Método Supervisado orientado a la clasificación automática de documentos. Caso Historias Clínicas

Roque E. López Condori[1] Dennis Barreda Morales[2] Javier Tejada Cárcamo[2]
Luis Alfaro Casas[1]

[1]Universidad Nacional de San Agustín, Arequipa-Perú
[2]Universidad Católica San Pablo, Arequipa-Perú

rlopezc27@gmail.com, dennis.barreda@ucsp.edu.pe, jtejadac@ucsp.edu.pe, casas@unsa.edu.pe

Resumen

En este artículo se presenta un método supervisado para la clasificación automática de documentos aplicado a historias clínicas. El clasificador propuesto consta de dos fases: la primera utiliza técnicas de procesamiento de lenguaje natural para la extracción de características de las historias clínicas. La segunda fase combina dos métodos de clasificación, el método Basado en Prototipos para reducir el número de clases iniciales seleccionando las más similares y el método Naive Bayes para clasificar la historia clínica en una sola clase. Los resultados obtenidos mejoran, o en el peor de los casos, igualan los resultados obtenidos por los métodos individuales en la clasificación de historias clínicas.

Palabras clave: clasificación automática de documentos, historias clínicas, procesamiento de lenguaje natural.

1. Introducción

La clasificación automática de textos consiste en asignar un documento dentro de un grupo de clases previamente definidas[6]. Si el documento pertenece sólo a una de las categorías, se trata de una *clasificación de una sola etiqueta*, caso contrario, es una *clasificación multi-etiqueta*[3][17]. En la clasificación de textos se distingue dos escenarios. En el primer escenario, conocido como *clasificación supervisada o categorización*, se parte de un conjunto de categorías diseñadas a priori, y la tarea del clasificador es asignar cada documento a la clase que le corresponda[7]. En el segundo escenario, conocido como *clasificación no supervisada o clustering*, no hay categorías definidas previamente, todos los documentos se agrupan en función de ellos mismos según su contenido u otro criterio similar[1].

En la actualidad existe una gran cantidad de historias clínicas disponible. Toda esta información es improductiva si no se cuenta con mecanismos adecuados para su acceso, clasificación y análisis. La necesidad de poder utilizar esta información ha llevado a la creación de diversos medios de manipulación de información, entre los que se encuentra la *clasificación*. Sin embargo, el incremento constante de las historias clínicas, hace que la tarea de clasificación manual sea costosa y además consuma mucho tiempo, por lo que ha surgido un interés en realizar la clasificación de forma automática. Una historia clínica es el conjunto de documentos que contiene los datos sobre la situación y evolución clínica de un paciente a lo largo del proceso asistencial[11].

En este artículo se presenta un método supervisado para la clasificación automática de historias clínicas. El clasificador propuesto combina el método Basado en Prototipos y Naive Bayes. El resto del documento se organiza de la siguiente manera. En la sección 2 se explican los pasos que se realizan antes de que una historia clínica sea procesada por un clasificador. La sección 3 introduce los métodos de clasificación empleados, en este caso el método Basado en Prototipos y Naive Bayes. En la sección 4 se explica el método propuesto para la clasificación de historias clínicas. En la sección 5 se describen los datos de pruebas utilizados en este trabajo para los experimentos. Finalmente en la sección 6 se exponen nuestras conclusiones.

2. Pre-procesamiento de las Historias Clínicas

El pre-procesamiento consiste en transformar las historias clínicas, de su formato original, a un modelo matemático adecuado para la tarea de clasificación. El pre-procesamiento consta de dos etapas: la extracción de las características principales y la parametrización de las historias clínicas.

2.1. Extracción de características

El objetivo de esta etapa es eliminar las partes de las historias clínicas que no sean importantes, es decir, que no aportan significado. Esta etapa consta de los siguientes pasos:

- Eliminación de palabras vacías (*stop words*). Elimina palabras que no transmiten informacón (pronombres, preposiciones, conjunciones, etc.).
- Eliminación de símbolos de puntuación.
- Lematización de palabras. Elimina sufijos y afijos de una palabra de tal modo que aparezca sólo su raíz léxica. Por ejemplo, los vocablos *medicina*, *médico* y *medicinal* tienen la raíz léxica *medic*. Para este paso se utiliza el algoritmo de Porter[14].

