
	Inglés	multi-tópico	500	1000	1.00	536
	Griego	multi-tópico	100	380	2.80	756
	Español	mixto	100	500	4.00	946
	Total		865	2332	2.3	

de cada autor eran homogéneas en cuanto a género textual y tema abordado, a diferencia de la colección presentada en 2015 [2], donde las muestras de cada autor podían ser de diferente género textual y tema.

En la tabla 1 y 2 se presentan las características de las colecciones que usaremos para los experimentos descritos en las próximas secciones. Se aprecia que la colección con mayor número de muestras por autores es la de español, además de contar con un número alto de palabras por documentos. Estos dos elementos son esenciales para los resultados que esperamos obtener en los experimentos que se describirán en las siguientes secciones.

La medida de evaluación que empleamos se corresponde con la medida de *accuracy* (8) propuesta y usada en las ediciones de las competencias PAN [3]

$$c@1 = \frac{n_c + \left(\frac{n_u * n_c}{n}\right)}{n}, \qquad (8)$$

donde n_c es la cantidad de respuestas correctas, n_u la cantidad de abstenciones y n el total de problemas a responder.

3.1. Centroide vs Promedio

Evaluamos inicialmente la efectividad de cada una de las estrategias de clasificación que proponemos, basada en instancias y en el centroide y en las siguientes secciones de experimentos, solo utilizaremos el enfoque que brinde mejores resultados.

Mostramos los resultados de las ejecuciones de las estrategias de centroide y promedio de semejanza sobre 4 colecciones de PAN que usamos, así como del empleo de las combinaciones de todos los pares funciones de comparación y rasgos. En la tabla 3 los valores obtenidos para la variante 1 de comparación. En la tabla se resaltan valores en los que los resultados son superiores con diferencia entre una estrategia de promedio de las instancias y la de centroide.

A modo de resumen se aprecia que la estrategia usando el promedio entre las instancias obtiene los mejores valores de *accuracy* que la estrategia de calcular el centroide. Con respecto a los idiomas se aprecian los mejores resultados en las colecciones de español y griego en PAN 2014 y español para PAN 2015, correspondiendo estas a colecciones con mayor cantidad de muestras conocidas por autores y textos con una longitud considerable.

3.2. Evaluando respuestas Sí, No y abstenciones

A continuación, vamos a ilustrar algunos resultados de las tres variantes propuestas de umbrales, especificando en el número de respuestas de Sí, No y las Abstenciones. Los resultados mostrados se corresponden a ejecuciones realizadas con la colección de textos en español de las datas de PAN 2014 test2.

En los resultados de la sección anterior no podemos apreciar donde se equivoca más la estrategia de clasificación, si en determinar los documentos que sí son redactados por el autor o en responder qué documento no fue redactado por el autor. Debemos señalar que en las colecciones de la competencia se brindan dos clases de problemas: clases de autores para los que el documento desconocido que se debe evaluar no fue redactado por él, y en este la respuesta positiva es decir No, y problemas en los que el documento desconocido si fue redactado por el autor y la respuesta positiva es decir Sí.

Los resultados mostrados en la tabla 4, son obtenidos solo considerando que el promedio de semejanza del desconocido supere la media de la semejanza entre todos. **Correcto** representa el total de respuestas positivas ya sea que se responda Sí cuando es Sí y No cuando no fue redactado. **Incorrecto** representa el total de respuestas negativas, o sea, dijo Sí cuando no fue redactado y viceversa. **Abstenciones** cuando la respuesta es una abstención, con esta estrategia, solo se da abstención en la combinación final, cuando entre los 10 tipos de rasgos, se alcancen 5 respuestas Sí y 5 No, pero para cada par rasgo-función siempre será 0. **Correcto Sí** cantidad de respuestas positivas en las que se debía responder Sí. **Correcto No** cantidad de respuestas positivas cuando se debía responder que No. **Accuracy** representa el valor de la medida *accuracy*, tal como se propone en las evaluaciones de las competencias PAN 2014 y 2015.

