Aplicación de redes neuronales artificiales en la clasificación de textos académicos según disciplina: Biometría, Filosofía y Lingüística informática.

Application of artificial neural networks in the classification of academic texts based on the disciplines of: Biometrics, Phylosphy and Computational Linguistics.

Celina Beltrán
Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina
beltranc36@yahoo.com.ar

Abstract

At present, there is a wide volume of documents in electronic form and which are easy to access from the web. Document classification is one of the essential tasks to make such amounts of information useful. The goal of automatic text classification is to categorize documents within a fixed number of predefined categories according to their content.

This work is intended to propose the Artificial Neural Network model with supervised learning: Perceptrón Multicapa, which uses the disciplanary area and the characterization of texts based on the distribution of frecuencies of the morphosyntactic categories as a sort and selection criterion.

The efectiveness of this network has been proved for the prediction of the disciplanary area, which determines the values of the set of parameters that correspond to this model.

The correct classification percentage in each discipline was 100%, 100%, 93.3% for Biometrics, Phylosophy and Computational Linguistics, respectively; whereas the global error was 2.2%

Key words: Neural neworks, multinomial logistic regression, multivariate análisis, text classification.

Resumen

En la actualidad existe un volumen de documentos dispuesto en formato electrónico de fácil acceso en la web. La clasificación de documentos es una de las tareas imprescindibles para brindar utilidad a tanta información. El objetivo de la clasificación automática de texto es categorizar documentos dentro de un número fijo de categorías predefinidas en función de su contenido.

En este trabajo se propone el modelo de Redes Neuronales Artificiales con aprendizaje supervisado: Perceptrón Multicapa, utilizando como criterio de clasificación el área disciplinar y la caracterización de los textos basada en distribución de frecuencias de las categorías morfo-sintácticas. Se comprobó la efectividad de esta red para la predicción del área disciplinar, determinando los valores del conjunto de parámetros correspondientes a este modelo. El porcentaje de clasificación correcta en cada disciplina fue 100%, 100% , 93.3% , para Biometría, Filosofía y Lingüística computacional respectivamente; mientras que el error global fue del 2.2%.

Palabras claves: Redes neuronales, Regresión logística multinomial, análisis multivariado, clasificación de textos.

1. INTRODUCCION

En la actualidad existe un gran volumen de documentos dispuesto en formato electrónico de fácil acceso en la web. La clasificación de documentos es una de las tareas imprescindibles para brindar utilidad a tanta información. El objetivo de la clasificación automática de texto es categorizar documentos dentro de un número fijo de categorías predefinidas en función de su contenido. Cuando se utiliza aprendizaje automático, el objetivo es aprender a clasificar a partir de ejemplos que permitan hacer la asignación a la categoría automáticamente. Durante el aprendizaje o entrenamiento del sistema se evalúan las condiciones de pertenencia a cada una de las categorías. Para realizar el entrenamiento es necesario disponer de conocimiento previo de expertos en forma de decisiones de categorización asignadas a cada uno de los documentos. Este conocimiento corresponde a un conjunto de documentos preclasificados de modo que el sistema pueda leer la categoría o grupo de pertenencia de cada uno de los documentos.

En este trabajo se propone el modelo de Redes Neuronales Artificiales con aprendizaje supervisado: Perceptrón Multicapa, utilizando como criterio de clasificación el área disciplinar y la caracterización de los textos basada en distribución de frecuencias de las categorías morfo-sintácticas.

2. MATERIAL Y METODOS

2.1. Diseño de la muestra

El marco muestral para la selección de la muestra está compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas, extraídos de internet pertenecientes a las disciplinas: Biometría, Lingüística informática y Filosofía. La unidad de muestreo fue el texto y la selección de la muestra se llevó a cabo empleando un diseño muestral estratificado con selección proporcional al tamaño, siendo la medida de tamaño el “número de palabras del texto”. Se seleccionaron 60 textos correspondiente a cada disciplina.