2.2. Parametrización de Historias Clínicas

Existen varias maneras de representar un documento; pero la más usada es el *modelo vectorial*[10] (ver figura 1). En este modelo, las historias clínicas se representan por vectores de palabras en un espacio de n dimensiones, siendo n el número de palabras en el texto. De esta manera, las historias clínicas quedan representados como un vector $d = (w_1, ..., w_n)$, donde cada término indexado corresponde a una palabra en el texto y tiene un peso w_i que refleja la importancia del término.

Existen varios mecanismos para hallar esta importancia, en nuestros experimentos, el peso de un término se representa por el mecanismo $tf \cdot idf$ [16], el cual combina la frecuencia del término en el documento con la

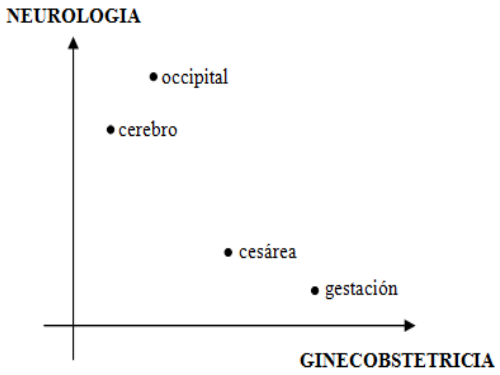

Figura 1: Espacio vectorial

frecuencia de éste en el resto de los documentos de la colección, y se calcula como:

$$w_i = tf_i \cdot log(\frac{N}{n_i}) \tag{1}$$

donde tf_i es la frecuencia de aparición del término *i-ésimo* en el documento, N es el número total de documentos y n_i es el número de documentos en los que aparece el término *i-ésimo*.

3. Métodos de clasificación

En la actualidad existe un número importante de métodos para la clasificación de textos. Con base a los resultados obtenidos en [4] y [13], se optó por el método Basado en Prototipos y Naive Bayes para la clasificación de historias clínicas.

3.1. Método Basado en Prototipos

Este método busca un documento representante de cada clase. El problema es elegir cuál de ellos tomar como representante o, si es necesario, crear un *documento virtual* que represente de mejor manera a la clase, a este representante de la clase se le conoce como *prototipo*[5].

Dado un conjunto de documentos $D = (d_1, ..., d_m)$ asociado a la clase C, el prototipo p se calcula hallando el peso de todas las palabras presentes en la clase C. El peso de cada palabra se obtiene a través de la ecuación 1.

La clasificación se lleva a cabo comparando el documento a clasificar con cada prototipo, tal como se muestra en la figura 2. Finalmente el documento se asigna a la clase del prototipo más similar.

3.2. Método Naive Bayes

Naive Bayes es uno de los modelos probabilistas ampliamente utilizados en la clasificación de textos, porque produce resultados tan buenos como otros modelos más sofisticados[2]. Si se tiene un conjunto de documentos $D = (d_1, ..., d_m)$ asociado a las clases predefinidas $C = (c_1, ..., c_n)$, cada documento es representado por un vector $d_j = (w_{1j}, ..., w_{|T|j})$ donde T es el conjunto de términos que pertenecen a c_i; el método bayesiano estima la probabilidad a posteriori de cada clase c_i dado el documento d_j.

Este clasificador se basa en la aplicación de la regla de Bayes para predecir la probabilidad condicional de que un documento pertenezca a una clase $P(c_i|d_j)$a partir de la probabilidad de los documentos dada la clase $P(d_j|c_i)$ y la probabilidad a priori de la clase en el conjunto de entrenamiento $P(c_i)$.

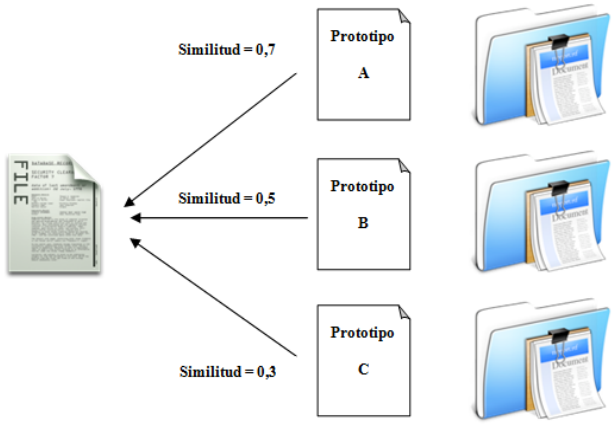

Figura 2: Método Basado en Prototipos

$$P(c_i|d_j) = \frac{P(c_i) \cdot P(d_j|c_i)}{P(d_j)} \tag{2}$$

4. Método Propuesto

La labor de un clasificador de historias clínicas se vuelve más difícil conforme se aumente el número de clases en las que una historia nueva puede ser clasificada. Más aún, pueden existir clases muy similares, esto haría más difícil la clasificación. Por lo anterior, una solución sería reducir el número de clases, es más sencillo elegir la clase a la que pertenece una historia clínica en un número reducido de posibilidades.