Tabla 3. Comparación de los resultados obtenidos con variante 1 entre promedio y centroide

Año	Colección	Idioma	Género	Promedio	Centroide
2014	Entrenamiento	Español	artículos	**0,84**	0,71
		Griego	artículos	0,53	0,46
		Inglés	novelas	**0,57**	0,4
		Inglés	ensayos	0,55	0,56
		Holandés	comentarios	0,49	0,5
		Holandés	ensayos	0,5	0,49
	Prueba	Español	artículos	**0,74**	0,64
		Griego	artículos	0,62	0,56
		Inglés	novelas	0,46	0,48
		Inglés	ensayos	0,6	0,54
		Holandés	comentarios	0,49	0,51
		Holandés	ensayos	0,58	0,45
2015	Entrenamiento	Español	mixto	**0,77**	0,69
		Griego	multi-tópico	0,58	0,55
		Inglés	multi-tópico	0,5	0,48
		Holandés	multi-género	0,57	0,54
	Prueba	Español	mixto	0,66	0,52
		Griego	multi-tópico	0,57	0,54
		Inglés	multi-tópico	0,5	0,5
		Holandés	multi-género	0,5	0,5

Tabla 4. Estrategia 1 empleando medida de comparación jaccard

Rasgos	Correcto	Incorrecto	Abstenciones	Correcto Sí	Correcto No	Accuracy
F1	67	33	0	22	45	0.67
F2	67	33	0	26	41	0.67
F3	69	31	0	27	42	0.69
F4	63	37	0	17	46	0.63
F5	59	41	0	18	41	0.59
F6	64	36	0	25	39	0.64
F7	68	32	0	23	45	0.68
F8	60	40	0	23	37	0.6
F9	67	33	0	23	44	0.67
F10	63	37	0	25	38	0.63
Combinación	65	31	4			0.67

Tabla 5. Estrategia 2 empleando medida de comparación jaccard

Rasgos	Correcto	Incorrecto	Abstenciones	Correcto Sí	Correcto No	Accuracy
F1	22	5	73	22	0	0.38
F2	26	9	65	26	0	0.42
F3	27	8	65	27	0	0.44
F4	17	4	79	17	0	0.30
F5	18	9	73	18	0	0.31
F6	25	11	64	25	0	0.41
F7	23	5	72	23	0	0.39
F8	23	13	64	23	0	0.37
F9	23	6	71	23	0	0.39
F10	25	12	63	25	0	0.40
Combinación	21	6	73			0.36

En la data que estamos mostrando del español, se cuenta con un total de 100 problemas de verificación y en cada problema un total de 5 muestras de documentos redactados por el autor y un documento desconocido.

Para la evaluación se conoce si el documento desconocido fue redactado o no por este autor. Se puede responder Sí redactado, No redactado o Abstenerse. Se presentan 50 problemas en los que la respuesta debe ser Sí y 50 en los que la respuesta debe ser No.

Como la restricción de esta estrategia es que solo se responda Sí cuando se supere la media de semejanza del grupo, se busca que el documento desconocido sea bien semejante a la mayoría de las muestras conocidas, según esta idea, debe responder positivo a todas las muestras desconocidas que no fueron redactadas, o sea decir No, y evaluar en qué grado es capaz de responder correctamente Sí, ya que para las respuestas de Sí es una restricción fuerte que supere la media. Se aprecia, en sentido general, que para la mayoría de las respuestas No, es positiva la respuesta y que en casi la mitad de las respuestas Sí, los documentos pasaban la frontera de la media.

Las principales respuestas negativas están en los Sí que se respondió que No por no superar la media y bastante interesante es ver cómo algunos documentos desconocidos en los que se debe responder No, superaron la media de su grupo de muestras de autor que en principio sería más semejante a la mayoría de las conocidas.

La estrategia en este caso (tabla 5) es responder que Sí, si el promedio del desconocido es mayor que la media del grupo, decir abstención (Abstención) si no supera la media del grupo, pero no es el menor promedio de semejanza del grupo y se responde que No cuando el promedio de semejanza del desconocido es el menor.