Las muestras de los tres estratos fueron evaluadas y comparadas respecto al número medio de palabras por texto. Esta comparación se requiere para evitar que la discriminación entre las disciplinas se vea afectada por el tamaño de los textos.

2.2. Etiquetado: Análisis morfológico de los textos

El software Smorph, analizador y generador morfosintáctico desarrollado en el Groupe de Recherche dans les Industries de la Langue (Universidad Blaise-.Pascal, Clermont II) por Salah Aït-Mokhtar (1998) realiza en una sola etapa la tokenización y el análisis morfológico. A partir de un texto de entrada se obtiene un texto lematizado con las formas correspondientes a cada lema (o a un subconjunto de lemas) con los valores correspondientes. Se trata de una herramienta declarativa, la información que utiliza está separada de la maquinaria algorítmica, en consecuencia, puede adaptarse a distintos usos. Con el mismo software se puede tratar cualquier lengua si se modifica la información lingüística declarada en sus archivos.

Smorph compila, minimiza y compacta la información lingüística que queda disponible en un archivo binario. Los códigos fuente se dividen en cinco archivos: Códigos ASCII, Rasgos, Terminaciones, Modelos y Entradas.

En el archivo **entradas**, se declaran los ítems léxicos acompañados por el modelo correspondiente. Este indicador de modelo oficia de enlace con el archivo modelos, en el que se especifica la información morfológica y las terminaciones que se requieren en cada ítem. El archivo **modelos**, es el que introduce la información correspondiente a los modelos de flexiones morfológicas, mientras que en el archivo **terminaciones** es necesario declarar todas las terminaciones que son necesarias para definir los modelos de flexión. Las etiquetas correspondientes a los rasgos morfológico-sintácticos son organizadas jerárquicamente en el archivo **rasgos**. Por último, en el archivo de códigos **ASCII** se especifican, entre otros, los caracteres separadores y las equivalencias entre mayúsculas y minúsculas.

El módulo post-smorph **MPS** es un analizador que recibe en entrada una salida Smorph (en formato Prolog) y puede modificar las estructuras de datos recibidos. Ejecuta dos funciones principales: la Recomposición y la Correspondencia, que serán útiles para resolver las ambigüedades que resulten del análisis de Smorph.

La información contenida en estos archivos es la presentada en Beltrán (2009) para implementar el etiquetador.

2.3. Diseño y desarrollo de la base de datos

El resultado del análisis de Smorph-Mps se almacena en un archivo de texto. Esta es la información que contendrá la base de datos.
Mediante una función definida en el sistema estadístico R se logra captar la información resultante del análisis morfológico y disponerla en una matriz de dimensión: tantas filas como cantidad de objetos lingüísticos tenga el texto y tantas columnas como ocurrencia+lema+valores. De esta manera se obtiene una base de datos con la estructura que se muestra en la tabla 1.

Tabla 1. Fragmento de la base de datos obtenida

MUESTRA	TEXTO	OCURRENCIA	LEMA	ETIQUETA
1	1	El	el	det
1	1	problema	problema	nom
1	1	de	de	prep
1	1	las	el	det
1	1	series	serie	nom
…	…	…	…	…
2	1	Uno	uno	pron
2	1	de	de	prep
2	1	los	el	det
2	1	agentes	agente	nom
2	1	que	que	rel
2	1	ha	haber	aux
…	…	…	…	…
3	1	permitió	permitir	v
3	1	el	el	det
3	1	análisis	análisis	nom
3	1	automático	automático	adj
…	…	…	…	…

Abreviaturas:
‘adj’: adjetivo ‘art’: artículo ‘nom’: nombre ‘prep’: preposición ‘v’: verbo ‘adv’: adverbio ‘cl’: clítico ‘aux’: auxiliar ´cop’: copulativo ‘pun’: signo de puntuación

Luego, a partir de esta base de datos por palabra (cada unidad o fila es una palabra analizada del texto), se confecciona la base de datos por documento que será analizada estadísticamente. Esta es una nueva base, donde cada unidad es el texto, que retiene la información de las variables indicadas en la tabla 2.a con la estructura presentada en la tabla 2.b.