Este método propone realizar el proceso de clasificación en dos etapas. En la primera etapa se utilizan las técnicas de lenguaje natural mencionadas en la sección 2 para extraer las características principales de las historias clínicas. Para la lematización de las palabras se utilizó el algoritmo de Porter[14], este algoritmo asegura que la forma de las palabras no penalice la frecuencia de estas mismas. En la parametrización de las historias clínicas se empleó el modelo vectorial.

La segunda etapa en primer lugar busca reducir el número de clases iniciales seleccionando las clases más similares a cada historia clínica nueva, este paso utiliza el método de Basado en Prototipos. La selección de las clases más similares se realiza en base a un umbral de similitud, el cual viene dado por el promedio de las similitudes de los centroides. Es decir, las clases que superan este umbral pasan al siguiente paso. Posteriormente se clasifica la nueva historia clínica, asignándola a una de las clases que se seleccionó en el paso anterior usando el método de Naive Bayes. La figura 3 muestra la arquitectura de este clasificador.

5. Evaluación Experimental

5.1. Datos de Prueba

Con el objetivo de probar si el rendimiento del método propuesto podía mejorar los resultados del método Basado en Prototipos y Naive Bayes se realizaron experimentos con historias clínicas, adicionalmente se utilizaron otras colecciones de textos como R(8) y WebKB.

5.1.1. Historias Clínicas

La colección Historias Clínicas está formada por 261 documentos los cuales están distribuidos de manera casi uniforme en 6 clases. Esta colección se elaboró con el fin de realizar pruebas sobre historias clínicas reales. La colección se creó de forma manual por profesionales expertos y cada historia clínica fue asignada a una sola categoría. La tabla 1 la composición de esta colección.

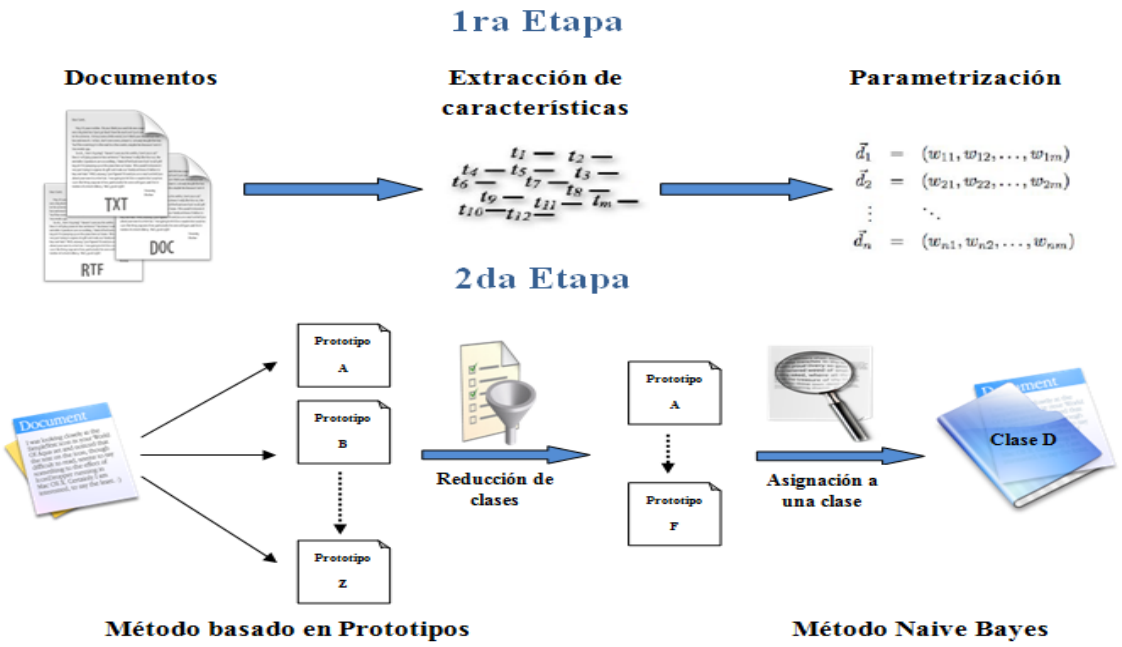

Figura 3: Arquitectura del clasificador propuesto

Tabla 1: Composición de la colección Historias Clínicas

Clases	Documentos
Gastroenterología	41
Ginecología	48
Neurología	42
Oncología	43
Traumatología	42
Urología	45
Total	261