En este experimento, podemos ver cómo la mayoría de las respuestas son de abstención, casi todas en las que debía responder que No y el resto de las que debía responder que Sí, esto identifica que casi todas las respuestas en que debe decir Sí, o están por encima de la media del grupo o por debajo de la media, pero sin ser el menor promedio de semejanza, y que es bastante fácil que, un objeto no redactado por el autor sea, incluso, más semejante a sus muestras que algunas de las conocidas, por lo que es bastante difícil que tengan un promedio de semejanza mayor a la media, pero no fueron las muestras con menor promedio.

Si se considera la abstención como una respuesta más favorable a equivocarse, entonces el resultado es positivo, porque la cantidad de respuestas **Correcto** es para la mayoría de los

Tabla 6. Estrategia 3 empleando función de comparación Jaccard

Rasgos	Correcto	Incorrecto	Abstenciones	Correcto Sí	Correcto No	Accuracy
F1	50	50	0	50	0	0.5
F2	50	50	0	50	0	0.5
F3	50	50	0	50	0	0.5
F4	50	50	0	50	0	0.5
F5	50	50	0	50	0	0.5
F6	50	50	0	50	0	0.5
F7	50	50	0	50	0	0.5
F8	50	50	0	50	0	0.5
F9	50	50	0	50	0	0.5
F10	50	50	0	50	0	0.5
Combinación	50	50	0			0.5

rasgos mucho más alta que los errores **Incorrecto.**

Esta estrategia (tabla 6), determina como respuesta Sí, cuando el promedio de semejanza del desconocido no es el menor. Se responde No en caso contrario. Estamos tomando como frontera de decisión el objeto con menor promedio de semejanza.

Se aprecia que para todos los documentos desconocidos en los que se debe decir que Sí, estos nunca tienen el menor promedio de semejanza, y entonces el **Correcto Si** es igual al total de **Correcto**, pero se equivocó en los que debe responder No, porque estos tampoco son los objetos de menor promedio de semejanza, contrario a lo que se debía esperar.

Esto ilustra que podemos estar en presencia de situaciones en las que tenemos documentos (outliers) en los bordes de la distribución en el espacio de característica de los rasgos, probablemente debido a la cantidad de rasgos que pueden ser redundantes con respecto a los documentos desconocidos.

A modo de resumen para situaciones prácticas forenses sería conveniente utilizar la segunda estrategia de decisión, puesto que se equivoca menos que las otras estrategias, aunque obtiene un volumen alto de abstenciones.

Consideramos que introduciendo estrategias para determinar los documentos menos representativos de las muestras, se pudiera discriminar mejor con la tercera estrategia (tabla 6) y que en nuestro trabajo no usamos métodos de selección de rasgos que pudieran permitir una diferencia mayor, entre las muestras conocidas y el documento desconocido cuando este no pertenece al autor en análisis.

3.3. Comparación con trabajos presentados en la edición de PAN 2014

Queremos a continuación comparar los resultados obtenidos con las propuestas presentadas en la edición del PAN 2014. Para esto se presentan problemas en los que solo se cuenta con una muestra de documento conocido para el autor y esto no permitiría realizar la comparación, porque se necesitan al menos dos documentos de muestra para el cálculo de los promedios o del centroide.

Ante esta situación decidimos dividir el documento a la mitad y generar dos documentos. Esta es una idea muy simple y burda y reconocemos que podemos utilizar estrategias de segmentado más elaboradas, pero nos quedará para trabajo futuro. La dificultad mayor se concentra cuando tenemos una sola muestra y esta es corta.