Tabla 2.a. Variables de la base de datos por documento

CORPUS	Corpus al que pertenece el texto
TEXTO	Identificador del texto dentro del corpus
adj	proporción de adjetivos del texto
adv	proporción de adverbios del texto
cl	proporción de clíticos del texto
cop	proporción de copulativos del texto
det	proporción de determinantes del texto
nom	proporción de nombres (sustantivos) del texto
prep	proporción de preposiciones del texto
v	proporción de verbos del texto
otro	proporción de otras etiquetas del texto
total_pal	total de palabras del texto

Tabla 2.b. Fragmento de la base de datos para análisis estadístico

CORPUS	TEXTO	adj	adv	cl	cop	det	nom	prep	v	OTRO	TOTAL_PAL
1	1	0,11	0,02	0,02	0,04	0,16	0,26	0,18	0,09	0,11	185
1	2	0,13	0,00	0,05	0,04	0,13	0,25	0,18	0,08	0,15	110
1	3	0,09	0,03	0,06	0,03	0,15	0,26	0,14	0,10	0,14	181
…	…	…	…	…	…	…	…	…	…	…	…
2	28	0,08	0,01	0,02	0,03	0,16	0,32	0,21	0,09	0,09	186
2	29	0,10	0,00	0,03	0,04	0,17	0,28	0,18	0,09	0,11	141
2	30	0,10	0,03	0,01	0,03	0,19	0,27	0,16	0,10	0,11	183
…	…	…	…	…	…	…	…	…	…	…	…
3	28	0,06	0,03	0,05	0,04	0,16	0,22	0,17	0,16	0,11	192
3	29	0,05	0,01	0,02	0,03	0,16	0,19	0,12	0,24	0,19	138
3	30	0,07	0,01	0,04	0,02	0,16	0,21	0,17	0,19	0,13	157

2.4. Redes Neuronales Artificiales: El Perceptrón Multicapa

2.4.1. El modelo

Las redes neuronales son sistemas pertenecientes a una rama de la inteligencia artificial que emulan al cerebro humano. Requieren un entrenamiento en base a un conocimiento previo del entorno del problema. Una red neuronal es un sistema compuesto por un gran número de elementos básicos, agrupados en capas que se encuentran totalmente interconectadas y que serán entrenadas para reaccionar de una determinada manera a los estímulos de entrada.

Las redes neuronales constituyen naturalmente una técnica de modelización multivariada, es decir, pueden hacer predicciones de dos o más variables simultáneamente. Pueden realizar predicciones

tanto de variables continuas como discretas, utilizando las implementaciones apropiadas. En este trabajo son utilizadas para predecir el grupo o categoría de procedencia del texto en función de la distribución porcentual de las categorías morfológicas, información derivada del análisis automático de los mismos.

El Perceptrón Multicapa (MLP, por sus siglas en inglés "Multi-Layer Perceptron") tiene como objetivo la categorización o clasificación de forma supervisada. Para este trabajo se ha utilizado esta red aplicado a la clasificación de textos en tres áreas disciplinares. Utilizando el algoritmo de aprendizaje supervisado Backpropagation, la red aprende la relación entre la proporción de las distintas categorías morfosintácticas y la categoría de pertenencia (disciplina), con el propósito de lograr clasificar un nuevo texto para el cual se cuenta con el análisis morfológico pero se desconoce su área de pertenencia.

En esta aplicación se consideraron 60 textos de cada una de las disciplinas consideradas. Cada una de estas muestras fue dividida aleatoriamente en dos submuestras de igual tamaño de modo de utilizar una de ellas en la fase de entrenamiento de la red y la otra en la etapa de validación.