5.1.2. R(8)

La colección Reuters-21578 es un recurso linguístico ampliamente utilizado en el marco de clasificación de textos[9]. R(8) es una subcolección de Reuters-21578, en la que todos los documentos sólo pertenecen a una clase, está conformada por 8 clases, en la tabla 2 se detalla la distribución de los documentos de esta colección.

5.1.3. WebKB

Para las pruebas realizadas en este trabajo, se utiliza una subcolección del original, el cual ya se encuentra distribuido en documentos de entrenamiento y prueba. Esta subcolección es utilizada por [1] y [3]. La distribución de esta colección se muestra en la tabla 3.

5.2. Resultados

5.2.1. Evaluación de la colección Historias Clínicas

Cada documento de esta colección se caracteriza por tener poco texto y también porque existen muchas palabras que aparecen en todas las categorías, como por ejemplo los términos: pacientes, dolor, enfermedad,

Tabla 2: Composición de la colección R(8)

Clase	Entrenamiento	Prueba
Trade	251	75
Grain	41	10
Acq	1596	696
Earn	2840	1083
Interest	190	81
Money-fx	206	87
Ship	108	36
Crude	253	121
Total	5845	2189

Tabla 3: Composición de la colección WebKB

Clase	Entrenamiento	Prueba
course	620	310
faculty	750	374
project	336	168
student	1097	544
Total	2803	1396

etc. El método propuesto soluciona estos dos problemas, y por tal motivo obtiene los mejores resultados. En primer lugar este clasificador toma en cuenta la importancia que tiene un término para cada clase y no para el documento. El segundo problema lo soluciona porque no considera las frecuencias de términos que aparecen en todas las clases. La figura 4 y 5 muestran la precisión y recall para esta colección.

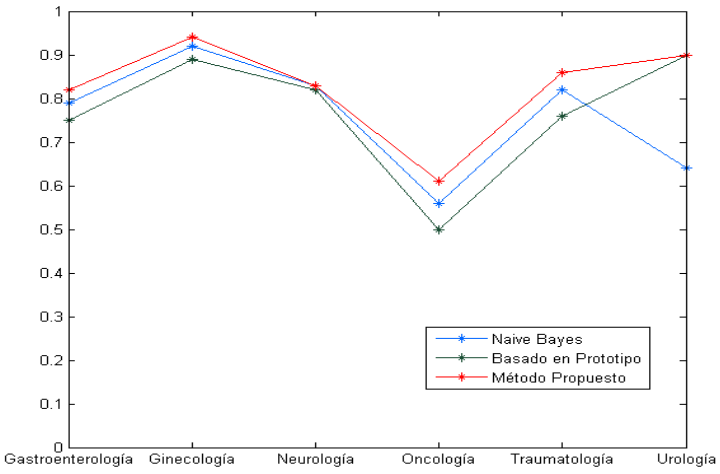

Figura 4: Precisión para la colección Historias Clínicas

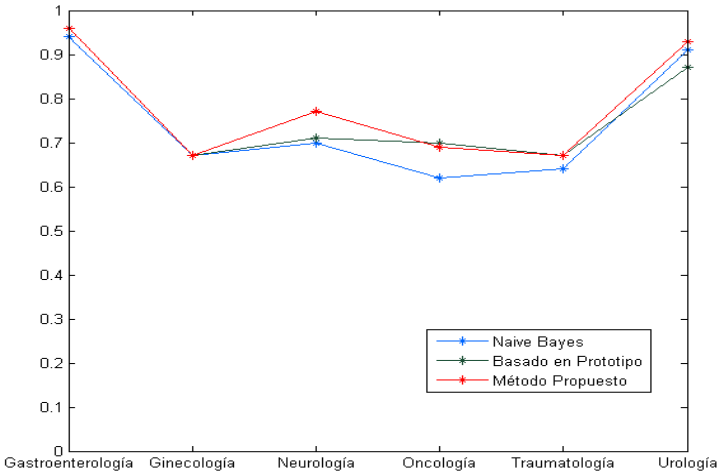

Figura 5: Recall para la colección Historias Clínicas

5.2.2. Evaluación de la colección R(8) y WebKB

De acuerdo a los resultados obtenidos, los clasificadores tienen mejor rendimiento para las clases que tienen más documentos de entrenamiento, esto se puede observar en las precisiones para las clases *acq* y *earn* de la colección R(8).