Tabla 7. Resultados alcanzados de los participantes en la edición PAN 2014 y nuestro enfoque

Posición	Holandés-ensayo		Holandés-comentario		Griego-artículos	
	Trabajo	**c@1**	**Trabajo**	**c@1**	**Trabajo**	**c@1**
1	Frery et al.	0,9	Satyam et al.	0,69	Khonji & Iraqi	0,81
2	Mayor et al.	0,88	Khonji & Iraqi	0,65	Mayor et al.	0,75
3	Castillo et al.	0,86	Moreau et al.	0,59	Castillo et al.	0,73
4	Khonji & Iraqi	0,84	Zamani et al.	0,59	Moreau et al.	0,7
5	Jankowska et al.	0,84	Frery et al.	0,57	Jankowska et al.	0,68
6	Moreau et al.	0,83	Jankowska et al.	0,56	Zamani et al.	0,66
7	BASELINE	0,79	Halvani & Steinebach	0,55	**Nuestro+**	0,66
8	Satyam et al.	0,75	BASELINE	0,53	Frery et al.	0,642
9	**Nuestro+**	0,73	Mayor et al.	0,525	BASELINE	0,64
10	Vartapetiance & Gillam	0,71	Layton	0,52	**Nuestro**	0,62
12	Modaresi & Gross	0,63	Modaresi & Gross	0,5	Halvani & Steinebach	0,6
13	Halvani & Steinebach	0,617	**Nuestro**	0,49	Satyam et al.	0,6
14	Harvey	0,615	Harvey	0,48	Modaresi & Gross	0,54
15	**Nuestro**	0,58	Castillo et al.	0,37	Vartapetiance & Gillam	0,53
16	Layton	0,56			Harvey	0

Esta situación se refleja fundamentalmente en las colecciones de documentos del holandés. En la tabla 7 incluimos los resultados para tres colecciones y el resto en la tabla 8.

Nuestros resultados se observan con el nombre **Nuestro** y además adicionamos un **Nuestro+** que se corresponde con evaluar problemas en los que se tiene más de una muestra conocida.

Observamos que los resultados más bajos se alcanzan en las colecciones de novela en inglés, a partir de que todos los problemas de esta colección contienen un solo documento conocido a pesar de ser documentos extensos, y para el holandés en comentarios, donde los textos son bien cortos y una muestra conocida por autor.

Podemos apreciar que en las colecciones donde eliminamos el análisis de los problemas de una sola muestra, se mejoran los valores de *accuracy* y se alcanzan los mayores valores en la colección de español donde se presenta un mayor número de documentos de muestra por autor.

3.4. Verificación de autoría para todos los idiomas de las colecciones

En las colecciones que se brindan en las competencias PAN, se incorporan muestras para la verificación de autoría en los idiomas inglés, holandés y griego. La propuesta que implementamos es dependiente de las Clases de Rasgos con las que se representan los documentos y, como se expone en la descripción de los rasgos empleados, estos se obtienen en dependencia de determinadas herramientas de PLN disponibles.

Realizamos experimentos para los 4 idiomas brindados: español, inglés, griego y holandés. Como salvedad, debemos mencionar que al no disponer de lematizador y etiquetador morfológico para el griego y el holandés, solo se utilizaron combinaciones de 6 clases de rasgo, [*F1- F6*] y para el inglés, al igual que para español, desde [*F1- F10*].

Anteriormente, comprobamos que esta aproximación del promedio es sensible cuando se

Tabla 8. Resultados alcanzados de los participantes en la edición PAN 2014 y nuestro enfoque

Posición	español,-artículos		inglés,-ensayos		inglés,-novelas	
	Trabajo	**c@1**	**Trabajo**	**c@1**	**Trabajo**	**c@1**
1	Khonji & Iraqi	0,77	Frery et al.	0,71	Modaresi & Gross	0,71
2	Castillo et al.	0,76	Satyam et al.	0,65	Zamani et al.	0,65
3	Moreau et al.	0,75	Layton	0,61	Castillo et al.	0,615
4	Frery et al.	0,75	**Nuestro**	0,6	Mayor et al.	0,614
5	**Nuestro**	0,74	Moreau et al.	0,6	Khonji & Iraqi	0,61
6	Jankowska et al.	0,73	Khonji & Iraqi	0,583	Frery et al.	0,58
7	Mayor et al.	0,71	Modaresi & Gross	0,58	Satyam et al.	0,57
8	Vartapetiance & Gillam	0,66	Castillo et al.	0,58	Moreau et al.	0,525
9	Harvey	0,65	Mayor et al.	0,557	Harvey	0,525
10	Modaresi & Gross	0,65	Zamani et al.	0,55	Halvani & Steinebach	0,515
11	Zamani et al.	0,64	Jankowska et al.	0,548	Layton	0,51
12	Halvani & Steinebach	0,64	Harvey	0,54	Vartapetiance & Gillam	0,49
13	Satyam et al.	0,56	Halvani & Steinebach	0,538	**Nuestro**	0,46
14	Layton	0,54	BASELINE	0,53	Jankowska et al.	0,45
15	BASELINE	0,53	Vartapetiance & Gillam	0,52	BASELINE	0,44