2.4.2. Arquitectura

Un perceptrón multicapa está compuesto por una capa de entrada, una capa de salida y una o más capas ocultas; aunque se ha demostrado que para la mayoría de problemas bastará con una sola capa oculta. En la figura 1 podemos observar un perceptrón típico formado por una capa de entrada con P neuronas, una capa oculta con L neuronas y una de salida con M neuronas. En este tipo de arquitectura, las conexiones entre neuronas son siempre hacia delante, es decir, las conexiones van desde las neuronas de una determinada capa hacia las neuronas de la siguiente capa; no hay conexiones laterales, ni conexiones hacia atrás. Este es, la información siempre se transmite desde la capa de entrada hacia la capa de salida. En dicho diagrama w_{ji} representa el peso de conexión entre la neurona de entrada i y la neurona oculta j, y v_{kj} es el peso de conexión entre la neurona oculta j y la neurona de salida k.

En esta aplicación las P neuronas de la capa de entrada corresponden a las proporciones de las P categorías morfológicas consideradas y la capa de salida estará constituida por las 3 neuronas que corresponden a las áreas disciplinares.

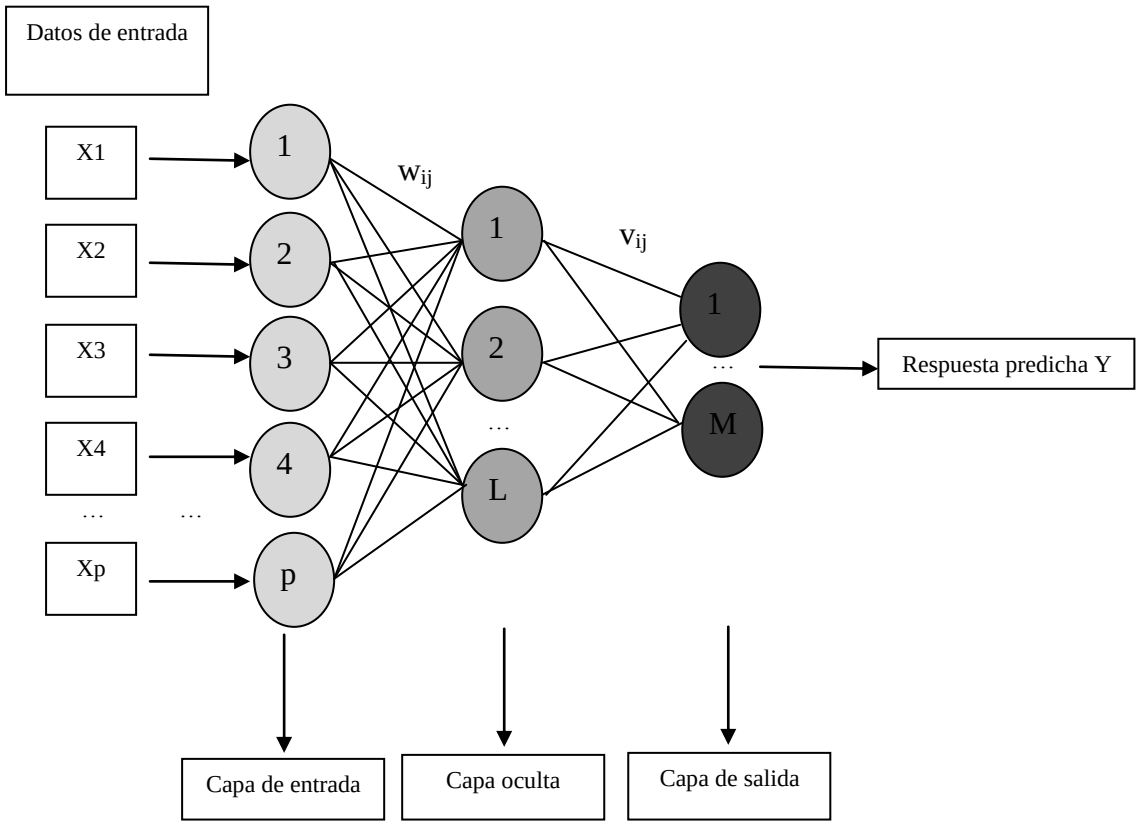

Figura 1: Perceptrón multicapa

2.4.3. Entrenamiento o aprendizaje de la red. Algoritmo backpropagation. Funcionamiento de la red.

Durante el aprendizaje o entrenamiento del sistema se evalúan las condiciones de pertenencia a cada una de las categorías. El aprendizaje supervisado se caracteriza por conocer la respuesta que debería tener la red frente a una determinada entrada. De esta manera, se compara la salida deseada con la salida de la red y si existen discrepancias se ajusta iterativamente los pesos considerando en cada paso la información sobre el error cometido.

El algoritmo *backpropagation* se basa en el ajuste de los pesos de las conexiones de la red en función de las diferencias entre los valores deseados (verdaderos) y los obtenidos por el sistema.

Así, la etapa de aprendizaje tiene por objeto hacer mínimo el error entre la salida brindada por la red y la salida deseada o verdadera. El aprendizaje se hace sobre un conjunto de datos, llamado conjunto de entrenamiento, que consta de un grupo de patrones asociados a sus correspondientes salidas.

Se pretende minimizar una función de error cuya expresión para el patrón j viene dada por

$$E_i = \frac{1}{2}\sum_{k=1}^{M}(d_{ik} - y_{ik})^2$$

donde la d_{ik} es la salida deseada para la neurona de salida k cuando se presenta el patrón i. La medida de error general se expresa como

