Trabajo seleccionado del CoNaIISI 2016

Clasificación automática de textos periodísticos usando SVM

C. Javier Izetta Riera[1]
Juan G. Salinas[2]

[1]E-mail: javierizetta@gmail.com
[2]E-mail: juansalinas90@gmail.com

Universidad Nacional de Jujuy

RESUMEN

En los últimos años el periodismo pasó de su formato clásico de publicación al electrónico. Así, las webs de noticias están obligadas a mejorar sus prestaciones mediante una mejor organización de la información disponible para el lector. En este trabajo se propone abordar la clasificación automática de textos periodísticos digitales a través del Aprendizaje Automatizado. Se presentan dos clasificadores de textos periodísticos basados en Support Vector Machine junto con dos técnicas nuevas de reducción de dimensionalidad del espacio de características. Estos clasificadores fueron evaluados con distintas colecciones de noticias extraídas de páginas webs demostrando un buen desempeño.

ABSTRACT

Recently the journals have begun to publish their online versions. This new format requires better ways to organize and present the information to readers. This article introduces a new approach based on Machine Learning in order to classify online news. Two classifiers using the Support Vector Machine method and two new dimensionality reduction techniques are presented. A good performance has been obtained when both classifiers were evaluated on several news collections extracted from different newspaper websites.

PALABRAS CLAVES:

Clasificación Automática de Textos, Support Vector Machine

INTRODUCCIÓN

Los avances tecnológicos junto a la reducción del costo de almacenamiento provocaron un aumento en la disponibilidad de información en formato digital. Además, la información que circula por la web crece exponencialmente con el paso del tiempo y con ello surge la necesidad de organización y clasificación de la misma. Es de esta manera que la Categorización Automática de Texto (CAT) surge de la necesidad de desarrollar herramientas que faciliten la manipulación de un gran volumen de información y tiene como objetivo hallar una función óptima de clasificación de documentos a partir de atributos constituidos por palabras que describen cada categoría específica [1]. Ya durante los 80´s tuvieron lugar soluciones basadas en reglas genera-

das manualmente por expertos, denominadas “Sistemas Basados en Conocimiento”. Una solución sencilla pero que requería un gran esfuerzo humano a la hora de la generación de las reglas. Durante los 90´s con otro tipo de perspectiva, se introducen soluciones que conducen a la CAT como un problema de clasificación supervisada, es decir, a partir de una muestra de documentos previamente etiquetados como pertenecientes a una clase o categoría, se procede a la extracción del conocimiento necesario para la clasificación automática de nuevos documentos. Los métodos computacionales desarrollados para tal fin forman parte de lo que se conoce como Aprendizaje Automatizado (AA). Con esta metodología se reduce considerablemente la intervención humana, la cual solo queda delegada a etapas de diseño. A partir de esta etapa diversos algoritmos de AA, fueron utilizados para dar solución a la problemática de CAT. Se pueden destacar aquellos con muy buenos resultados como Redes Neuronales Artificiales [2], Árboles de Decisión [3], Naive Bayes [4] y K-vecinos más cercanos [5]. Un método que alcanzó gran interés en los últimos años dentro del área de AA son las Máquinas de Vectores Soporte (SVM. por sus siglas en inglés Support Vector Machine) y es posible encontrar diversos trabajos que evidencian que constituyen una buena solución a una amplia gama de problemas de clasificación, demostrándose sobre todo buen desempeño [6] [7]. En este trabajo se propone utilizar el método SVM para la clasificación automática de textos periodísticos extraídos de webs del noroeste argentino.Para textos en español, en la literatura existen algunas propuestas que aplican SVM a CAT, entre ellas pueden mencionarse el trabajo de Varguez Moo y colaboradores [8] que destaca la robustez del método SVM en la clasificación de documentos, el trabajo de Villasana y colaboradores [9] que demuestra el excelente desempeño de SVM y el uso de un kernel de cadenas aplicado a la CAT y el trabajo de Hidalgo y colaboradores [10] que realiza una evaluación comparativa de distintos algoritmos de aprendizaje en CAT obteniendo los mejores resultados con SVM.