dispone de una sola muestra conocida, y en las colecciones de los idiomas griego, holandés ensayo e inglés ensayo se presentan problemas (autores) en los que se dispone de una sola muestra conocida.

Para estos casos, elaboramos una sub-colección eliminando esos problemas y en la tabla de los resultados se llaman igual que la anterior pero con un +. Incluiremos los valores obtenidos en las dos primeras variantes y usando la combinación de los 30 pares de función-rasgo. Ver tabla 9.

Es interesante en estos resultados, apreciar los valores obtenidos para las colecciones de español y holandés ensayo con más de una muestra. En estos se reduce en gran medida en la variante 2 de decisión el número de respuestas en que se equivoca, aunque se incrementan considerablemente las abstenciones. No obstante considero para una situación práctica pericial que es preferible que se abstenga a que dé respuestas equivocadas.

3.5. Influencia de las funciones de comparación

Al ser tres funciones de comparación las propuestas a usar, debemos analizar la influencia de cada una, o sea, evaluar cuál aporta en las decisiones correctas, siempre empleando todas las clases de rasgos con los que se representan los documentos.

En la tabla 10, podemos ver los valores de *accuracy* en las diferentes colecciones, cuando se emplean todas las funciones (30 pares función-rasgo), dos funciones (20 pares función-rasgo) y solo una función de comparación (10 pares función-rasgo). Los resultados presentados se corresponden con la variante 1 propuesta.

Se observa, como resumen, que los valores alcanzados, cuando utilizamos las tres funciones de comparación, en su mayoría son superiores a los alcanzados cuando se emplean dos o una, pero no son significativamente más altos. De todas las funciones de comparación se pueden resaltar los valores obtenidos cuando empleamos la

Tabla 9. Valores de accuracy para todas las colecciones y todas las combinaciones de pares rasgo-función. Variantes 1 y 2. En la variante 2 se expone accuracy(respuestas Positivas, Negativas, Abstenciones)

Año	Colección	idioma	género	todo (variante1)	todo (variante2)
2014	Entrenamiento	Español	artículos	**0,84**	**0.58(36/2/62)**
		Griego	artículos	0,53	0.26(15/11/73)
		Griego +	artículos	0,55	0.16(7/0/73)
		Inglés	novelas	0,57	0.47(33/23/44)
		Inglés	ensayos	0,55	0.43(58/45/97)
		Inglés +	ensayos	0,53	0.34(31/28/85)
		Holandés	comentarios	0,49	0.49(49/49/1)
		Holandés	ensayos	0,5	0.47(37/37/21)
		Holandés +	ensayos	0,67	0.57(12/1/21)
	Prueba	Español	artículos	**0,74**	**0.47(28/4/68)**
		Griego	artículos	0,62	0.41(26/13/61)
		Griego +	artículos	0,66	0.34(15/2/61)
		Inglés	novelas	0,46	0.33(42/40/118)
		Inglés	ensayos	0,6	0.46(62/40/98)
		Holandés	comentarios	0,49	0.49(49/50/1)
		Holandés	ensayos	0,58	0.5(37/29/29)
		Holandés +	ensayos	**0,73**	**0.5(13/1/29)**
2015	Entrenamiento	Español	mixto	**0,77**	**0.45(26/0/74)**
		Griego	multi-tópico	0,58	0.33(20/11/69)
		Griego +	multi-tópico	0,63	0.35(18/2/69)
		Inglés	multi-tópico	0,5	0.5(48/46/6)
		Holandés	multi-género	0,57	0.49(37/30/33)
		Holandés +	multi-género	0,6	0(0/0/33)
	Prueba	Español	mixto	0,66	0,63(62/31/7)
		Griego	multi-tópico	0,57	0,57(57/39/4)
		Inglés	multi-tópico	0,5	0,5(250/250/0)
		Holandés	multi-género	0,5	0,5(83/82/0)

función de distancia MinMax y la semejanza Coseno.