$$E = \sum_{i=1}^{N} E_i$$

Este algoritmo realiza la modificación de los pesos basándose en la técnica del gradiente decreciente. Considerando al conjunto de pesos en un espacio de tantas dimensiones como pesos se tenga, el algoritmo busca obtener información sobre la pendiente de la superficie y modificar iterativamente los pesos de modo de hallar el mínimo global.

Una vez que se tiene la red estimada, al presentarse un patrón de entrada X_i, se transmite mediante los pesos w_{ik} desde la capa de entrada hacia la capa oculta de la red. Las neuronas de esta capa oculta aplican la función de activación a las señales recibidas obteniendo un valor de salida. Estos valores son transmitidos por los pesos v_{jk}, quienes, mediante la aplicación de la misma función anterior, obtienen los valores de salida de la red correspondientes a las neuronas de la última capa.

Esta función de activación que se aplica sobre la entrada de cada neurona para obtener el valor de salida debe ser una función continua y derivable. En este trabajo la función de activación utilizada es del tipo sigmoidal logística.

2.4.4. Evaluación del modelo y selección de variables

Para realizar la validación del modelo obtenido con los datos del conjunto de entrenamiento, es necesario considerar el error que se comete cuando la red es aplicada sobre un nuevo conjunto de datos, el conjunto de prueba. Esta nueva aplicación brindará como resultado de clasificación la matriz de confusión. La matriz de confusión que muestra el tipo de las predicciones correctas e incorrectas cuando se aplica el modelo sobre el conjunto de prueba. La misma permite comprender en qué sentido se equivoca la red al intentar clasificar los nuevos textos. En el gráfico de esta matriz, las predicciones correctas están representadas por las barras que aparecen sobre la diagonal, mientras que el resto de las barras indican el tipo de error cometido (qué valor ha predicho el modelo y cuales el valor verdadero). La altura de las barras es proporcional al porcentaje de los registros que representan.

En esta aplicación se evaluó la participación de cada variable considerando el porcentaje de clasificación correcta en los datos de prueba. Se retuvieron aquellas variables cuya ausencia en la red provocaba un incremento considerable en el porcentaje de error global.
Se calculan los porcentajes de clasificación correcta en cada categoría de texto y también en forma global.