MÁQUINAS DE VECTORES SOPORTE

El concepto de Máquinas de Vectores Soporte (SVM) se introduce en los años 90´s por Vapnik y colaboradores [11]. En sus principios solo se destinó a resolver problemas de clasificación del tipo binaria aunque posteriormente también ha llevado a la resolución problemas de multiclasificación. Dentro de las técnicas y herramientas para los problemas de clasificación, SVM han ganado popularidad por su capacidad de generalización a la hora de clasificar nuevos elementos con un bajo costo computacional. Desde un punto de vista experimental e interpretativo, SVM busca construir un separador lineal de clases o un hiperplano en el espacio de características original. Así, el clasificador obtenido constituye un modelo que servirá para predecir la clase de nuevos casos. Cuando el conjunto de datos es separable linealmente (Figura 1a), una manera formal de describir el método SVM, es la siguiente:

Se parte de un conjunto de ejemplos muestrales $S=\{x_1,x_2,...,x_n\}$, todos pertenecientes a un espacio característico $S\subset X\subseteq R^m$, (m indica la dimensión del espacio muestral o espacio de características), de los cuales algunos pertenecen a la clase de muestras positivas (+1) y otros a la clase de muestras negativas (-1). Entonces cada ejemplo de entrenamiento se define a partir de un par (x_i, y_i) $con\ x_i \in R^m\ e\ y_i \in$ {+1,-1}, de manera que el conjunto de entrenamiento queda expresado como

$L=\{(x_1,y_1),\ldots,(x_1,y_1)\}$.

El objetivo de SVM en la clasificación binaria consiste en construir un hiperplano de dimensión $(m-1)$ que separe los ejemplos etiquetados con -1 de los etiquetados con +1 con un margen máximo. Ya que, como se aprecia en la Figura 1b, podrían existir infinitos hiperplanos que separen los ejemplos, SVM buscará aquel que lo haga con un máximo margen geométrico.

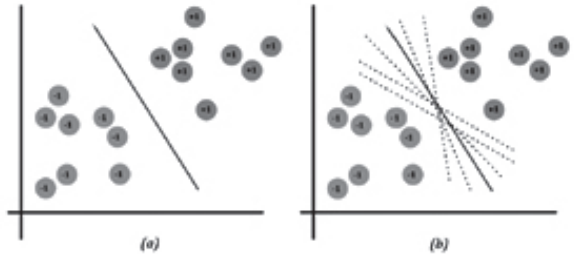

Figura 1: Hiperplanos de separación en un espacio bidimensional de un conjunto de ejemplos separables en dos clases.

SVM multiclase

En los últimos años se han desarrollado varios métodos para poder resolver problemas multiclase utilizando una combinación apropiada de clasificadores binarios. Las estrategias más utilizadas para poder aplicar SVM en problemas multiclase son: Uno Contra Uno (OVO-SVM) y Uno Contra Todos (OVA-SVM). Ambas consisten en convertir un problema de múltiples clases a varios problemas de dos clases, procediendo de la siguiente manera:

Uno Contra Todos (OVA-SVM): Para un problema con *c* clases, se construyen *c* clasificadores binarios SVM. El *i*-ésimo SVM es entrenado usando todos los ejemplos re-etiquetados de manera que la *i*-ésima clase es positiva y las demás clases son negativas.

Uno Contra Uno (OVO-SVM): En este caso, para un problema con *c* clases, se construye $(c-1)/2$ clasificadores binarios SVM, cada uno para discriminar un par de clases. Cada clasificador es entrenado solo con los ejemplos de las dos clases.