3.6. Influencia de cada clase de rasgo empleado

Otro aspecto importante que evaluamos es la influencia o aporte de las representaciones con cada clase de rasgo. Para esto, analizamos la variación de los resultados de *accuracy* cuando mantenemos la combinación de los resultados de emplear una función de comparación y solo eliminamos una clase de rasgo. Los resultados se aprecian en las tablas 11, 12 y 13.

En la columna se denota como **No 1** a no considerar el empleo del Tipo de Rasgo *F1*, de forma similar el resto de las columnas.

En los resultados, no se aprecia una marcada disminución de los valores de *accuracy* cuando

Tabla 10. Variante 1 de promedio para colecciones de español, variando las funciones de comparación, y manteniendo todas las clases de rasgos

Año	Colección	todo	Jacc-coseno	Jacc-Minmax	Cose-Minmax	Jaccard	Coseno	MinMax
2014	Entrenamiento	**0,84**	0,8	0,85	0,8	0.8	0.82	0.73
	Prueba	**0,74**	0,67	0,73	0,72	0.67	0.71	0.73

Tabla 11. Variante 1 de promedio para colecciones de español, usando Jaccard como función de comparación y dejando de usar un Rasgo en la combinación

Año	Colección	todo	No 1	No 2	No 3	No 4	No 5	No 6	No 7	No 8	No 9	No 10
2014	Entrenamiento	0,8	0,82	0,8	0,81	0,78	0,8	0,78	0,78	0,79	0,77	0,77
	Prueba	0,67	0,64	0,64	0,64	0,66	0,7	0,68	0,64	0,68	0,66	0,66

Tabla 12. Variante 1 de promedio para colecciones de español, usando Coseno como función de comparación y dejando de usar un Rasgo en la combinación

Año	Colección	todo	No 1	No 2	No 3	No 4	No 5	No 6	No 7	No 8	No 9	No 10
2014	Entrenamiento	0,82	0,83	0,81	0,82	0,8	0,82	0,81	0,84	0,81	0,8	0,81
	Prueba	0,71	0,71	0,71	0,69	0,7	0,72	0,7	0,68	0,7	0,69	0,7

Tabla 13. Variante 1 de promedio para colecciones de español, usando MinMax como función de comparación y dejando de usar un Rasgo en la combinación

Año	Colección	todo	No 1	No 2	No 3	No 4	No 5	No 6	No 7	No 8	No 9	No 10
2014	Entrenamiento	0,73	0,73	0,71	0,7	0,73	0,71	0,7	0,72	0,71	0,72	0,72
	Prueba	0,73	0,72	0,7	0,72	0,73	0,72	0,71	0,7	0,7	0,72	0,73

dejamos de emplear alguno de los Tipos de rasgo propuestos.

En resumen, en las secciones de los experimentos, cuando evaluamos el uso de algunas funciones de comparación y cada uno de los Tipos de Rasgos, se observa que la combinación de varios rasgos o de varias funciones de comparación, nos permite obtener valores similares sin mucha afectación. Se debe analizar en detalle cada uno de los rasgos de forma independiente.

3.7. Resultados según la cantidad de muestras conocidas por autor

En los experimentos realizados ocurre que todos los autores presentan la misma cantidad de muestras de documentos conocidos, a pesar de ser pocas.

Con esto, no podemos analizar el impacto que se produce cuando se varía la cantidad de las muestras. La idea que subyace es que mientras mayor sea la cantidad de muestras conocidas, debe equivocarse menos el método, pero también pasa que la dispersión de los objetos en el espacio es mayor.