2.5. Software utilizado: Herramienta Weka

Para realizar la clasificación de los textos según el área disciplinar se empleó la herramienta Weka. Esta herramienta es un conjunto de librerías implementadas en Java. Dado que la licencia de Weka es GPL[1], este programa es de libre distribución y difusión.

3. RESULTADOS

[1] ***G**NU **P**ublic **L**icense*. http://www.gnu.org/copyleft/gpl.html

3.1. Análisis preliminar.

La primera comparación que se realiza, como ya se mencionó al describir la muestra, es la del número de palabras por texto. La misma se lleva a cabo mediante el test no paramétrico de Kruskal Wallis, arrojando una probabilidad asociada p=0.16, evidenciando que no existen diferencias significativas entre los corpus respecto al tamaño de los textos.

Comparaciones similares entre los corpus se llevan a cabo para las restantes variables hallando diferencias significativas ($p<0.05$) para el número de clíticos y de adverbios en los documentos analizados (Tabla 3). El número de clíticos es mayor en los textos de biometría y el número de adverbios es superior en los textos de filosofía.

Tabla 3. Comparación mediante test de Kruskal Wallis

Número promedio de:	**BIOMETRIA**	**FILOSOFIA**	**LINGÜÍSTICA INFORMÁTICA**	**Valor de p**
adjetivos	17,9	21,3	11,1	**0.0031**
adverbios	2,9	5,9	2,33	**0.0007**
Clíticos	4,1	2,7	2,44	**0.0072**
copulativos	4,7	6,0	4,0	**0.0122**
determinantes	26,8	32,4	20,9	**0.0031**
Nombres	44,6	45,0	30,2	**0.0010**
preposición	30,0	29,7	21,5	**0.0077**
Verbos	16,1	18,4	24,0	**0.2592**
Otro	18,8	21,4	16,7	**0.6324**
TOTAL_PALABRAS	165,8	182,9	155,1	**0.1664**

3.2. Modelo perceptrón multicapa

Para construir el modelo Perceptrón se utilizó la herramienta Weka y se analizaron 3 aspectos importantes en el proceso de elaboración de la red: arquitectura, entrenamiento y estimación del error cometido durante la generalización.

Para decidir el número de neuronas ocultas de la red se estimaron los modelos considerando de 1 a 10 neuronas ocultas y en cada caso se estimó el error global de clasificación. Se seleccionó el número de neuronas cuyo error resultó significativamente menor.

El entrenamiento de la red se realizó con un conjunto de textos ($n_1=30$) y la evaluación del mismo como clasificador se llevó a cabo sobre otro conjunto de textos diferente al anterior ($n_2=30$).

El modelo final seleccionado corresponde a una red con 7 neuronas en la capa oculta cuya matriz de confusión resultante se encuentra presentada en la tabla 4. Esta tabla presenta el resultado de la aplicación de la red estimada sobre el conjunto de textos de prueba.

Tabla 4. Matriz de confusión

	Corpus predicho			
Corpus	BIOMETRIA	FILOSOFIA	LINGÜÍSTICA	Total general
BIOMETRIA	30	0	0	30
FILOSOFIA	0	30	0	30
LINGÜÍSTICA	1	1	28	30
Total general	31	31	28	90