CLASIFICADORES PROPUESTOS

La construcción de los clasificadores propuestos en este trabajo se abordó a través de dos etapas claramente delimitadas. La primera etapa, a la que suele llamarse "etapa de entrenamiento", se inicia con la recopilación manual de una serie de textos periodísticos de diarios digitales del noroeste argentino extraídos de la web (documentos de entrenamiento). Esta colección se procesa para lograr una representación adecuada para el entrenamiento de los clasificadores. Luego se realiza una reducción del conjunto de características generado por la colección (reducción de dimensionalidad) con el fin de mejorar el rendimiento durante el aprendizaje de los clasificadores. Una vez que éstos fueron entrenados, tiene lugar la segunda etapa, llamada "etapa de prueba", que consiste en la evaluación del desempeño de los clasificadores con nuevos documentos no considerados durante la etapa anterior. En la Figura 2 es posible observar con más detalle los pasos para la construcción de los clasificadores propuestos.

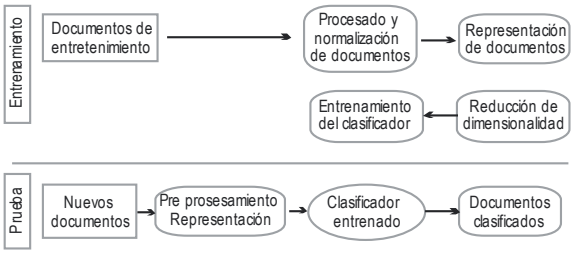

Figura 2: Esquema básico de la construcción de los clasificadores.

PREPROCESADO Y NORMALIZACIÓN DE DOCUMENTOS

En este paso se busca definir tokens o términos, para ello, en los clasificadores propuestos un token queda conformado por aquella cadena de caracteres delimitados por espacios en blanco. Además, en esta instancia, se descartan aquellos caracteres tales como símbolos y números ya que no aportan información alguna para la clasificación. Y también aquellos tokens o términos identificados como "palabras de parada" (stopwords), este conjunto de palabras está constituido por preposiciones, artículos, pronombres, conjunciones, contracciones y ciertos verbos y adverbios. En este trabajo se usó para tal fin el conjunto de stopwords para el español definido por el proyecto Snowball disponible en [12].

REPRESENTACIÓN DE DOCUMENTOS

Este paso consiste en la transformación de los documentos en una representación adecuada para que el algoritmo de aprendizaje sea capaz de procesarlos. En este trabajo

se propone utilizar para la representación de los textos periodísticos el modelo vectorial propuesto por Salton [13] y el esquema de pesado TF-IDF (Term Frequency - Inverse Document Frequency) [14]. En CAT este modelo de representación es uno de los más utilizados debido a sus altas prestaciones cuando se combina a esquemas de pesado y normalización de longitud de documentos [15]. En el modelo vectorial los documentos son formalmente representados a través de vectores cuya dimensión estará dada por la cantidad de términos del vocabulario generado por la colección de documentos. Cada componente del vector representa la importancia que tiene ese término en el documento y en la colección. Salton propone en [15] calcular los pesos mediante la combinación de la frecuencia relativa de los términos (TF) con la frecuencia inversa en los documentos (IDF), de manera que se tiene:

$TF{-}(t_j,d_i) = f_{ij} \times \log(N/df(t_j))$ (1)

Dónde t_j es el número de documentos en los que aparece el término, f_{ij} es la frecuencia del término t_j en el documento d_i y *N* es la cantidad de documentos en la colección.

REDUCCIÓN DE DIMENSIONALIDAD

Es importante aclarar que los dos pasos anteriores (preprocesado y representación documentos) se realizaron de la misma manera para la construcción de los dos clasificadores propuestos. En este paso se proponen dos variantes para la reducción de dimensionalidad. Una basada en la selección de un subconjunto del conjunto de términos originales, alternativa a la que denominaremos CATST. Y otra basada en la transformación del conjunto de términos originales a la que llamaremos CATLT. A continuación se exponen ambas alternativas:

PROPUESTA CATST

La ley del mínimo esfuerzo de Zipf [16], comprueba que en una colección de documentos coexisten términos muy pocos frecuentes y específicos para determinados documentos, junto con aquellos términos muy frecuentes que representan la colección de documentos en general. En base a esta ley, Luhn [17] afirma que existe un rango de términos que son relevantes para un determinado documento, cuando la tarea es la recuperación de documentos a través de una consulta. Esta misma idea se puede aplicar a CAT, es decir, es posible hallar un rango de términos relevantes para cada categoría. En un problema de clasificación de texto lo que se pretende es encontrar términos que tengan el mayor poder de discriminación entre las categorías. Esto implica centrarnos en términos que sean característicos de cada grupo de documentos pertenecientes a cada categoría, es decir, términos de frecuencia media que no son exclusivamente específicos de uno o muy pocos documentos ni absolutamente generales a toda la colección de documentos. Para encontrar este rango de términos proponemos realizar los siguientes pasos:

1. Particionar el conjunto de términos originales ordenados de manera decreciente según su frecuencia, en 4 partes iguales.

2. Tomar como punto de partida para la determinación del rango, aquel término que se ubica en la parte media del primer cuarto tal como se puede apreciar en la Figura 3.

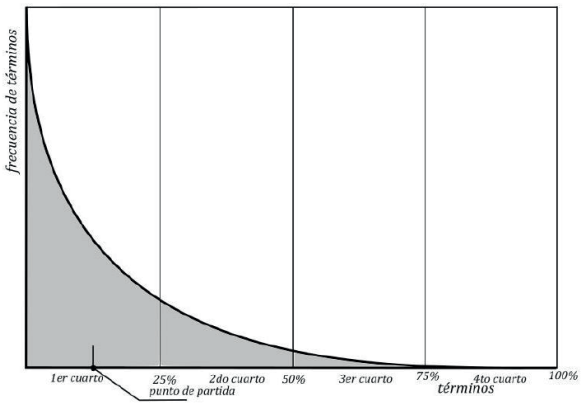

Figura 3: Punto de partida, propuesta CATST.

3. Tomar los términos correspondientes al 10, 20, 30 y sucesivamente hasta un 90% hacia la izquierda (cut-on) y derecha (cut-off) de este punto de partida, tal como se aprecia en la Figura 4 para formar nueve rangos candidatos de términos.

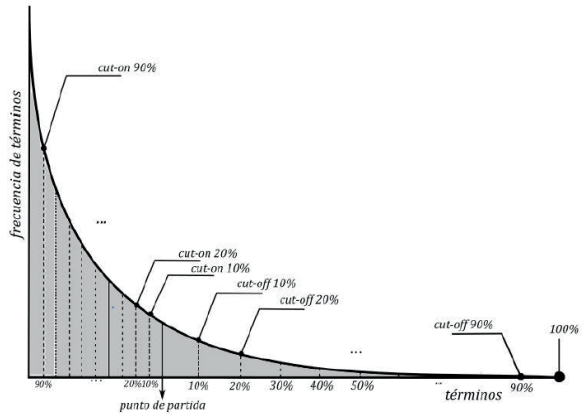

Figura 4: Cortes para los 9 rangos candidatos, propuesta CATST.

4. Entrenar una SVM utilizando cada rango candidato.

5. Evaluar y seleccionar el rango que mejor desempeño obtenga.

PROPUESTA CATLT

Otra alternativa que se propone en este trabajo, es utilizar una técnica de reducción de dimensionalidad basada en la transformación del conjunto de términos originales a través del concepto de lematización o por su terminología en inglés, stemming. Los algoritmos de lematización de términos son capaces de extraer prefijos y sufijos de palabras que son literalmente diferentes, pero que tienen una raíz en común y que pueden ser consideradas como un mismo término. Cada palabra es "truncada" a su lema o raíz equivalente.

Para tal fin en este trabajo se utilizó una adaptación al español del algoritmo de Porter [18] [19]. A pesar de que al transformar el espacio de términos en un espacio de raíces este conjunto se reduce notoriamente, se debería considerar solo aquellas raíces que tengan mayor poder de discriminación entre las categorías. A diferencia de los términos originales, cuando se trabaja con raíces, estas últimas tienen mayor poder de discriminación cuando su frecuencia es alta. Para ello proponemos encontrar un rango de raíces de la siguiente manera:

1. Ordenar las raíces en forma decreciente según su frecuencia de aparición.

2. Tomar las raíces correspondientes al 10, 20, 30 y sucesivamente hasta un 90% a partir de aquella raíz cuya frecuencia de aparición sea máxima, para formar nueve rangos candidatos, como se muestra en la Figura 5.

