Redes neuronales artificiales. Una aplicación a la clasificación de textos según el género: Científicos – No científicos.

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

Abstract

This work proposes the Artificial Neuronal Network model using supervised learning: Multilayer Perceptron, employing as criterion of classification the gender of the text (Scientific/Non Scientific) and the characterization of the texts based upon the frequency distribution of the morphosyntactic categories.

This network proved its effectiveness in predicting the gender, determining the values of its own parameter set.

The correct percentage classification for each gender was 99% and 95%, respectively for Scientific and Non Scientific texts, while the global error was 2.7%.

Key words: neuronal networks, multivariate analysis, text classification

Resumen

En este trabajo se propone el modelo de Redes Neuronales Artificiales con aprendizaje supervisado: Perceptrón Multicapa, utilizando como criterio de clasificación el género al que pertenece el texto (Científico / No Científico) 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 género, determinando los valores del conjunto de parámetros correspondientes a la misma.

El porcentaje de clasificación correcta en cada género fue 99%, 95%, para Científicos y No Científicos respectivamente; mientras que el error global fue del 2.7%.

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

1. INTRODUCCION

La clasificación de documentos es una de las tareas imprescindibles para brindar utilidad a tanta información disponible actualmente en la web. El objetivo de la clasificación automática es categorizar documentos dentro de un número fijo de categorías definidas previamente 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. Para realizar el entrenamiento del sistema es necesario disponer de conocimiento previo de expertos en forma de decisiones de categorización asignadas a cada uno de los documentos. Durante el mismo, se evalúan las condiciones de pertenencia a cada una de las categorías. 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 género al que pertenece el texto (científico / no científico) 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 conjunto de textos que participan en la investigación, el corpus, corresponde a distintos tipos de acuerdo a los requerimientos de los objetivos planteados. Estos textos fueron agrupados de la siguiente manera:

- Noticias de tipo general, en español.
- Resúmenes, en español, de trabajos científicos presentados a congresos o revistas de Biometría/Estadística.
- Resúmenes, en español, de trabajos científicos presentados a congresos o revistas de Lingüística.
- Resúmenes, en español, de trabajos científicos presentados a congresos o revistas de Filosofía.

Los textos periodísticos fueron seleccionados de un corpus mayor utilizado por el equipo de investigación INFOSUR utilizado en mi tesis de doctorado. Este corpus se construyó con noticias extraídas de las páginas web de periódicos argentinos. Por otro lado, los textos científicos fueron seleccionados a partir de un marco muestral compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas, extraídos de internet pertenecientes a las disciplinas: Biometría, Lingüística y Filosofía. La unidad de muestreo fue el texto y la selección de la muestra se llevó a cabo empleando un muestreo aleatorio estratificado.

En esta etapa se han seleccionado, para los textos académicos, 60 textos de cada estrato de modo de poder utilizar 30 de ellos para estimar los modelos o entrenar los sistemas y los restantes para evaluar la tasa de error de clasificación en cada caso; mientras que para los textos periodísticos se seleccionaron 120, de modo de utilizar 60 de ellos durante el entrenamiento de cada sistema y los restantes para la etapa de evaluación.

Se dispusieron las muestras en un archivo de texto plano, archivo de texto sin formato, agregando las etiquetas “TEXTO 1”, “TEXTO 2”, etc., al inicio de cada uno de los textos para identificarlos luego del análisis morfológico automático. La base actual contiene 300 textos y 42.491 palabras.

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
…	…	…	…	…
2	1	Uno	uno	pron
2	1	de	de	prep
…	…	…	…	…

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

GÉNERO	TEXTO	adj	adv	cl	cop	det	nom	prep	v	OTRO	TOTAL_PAL
C	1	0,11	0,02	0,02	0,04	0,16	0,26	0,18	0,09	0,11	185
C	2	0,13	0,00	0,05	0,04	0,13	0,25	0,18	0,08	0,15	110
C	3	0,09	0,03	0,06	0,03	0,15	0,26	0,14	0,10	0,14	181
…	…	…	…	…	…	…	…	…	…	…	…
NC	1	0,08	0,01	0,02	0,03	0,16	0,32	0,21	0,09	0,09	186
NC	2	0,10	0,00	0,03	0,04	0,17	0,28	0,18	0,09	0,11	141
NC	3	0,10	0,03	0,01	0,03	0,19	0,27	0,16	0,10	0,11	183
…	…	…	…	…	…	…	…	…	…	…	…

2.4. Redes Neuronales Artificiales: El Perceptrón Multicapa

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 dos géneros: Científicos y no Científicos. 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 (género), con el propósito de lograr clasificar un nuevo texto para el cual se cuenta con el análisis morfológico pero se desconoce su género.