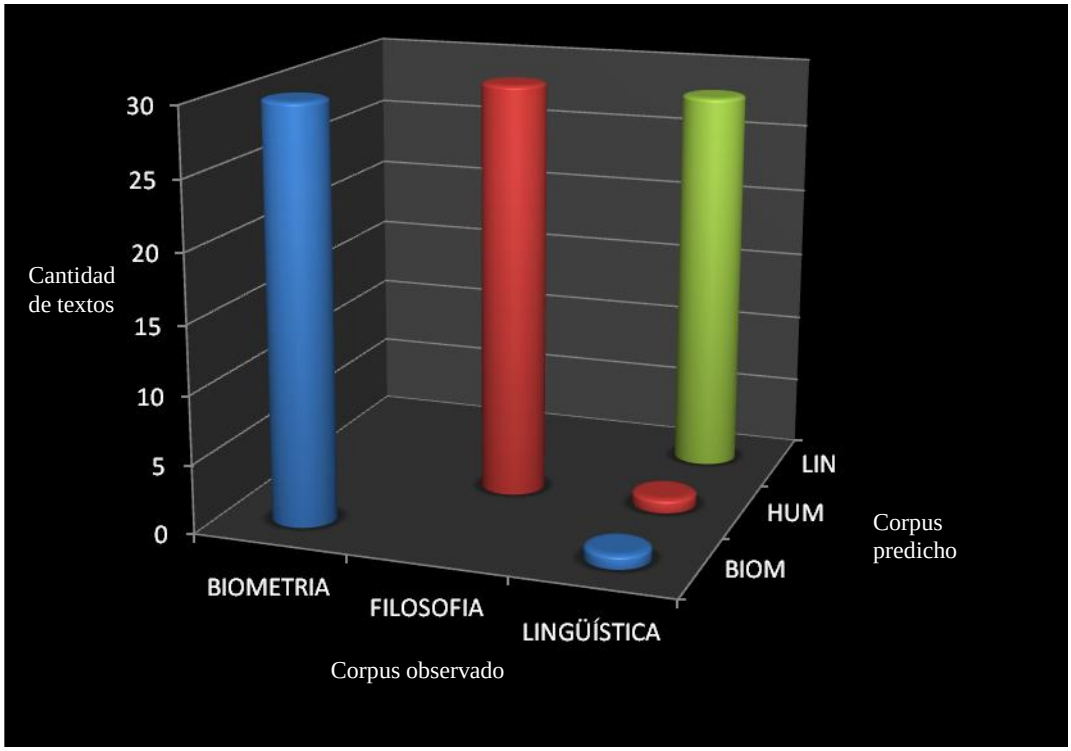

Gráfico 1: Matriz de confusión

Tabla 5: Tasa de error estimada por corpus en el MLP

Tasa de error por corpus				
	BIOMETRIA	**FILOSOFIA**	**LINGÜÍSTICA**	**Total**
Tasa	0%	0%	6.7%	2.2%

Se observa un alto porcentaje de clasificación correcta. Los errores de clasificación corresponden únicamente al corpus de Lingüística Computacional donde un texto se clasifica erróneamente en Biometría y otro en Filosofía.
Si se comparan estos porcentajes con los hallados en Beltrán (2011) mediante la aplicación de un modelo de regresión logística multinomial (Tabla 6) se evidencia un desempeño significativamente superior del MLP.

Tabla 6: Tasa de error estimada en el modelo de regresión logística

Tasa de error por corpus				
	BIOMETRIA	**FILOSOFIA**	**LINGÜÍSTICA**	**Total**
Tasa	16%	8%	17%	13.7%

6. CONCLUSIONES

Los resultados del análisis morfológico de los textos se analizaron teniendo en cuenta simultáneamente todas las mediciones realizadas sobre ellos sin aplicar ninguna transformación a las variables. Las variables predictoras fueron las proporciones de las categorías morfológicas halladas en el análisis automático del texto.

Este trabajo tuvo por objeto modelar el problema de la clasificación de textos según el área disciplinar a la que pertenecen: BIOMETRIA, FILOSOFIA, LINGÜÍSTICA INFORMÁTICA.

Mediante la utilización de la herramienta Weka se ha logrado comprobar la utilidad que tiene el uso de las Redes Neuronales Artificiales, en este caso específico el modelo Perceptrón Multicapa (MLP), para predecir el área de pertenencia de un texto. Las clasificaciones realizadas evidencian que la aplicación de este modelo es adecuada para predecir la disciplina.