3. Entrenar una SVM utilizando cada rango candidato.

4. Evaluar y seleccionar el rango que mejor desempeño obtenga.

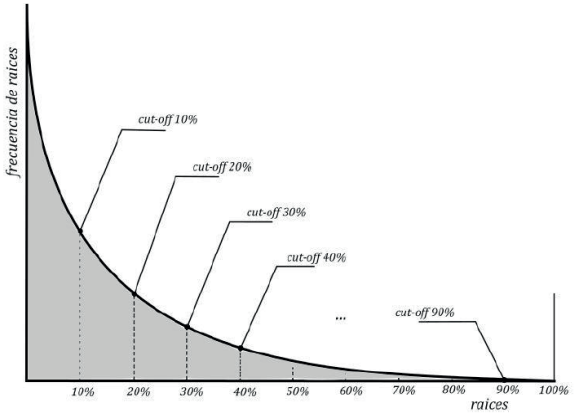

Figura 5: Cortes para los 9 rangos candidatos, propuesta CATLT.

ENTRENAMIENTO Y PRUEBA DEL CLASIFICADOR

El entrenamiento y prueba de los clasificadores se llevó a cabo mediante un proceso que realiza a partir de dos bucles anidados (Figura 6).

En el bucle externo (recuadro externo en líneas punteadas) se realiza 30 veces la partición de la colección de documentos en un subconjunto de documentos para entrenamiento, seleccionando aleatoriamente un 70% del total de documentos. El subconjunto de documentos restante (30% del total de documentos), es utilizado para prueba. De esta manera es posible obtener una mejor estimación del desempeño de los clasificadores propuestos. En el bucle interno (recuadro interno en líneas punteadas) se entrena una SVM por cada uno de los nueve rangos candidatos. Se evalúa el desempeño de cada rango a partir del subconjunto de documentos de prueba generado por el bucle externo. Al finalizar las iteraciones de ambos bucles se selecciona la SVM entrenada a partir del rango con tasa de error de clasificación más baja.

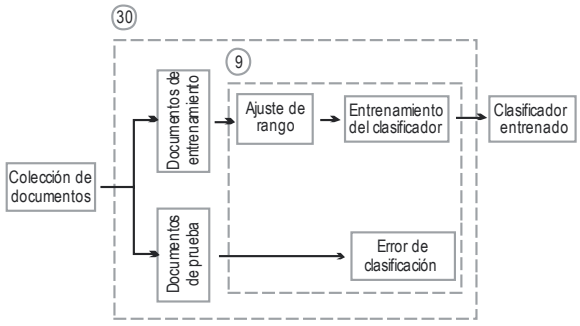

Figura 6: Entrenamiento y prueba de los clasificadores propuestos.

EXPERIMENTACIÓN

MEDIDA DE DESEMPEÑO

Para determinar qué tan bueno es un clasificador es posible recurrir a una medida de precisión denominada tasa de error. Esta medida consiste en considerar un éxito cuando una instancia es clasificada correctamente, y como un error cuando ocurre lo contrario. Dado un conjunto de documentos

$D = \{d_1, \ldots, d_i, \ldots, d_n\}$ una estimación del error del clasificador f sería:

$error\ (f) = N_e / N_t$ (2)

Donde N_e representa el número de clasificaciones erróneas de f sobre D y N_t el número total de documentos clasificados.