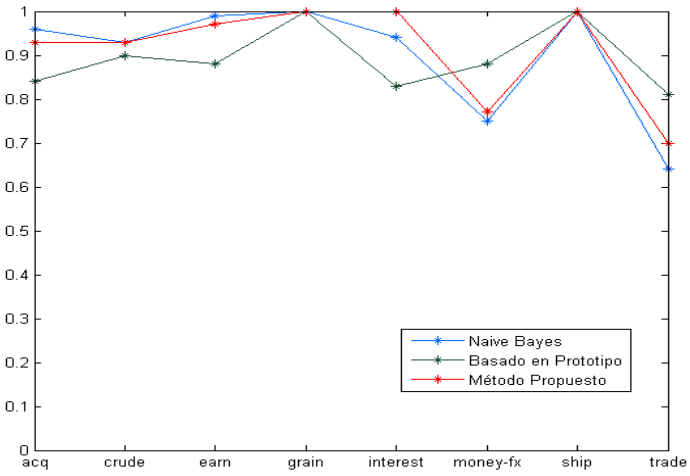

Figura 6: Precisión para la colección R(8)

Al igual que en la colección R(8) se mantiene la tendencia de los clasificadores, es decir, el clasificador propuesto arroja los mejores resultados. Sin embargo el desempeño en general de los clasificadores ha disminuido, esto se debe a que la cantidad de documentos de entrenamiento es mucho menor en comparación a la colección R(8). En la figura 7 se muestra la precisión para la colección WebKB.

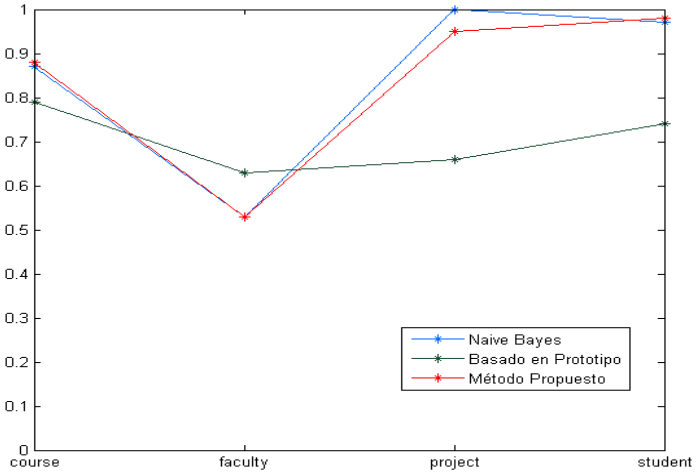

Figura 7: Precisión para la colección WebKB

En la figura 8 se muestra en resumen los resultados obtenidos por los tres métodos de clasificación, el método propuesto en este artículo supera, o en el peor de los casos, iguala el rendimiento de los demás métodos.