La arquitectura y características de la red MLP, que brindan mejores resultados y hacen que la red tenga un comportamiento estable por lo que logra la habilidad de generalizar fueron los siguientes:

• Número de capas: 3

• Número de neuronas: 9 en la capa de entrada, 7 en la capa oculta y 3 en la capa de salida

• Los atributos corresponden a las proporciones de categorías morfológicas en el texto.

En este trabajo se observa que no se clasifican correctamente todos los registros, aunque el porcentaje de las clasificaciones incorrectas es muy bajo y corresponden a la disciplina Lingüística Informática. Esto indica que el error cometido es bajo evidenciando un buen desempeño de la red para discriminar los textos por su área disciplinar.

Referencias

Beltrán, C., Bender, C., Bonino, R., Deco, C., Koza, W., Méndez, B., Moro, Stella Maris. 2008 Recursos informáticos para el tratamiento lingüístico de textos. Ediciones Juglaría. Rosario.

Beltrán, C. 2009 Modelización lingüística y análisis estadístico en el análisis automático de textos. Ediciones Juglaría. Rosario.

Beltrán, C. 2010 Estudio y comparación de distintos tipos de textos académicos: Biometría y Filosofía. Revista de Epistemología y Ciencias Humanas. Grupo IANUS. Rosario.

Beltrán, C. 2010 Análisis discriminante aplicado a textos académicos: Biometría y Filosofía. Revista INFOSUR. Grupo INFOSUR. Rosario.

Beltrán, C. 2011. Aplicación del análisis de regresión logística multinomial en la clasificación de textos académicos: Biometría, Filosofía y Lingüística informática. Revista INFOSUR. Grupo INFOSUR. Rosario.

Bès,Gabriel, Solana, Z y Beltrán, C. 2005 Conocimiento de la lengua y técnicas estadísticas en el análisis lingüístico en Desarrollo, implementación y uso de modelos para el procesamiento automático de textos (ed. Víctor Castel) Facultad de Filosofía y Letras, UNCUYO

Catena, A.; Ramos, M.M; Trujillo, H.M. 2003. ANALISIS MULTIVARIADO. UN MANUAL PARA INVESTIGADORES. Bibiloteca Nueva S.L. España.

Cuadras, C.M. 2008 NUEVOS MÉTODOS DE ANÁLISIS MULTIVARIANTE. CMC Editions. Barcelona, España.

Flórez López, R.; Fernández Fernández, J.M. 2008. LAS REDES NEURONALES ARTIFICIALES. FUNDAMENTOS TEORICOS Y APLICACIONES PRACTICAS. Netbiblio S.L. España.

Johnson R.A. y Wichern D.W. 1992 Applied Multivariate Statistical Análisis. Prentice-Hall International Inc.

Khattre R. y Naik D. (2000) Multivariate Data Reduction and Discriminatio with SAS Software. SAS Institute Inc. Cary, NC. USA

Pogliano, A.M. (2010) "Análisis Estadístico de Datos Aplicados al Estudio de Calidad en Servicios de Traducción". Tesis Lic. en estadística. Facultad de Cs. Económicas y estadística. UNR.

Rodrigo Mateos, José Lázaro y Bès, Gabriel G. 2004 Análisis e implementación de clíticos en una herramienta declarativa de tratamiento automático de corpus. En VI Congreso de Lingüística General, Santiago de Compostela.

Solana, Z. Beltrán, C., Bender, C., Bonino, R., Deco, C., Koza, W., Méndez, B., Rodrigo, A., Tramallino, C. 2009 La interlengua de los aprendientes de español como L2. Aportes de la Lingüística Informática. GRUPO INFOSUR- Ediciones Juglaría.

Stokes, M. E., Davis, C.S., Koch, G.G. 1999 Categorical Data Analysis using SAS® System. WA (Wiley-SAS).