COLECCIONES DE NOTICIAS

El entrenamiento y evaluación de los clasificadores propuestos se realizó a partir de colecciones de noticias confeccionadas manualmente. El motivo de esta decisión, se debe a la falta de disponibilidad de algún repositorio que contenga alguna con categorías explicitas y en idioma español. Además utilizando los textos originales extraídos de portales de noticias del NOA, se pretende que los clasificadores "aprendan" el estilo de redacción de la zona. Las noticias se obtuvieron a partir de los periódicos digitales más leídos de la región noreste de Argentina, como ser: Todo Jujuy; Jujuy al Momento; Jujuy al día; Notinor; Pregón; El Tribuno de Jujuy; El Tribuno de Salta; Informate Salta; Nuevo Diario de Salta; Que pasa Salta; El Intransigente; El Diario Noticias y La Gaceta [20]. Cada noticia seleccionada aleatoriamente corresponde a un período comprendido entre Octubre de 2015 y Marzo de 2016. Se crearon cuatro colecciones de documentos, la Tabla 1 muestra los detalles de cada colección. La creación de las colecciones tiene la finalidad de evaluar el desempeño de cada clasificador propuesto en diferentes situaciones. Por un lado, el clasificador puede enfrentarse a un problema de clasificación binaria o multiclase, en esta última situación se aplicó la técnica Uno Contra Uno (OVO-SVM) para extender el método SVM a problemas multiclase. Por otro lado, las categorías podrían tener muchos términos en común, adicionando complejidad a la colección. Por ejemplo en la colección C3PES, que tiene textos informativos sobre Política y Economía, ambas categorías suelen utilizar un vocabulario muy similar. Términos tales como "gobierno", "medidas", "funcionarios", "nacional", "reunión", etc.; podrían estar presentes tanto en textos políticos como económicos.

Tabla1

Colecciones de noticias. N= cantidad total de noticias, T=cantidad total de términos y C=categorías.

Nombre	N	T	C
C2PD	200	9084	Policial – Deportes
C2PE	200	10611	Política – Economía
C3PDT	300	13029	Policial–Deportes Tecnología
C3PES	300	13696	Política – Economía – Salud

RESULTADOS Y EVALUACIÓN

Colecciones binarias (C2PD y C2PE): En la Tabla 2 se exponen los resultados obtenidos a partir de las experimentaciones realizadas sobre las colecciones C2PD y C2PE, en particular, se muestra el error medio de clasificación de 30 corridas de cada clasificador con el mejor rango de términos o raíces encontrado.

Tabla 2

Resultados de los clasificadores con el mejor rango de términos o raíces encontrado sobre las colecciones C2PD y C2PE.

Colección C2PD:	CATST	CATLT
Cantidad de características total	8733	5202
Error medio de clasificación	0.01	0.0083
Cant. de características mejor rango	7859	4161
Colección C2PE:	CATST	CATLT
Cantidad de características total	10194	5559
Error medio de clasificación	0.1589	0.0733
Cant. de características mejor rango	7136	5003

En primer lugar se puede observar que la propuesta CATLT al aplicar lematización de términos trabaja con una cantidad de características considerablemente menor que CATST. En segundo lugar al analizar el desempeño de los clasificadores (error medio de clasificación), se puede observar que CATLT (propuesta basada en lematización), obtiene una tasa de error menor a CATST (propuesta basada en la selección de un subconjunto de términos originales). La razón es que este último incluye en el mejor rango encontrado algunos términos con poco poder de discriminación entre las clases. Esto se debe a la dificultad de encontrar un rango que solo contenga términos altamente discriminativos. Para ello, lo que se busca son los términos con una frecuencia de aparición media, ya que éstos son los más informativos para cada clase. Aun así, no todos estos términos van a aportar buena información, llevando en algunos casos a un entrenamiento menos eficaz. Por el contrario en CATLT al trabajar con raíces en vez de términos, el proceso de ajuste del rango es más sencillo ya que solo se debe descartar las raíces con menor frecuencia de aparición. Este proceso lleva a encontrar un rango de raíces con un alto poder de discriminación entre las clases, favoreciendo el entrenamiento del clasificador. Este comportamiento se puede observar en ambas colecciones binarias a pesar que la diferencia entre ellas es que la colección C2PE es más compleja de clasificar debido a que contiene muchos términos en común en sus categorías. En la Figura 7, se muestran para las colecciones C2PE y C2PD los errores de clasificación obtenidos de 30 corridas en un diagrama de cajas, que corresponden a los mejores rangos encontrados por cada clasificador. Se puede observar en la gráfica que CATLT produce los errores más bajos y esta diferencia es significativa.