Con el próximo experimento vamos a evaluar los valores de *accuracy* a medida que incrementamos la cantidad de muestras. Comenzamos con una sola muestra conocida, hasta el total de las muestras. Para esto, promediamos los valores de accuracy para cada autor con una estrategia de validación Leave-one out.

Tabla 14. Variante 1 de promedio para colecciones de español, variando la cantidad de documentos conocidos de muestra y variando las funciones de comparación

# de Docs	todo	Jacc-coseno	Jacc-Minmax	Cose-Minmax	Jaccard	Coseno	MinMax
uno	0,5	0,5	0,5	0,51	0,5	0,54	0,49
dos	0,71	0,7	0,71	0,7	0,68	0,69	0,66
tres	0,77	0,76	0,71	0,73	0.74	0.74	0.73

Tabla 15. Variante 1 de promedio para colecciones de español, variando la cantidad de documentos conocidos de muestra y variando las clases de rasgo empleadas

# de Docs	todo	No 1	No 2	No 3	No 4	No 5	No 6	No 7	No 8	No 9	No 10
uno	0,5	0,5	0,52	0,52	0,51	0,52	0,52	0,5	0,5	0,5	0,5
dos	0,71	0,71	0,71	0,7	0,72	0,72	0,71	0,71	0,72	0,71	0,73
tres	0,71	0,7	0,7	0,7	0,72	0,72	0,7	0,71	0,71	0,72	0,71

Vamos a realizar dos corridas, una en la que empleamos todos los rasgos y vamos eliminando funciones de comparación y otra, en la que mantenemos todas las funciones de comparación y eliminamos un rasgo a la vez. Los valores se reflejan en las tablas 14 y 15. Según las estrategias planteadas para el cálculo del promedio de semejanza y la definición de los umbrales de decisión en base a estos promedios, mínimo necesitamos contar con dos documentos. Para las evaluaciones en que dejamos un solo documento conocido, lo que hicimos fue dividir el documento a la mitad y construir dos documentos. Intuitivamente, esto conformaría dos documentos bien parecidos por lo que el promedio de semejanza debe ser bien alto.

Se observa, como era de esperar, que el cambio en los valores de *accuracy* entre tener un solo documento y más de uno es significativo, por lo que se debe trabajar en estrategias más elaboradas cuando se presenta un problema de un solo documento conocido. A partir de contar con dos documentos o más, no se evidencian diferencias de los resultados.

Tendríamos que estudiar otros fenómenos con respecto a la distribución de las muestras en el espacio de representación, en nuestra aproximación podríamos estudiar la desviación que se experimenta en los valores de los promedios de semejanza de los documentos con respecto al resto, evaluando la dispersión de los documentos de muestra. Esto permitiría definir el uso de algunas de las decisiones de comparación atendiendo a la desviación de las muestras en la clase.

4. Conclusiones y trabajo futuro

Implementamos un método de Verificación de Autoría, atendiendo solo a las muestras conocidas de un autor y sin la calibración de parámetros en fases de entrenamiento. Para este implementamos dos estrategias de representación de las muestras, una basada en instancias y la segunda en un centroide. Debemos evaluar con mayor detalle la aproximación usando centroide.

Definimos tres variantes para calcular la decisión de cuando un documento desconocido pertenece a las muestras del autor, o sea que fue redactado por este o no y consideramos que las variantes 1 y 2 son las más adecuadas, aunque de estas para una situación práctica pericial la variante 2 es menos estricta que la 1 presentándose menos equivocaciones aunque un número alto de abstenciones.

Consideramos que es necesario dedicar esfuerzos en la incorporación de técnicas de

selección de rasgos, que permitan diferenciar mejor los documentos no redactados por el autor de sus muestras conocidas y que la semejanza de sus muestras conocidas sea mayor entre ellas.

Se resalta que la propuesta es sensible al número de muestras conocidas y al tamaño de las mismas. La combinación de varias funciones de comparación y tipos de rasgos para la representación permite que el modelo no se afecte cuando con alguna de estas no se obtienen valores similares al resto.