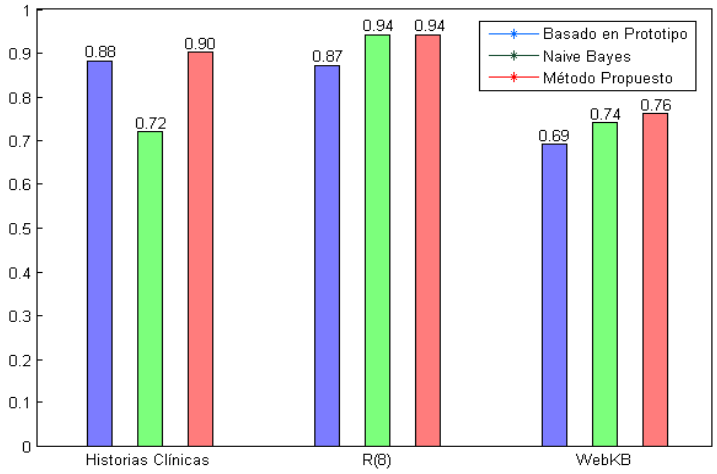

Figura 8: Exactitud para cada colección

6. Conclusiones y Trabajos Futuros

En este trabajo se presentó un método de clasificación de documentos aplicado a historias clínicas el cual mejora los resultados del método Basado en Prototipos y Naive Bayes. Este método consta de dos etapas, en la primera se utilizan algunas técnicas de procesamiento de lenguaje natural para extraer las características principales de las historias clínicas, así como también para la parametrización de las mismas. En la segunda etapa se hallan las clases más similares al documento a clasificar, por medio del método Basado en Prototipo utilizando un umbral de similitud. Posteriormente con el método de Naive Bayes se elige una de las clases más similares.

Los puntos más importantes a resaltar del presente trabajo son, en primer lugar, que el método propuesto obtuvo resultados aceptables en la clasificación automática de documentos (colección Historias Clínicas, R(8) y WebKB). En segundo lugar, que la mayoría de los errores surgieron generalmente en las clases que contaban con una menor cantidad de documentos de entrenamiento. Sin embargo, quedan por mejorar algunos aspectos, como por ejemplo, contar con características que puedan definir muy bien las categorías a las que pertenecen.

En el presente trabajo también se proporciona una comparación entre los clasificadores anteriormente mencionados y las combinaciones de estos mismos. Estas comparaciones nos dan la conclusión de que las combinaciones de estos métodos pueden mejorar los resultados obtenidos por los métodos individuales. Entre los principales trabajos futuros se encuentran: (1) Experimentar con otros métodos de parametrización para las historias clínicas, como por ejemplo a través de n-gramas. (2) Establecer otro tipo de umbral de similitud y en base a este elegir las clases más similares, para luego hacer la clasificación.

Referencias

[1] Álvarez, J. D., *Clasificación Automática de Textos usando Reducción de Clases basada en Prototipos*, Tesis de Maestría, Instituto Nacional de Astrofísica, Óptica y Electrónica, México, 2009.

[2] Anguiano, E., *Naive Bayes Multinomial para Clasificación de Texto Usando un Esquema de Pesado por Clases*, 2009, 1–3.

[3] Cardoso, A. M., *Improving Methods for Single-label Text Categorization*, Ph.D. thesis, Universidade Técnica de Lisboa, Portugal, 2007.

[4] Cardoso, A., and Arlindo Oliveira, *Semi-supervised single-label text categorization using centroid-based classifiers*, ACM (2007), 844–851.

[5] Cardoso, A., and Arlindo Oliveira, *Empirical evaluation of centroid-based models for single-label text categorization*, 2006, 3–7.

[6] Coyotl, R. M., *Clasificación automática de textos considerando el estilo de redacción*, Tesis de Maestría, Instituto Nacional de Astrofísica, Óptica y Electrónica, México, 2007.

[7] Debole, F., and Fabrizio Sebastiani, *Supervised term weighting for automated text categorization*, ACM (2003), 784–788.

[8] Figuerola C., José Berrocal Alonso Berrocal, Ángel Zazo and Emilio Rodríguez, *Algunas Técnicas de Clasificación Automática de Documentos*, 2004, 1–3.

[9] García, M. A., and Alfonso Ureña, *S-QUAMUS: Un Sistema de Búsqueda de Respuestas Multilinge*, 2006, 9–10.

[10] Gerard Salton, A. Wong, and C. S. Yang, *A vector space model for automatic indexing*, Communication of the ACM, 1975, 613–620.

[11] Gisbert, J. A. and Villanueva, Enrique, *Médicina legal y toxicología*, 2004, 102–103.

[12] Han E. H., and George Karypis, *Centroid-Based Document Classification: Analysis & Experimental Results*, 2000, 2–4.

[13] Olszewski, R., *Bayesian classification of triage diagnoses for the early detection of epidemics*, Proceedings of the FLAIRS Conference, 2003, 412–416.

[14] Porter, M. F., *An Algorithm for Suffix Stripping*, Program, vol.14, no. 3, 130-137, 1980.

[15] Ramirez G., Manuel Montes and Luis Villaseñor, *Enhancing Text Classification by Information Embedded in the Test Set*, Springer (2010), 627–637.

[16] Robertson S., *Understanding inverse document frequency: On theoretical arguments for IDF*, Journal of Documentation, 2004.

[17] Sebastiani, F., *Machine learning in automated text categorization*, ACM **34** (2002), 1–47.