Colecciones multiclase (C3PDT y C3PES): Por un lado, con la finalidad de observar si el desempeño de los clasificadores se ve afectado al aplicar la técnica OVO (técnica para extender SVM a problemas multiclase), se realizó una evaluación del comportamiento de los clasificadores propuestos en dos colecciones de tres categorías. Se puede observar en la Tabla 3 los resultados de los clasificadores con el mejor rango de términos o raíces encontrado, éstos sugieren que la técnica OVO-SVM no afecta en forma significativa el desempeño de SVM. Debido a que en las experiencias realizadas en estas colecciones ambos clasificadores muestran comportamientos muy similares a las experiencias realizadas en las colecciones binarias, en términos generales.

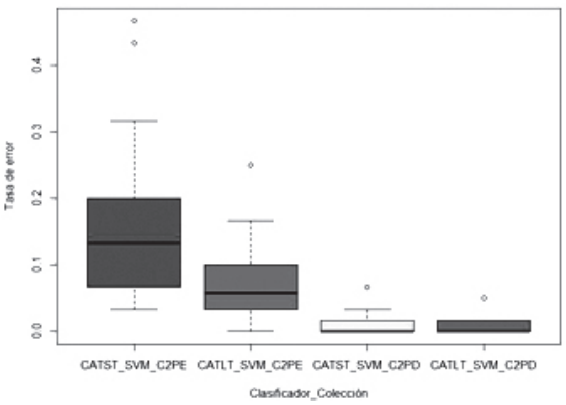

Figura 7: Errores de clasificación obtenidos de 30 corridas en colecciones C2PE y C2PD.

Se pudo comprobar que los clasificadores propuestos son robustos, ya que el desempeño de los mismos se mantiene en las distintas situaciones.

Tabla 3

Resultados de los clasificadores con el mejor rango de términos o raíces encontrado sobre las colecciones C3PDT y C3PES.

Colección C3PDT:	CATST	CATLT
Cantidad de características total	12634	7107
Error medio de clasificación	0.0192	0.0159
Cant. de características mejor rango	11371	5685
Colección C3PES:	CATST	CATLT
Cantidad de características total	13264	6961
Error medio de clasificación	0.1407	0.0707
Cant. de características mejor rango	11938	3480

CONCLUSIONES Y TRABAJOS FUTUROS

En este trabajo se presentó dos clasificadores automáticos de textos periodísticos del noroeste argentino usando SVM. Los clasificadores desarrollados implementan dos técnicas propuestas para la reducción de dimensionalidad del espacio de características, denominadas CATST (basada en la selección de un sub conjunto de características) y CATLT (basada en lematización).

Después de demostrar el buen desempeño de ambos clasificadores en las diferentes colecciones creadas, nuestros resultados sugieren que:

La propuesta CATLT produce tasas de errores más bajas que la propuesta CATST en todas las experimentaciones realizadas. Dado que el proceso de ajuste del rango de CATLT, al trabajar con raíces, resulta más sencillo que al trabajar con términos como en CATST. Esto se debe a que una raíz con frecuencia alta implica que aparece muchas veces en documentos pertenecientes a una determinada clase de la colección, siendo esa raíz representativa para esa clase. Entonces, para encontrar un rango óptimo solo se debe descartar las raíces con menor frecuencia de aparición. Por el contrario en CATST, encontrar un rango de términos informativos para la clasificación implica centrarse en términos de frecuencia media, que no son exclusivamente específicos de uno o muy pocos documentos ni absolutamente generales a toda la colección de documentos, aun así, no todos estos términos van a aportar buena información haciendo más complicada la determinación de este rango.

La técnica OVO-SVM no afecta en forma significativa el desempeño de SVM en las colecciones multiclase.