Debemos además, evaluar en qué medida se obtienen mayorías simples o altas tanto para responder que sí fue redactado, como para responder que no y devolver con esto, un grado de certeza de la respuesta.

No es suficiente con los resultados alcanzados para determinar, con absoluta certeza, cuando un documento No es escrito por un autor, siendo este el detalle en que más debemos trabajar, puesto que con la variante 1 y 2 de decisión, se equivoca poco en responder que sí fue redactado.

Referencias

1. **Manning, C.D., Raghavan, P., & Schütze, H. (2008)**. *Introduction to information retrieval.* Cambridge University Press, Vol. I-XXI, pp. 1–482.
2. **Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., Sánchez-Pérez, M.A., & Barrón-Cedeño, A. (2014).** Overview of the Author Identification Task at PAN 2014. *CLEF Working Notes,* pp. 877–897.
3. **Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., & Stein, B. (2015).** Overview of the Author Identification Task at PAN 2015. *CLEF Working Notes.*
4. **Castillo, E., Vilariño-Ayala, D., Cervantes, O., & Pinto, D. (2015):** Author attribution using a graph based representation. *CONIELECOMP*, pp. 135–142. DOI: 10.1109/CONIELECOMP.2015.7086940.
5. **Castillo, E., Cervantes, O., Vilariño-Ayala, D., Pinto, D., & León, S. (2014)**. Unsupervised Method for the Authorship Identification Task. *CLEF Working Notes*, pp. 1035–1041.
6. **Ghaeini, M.R. (2013).** Intrinsic Author Identification Using Modified Weighted KNN. *Notebook for PAN at CLEF Working Notes.*
7. **Jankowska, M., Keselj, V., & Milios, E. (2014).** Ensembles of Proximity-Based One-Class Classifiers for Author Verification. *Notebook for PAN at CLEF CLEF Working Notes,* pp. 1069–1072.
8. **López-Monroy, A.P., Montes-y-Gómez, M., Pineda-Villaseñor, L., Carrasco-Ochoa, J.A., & Martínez-Trinidad, J.F. (2012)**. A New Document Author Representation for Authorship Attribution. *Pattern Recognition 4th Mexican Conference, MCPR,* Huatulco, Mexico, Springer, pp. 283–292. DOI: 10.1007/978-3-642-31149-9_29.
9. **López-Monroy, A.P., Montes-y-Gómez, M., Pineda-Villaseñor, L., Carrasco-Ochoa, J.A., & Martínez-Trinidad, J.F. (2012)**. A new document author representation for authorship attribution. *Lect Notes Comput Sci*, Vol. 7329, pp. 283–292. DOI:10.1007/978-3-642-31149-9_29.
10. **Halvani, O., Steinebach, M., & Zimmermann, R. (2013).** Authorship Verification via k-Nearest Neighbor Estimation. *Notebook PAN at CLEF*.
11. **Juola, P. (2006)**: Authorship Attribution. *Foundations and Trends in Information Retrieva*l, Vol. 1, No. 3, pp. 233–334.
12. **Juola, P. & Stamatatos, E. (2013).** Overview of the Author Identification Task at PAN. *CLEF, Working Notes*.
13. **Ruiz-Shulcloper, J. (2009)**. *Reconocimiento Lógico Combinatorio de Patrones: Teoría y Aplicaciones.* Tesis en opción al grado científico de Doctor en Ciencias, La Habana.
14. **Seidman, S. (2013).** Authorship Verification Using the Impostors Method. *Notebook for PAN at CLEF, CLEF Working Notes.*
15. **Stamatatos, E. (2009)**. A survey of modern authorship attribution methods. *J Am Soc. Inf. Sci. Technol.,* Vol. 60, No. 3, pp. 538–556. DOI:10.1002/asi.21001.
16. **Sapkota, U., Bethard, S., Montes-y-Gómez, M., & Solorio, T. (2015).** Not All Character N-grams are Created Equal: A Study in Authorship Attribution. *HLT-NAACL*, pp. 93–102.

Artículo recibido el 14/11/2016; aceptado el 17/03/2017.
Autor de correspondencia es Daniel Castro.