Para concluir, se pudo comprobar la robustez de los prototipos propuestos al mantener el buen desempeño en las distintas colecciones.

Varias vías están abiertas para continuar este trabajo, por supuesto se necesita una evaluación más en profundidad de los clasificadores propuestos incluyendo más colecciones y un análisis comparativo con otras técnicas de reducción de dimensionalidad. Además se podrían analizar otros aspectos como por ejemplo el ajuste de parámetros del método SVM.

Se pueden nombrar algunas direcciones en las cuales extender los clasificadores presentados en este trabajo, tales como:

Emplear otros esquemas de pesado para la representación de relevancia de un término dentro de la colección.

Modificar los prototipos propuestos utilizando otros métodos de clasificación, tales como Redes Neuronales Artificiales.

Para finalizar se podría extender los prototipos propuestos a otros problemas de clasificación, como la clasificación de páginas web, o detección de correos no deseados.

REFERENCIAS

[1] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), 1-47.

[2] Hornik, K.; Stinchcombe, M.; White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5), 359-366.

[3] Quinlan, J.R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.

[4] Lewis, D.D.; Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. *In Third annual symposium on document analysis and informa-*

tion retrieval, 33, 81-93.
[5] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2), 69-90.
[6] Drucker, H.; Wu, D.; Vapnik, V.N. (1999). Support vector machines for spam categorization. Neural Networks, *IEEE Transactions*, 10(5), 1048-1054.
[7] Osuna, E.; Freund, R.; Girosi, F. (1997). Training support vector machines: an application to face detection. In Computer vision and pattern recognition. *Proceedings., 1997 IEEE computer society conference*, 130-136.
[8] Varguez-Moo, M.; Uc-Cetina, V.; Brito-Loeza, C. (2014). Clasificación de documentos usando Máquinas de Vectores de Apoyo. *Abstraction and Application Magazine*, 6.
[9] Villasana, S.; Seijas, C.; Caralli, A.; Jiménez, J.; Pacheco, J. (2008). Categorización de documentos usando máquinas de vectores de soporte. *Revista Ingeniería UC*, 15(3), 45-52.
[10] Hidalgo, J.G.; Sanz, E.P.; García, F.C.; de Buenaga Rodríguez, M. (2003). Categorización de texto sensible al coste para el filtrado de contenidos inapropiados en Internet. *Procesamiento del lenguaje natural*, 31, 13-20.
[11] Vapnik, V.N., Vapnik. V. (1998) *Statistical learning theory*, New York: Wiley, 1.
[12] Stopword Spanish Snowball. URL:http://snowball.tartarus.org/algorithms/spanish/stop.txt
[13] Salton, G. (1971). The SMART retrieval system. *Experiments in automatic document processing*.
[14] Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval*. Reading: Addison-Wesley.
[15] Salton, G.; Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513-523.
[16] Zipf, G.K. (2016). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.
[17] Luhn, H.P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2), 159-165.
[18] Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.
[19] Bordignon, F.R.A.; Panessi, W. (2011). Procesamiento de variantes morfológicas en búsquedas de textos en castellano. *Revista Interamericana de Bibliotecología*, 24(1).
[20] Todo Jujuy URL: http://www.todojujuy.com/, Jujuy al momento URL: http://www.jujuyalmomento.com/, Jujuy al día URL: http://www.jujuyaldia.com.ar/, Notinor URL: http://notinor.com/jujuy/, Pregón URL: http://www.pregon.com.ar/, El Tribuno de Jujuy URL: http://www.eltribuno.info/jujuy/, El Tribuno de Salta URL: http://www.eltribuno.info/salta/, Informate Salta URL: http://informatesalta.com.ar/, Nuevo Diario de Salta: URL: http://www.nuevodiariodesalta.com.ar/, Que Pasa Salta. URL: http://www.quepasasalta.com.ar/, El Intransigente. URL: http://www.elintransigente.com/, El Diario Noticias. URL: http://www.eldiarionoticias.com.ar/, La Gaceta: URL: http://www.lagaceta.com.ar/.