Respecto a la arquitectura y el entrenamiento se trabajó de la misma manera que en Beltrán (2012).

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 muestra las predicciones correctas e incorrectas cuando se aplica el modelo sobre el conjunto de prueba y permite comprender en qué sentido se equivoca la red al intentar clasificar los nuevos textos.

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 género de texto y también en forma global.

2.5. Software utilizado: Herramienta Weka

Para realizar la clasificación de los textos según el género 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

3.1. Análisis preliminar

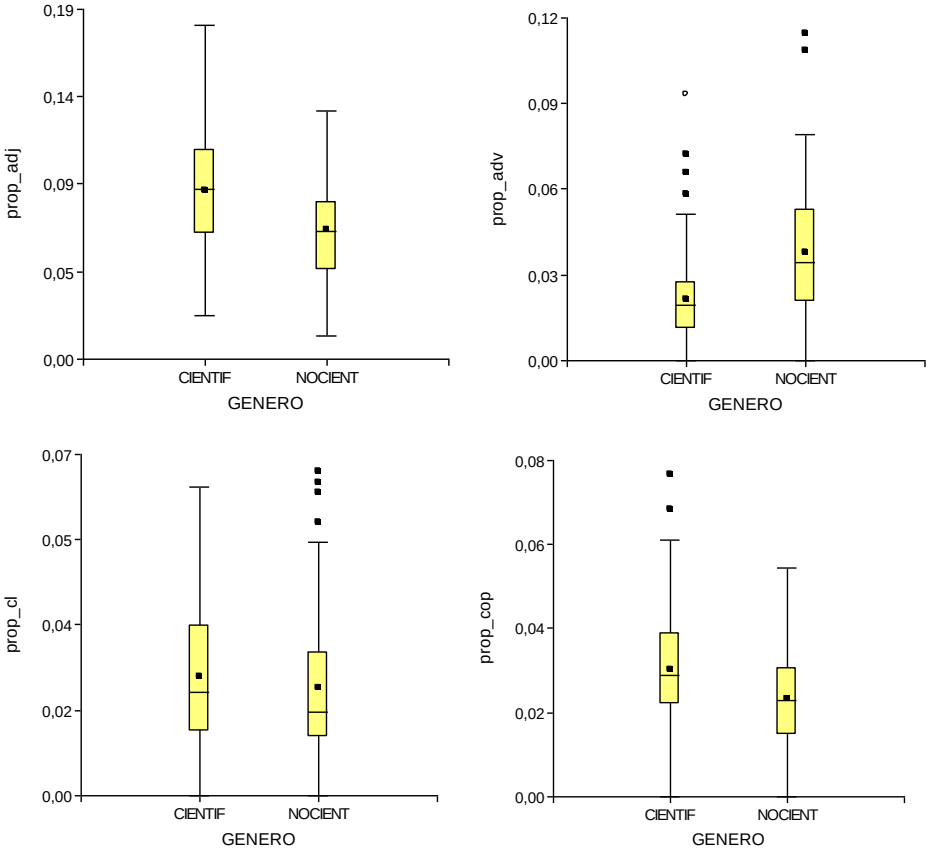
Esta comparación se llevó a cabo mediante la utilización del test no paramétrico de Wilcoxon ya que se trata de dos tipos de textos. En este caso se hallaron diferencias significativas en el número de palabras por texto (p=0.0062) por lo cual las comparaciones para las categorías morfológicas se realizaron sobre las proporciones de cada una de ellas, de modo que la diferencia en el tamaño de los textos no afecte los resultados (Tabla 3).

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

Tabla 3. Comparación mediante test de Wilcoxon

Proporción de:	CIENTIFICO	NO CIENTIFICO	Valor de p
adjetivos	0,09	0,07	0,0001
adverbios	0,02	0,04	0,0001
clíticos	0,02	0,02	0,2365
c. copulativas	0,03	0,02	0,0011
determinantes	0,16	0,15	0,0008
nombres	0,24	0,26	0,0109
preposición	0,17	0,15	0,0117
verbos	0,14	0,14	0,3195
otro	0,12	0,15	0,0001

La tabla 3 muestra que existen diferencias estadísticamente significativas ($p<0.05$) entre los textos científicos y no científicos para las proporciones de todas las categorías morfológicas excepto para verbos y clíticos cuyas probabilidades asociadas son $p=0.3195$ y $p=0.2365$, respectivamente.



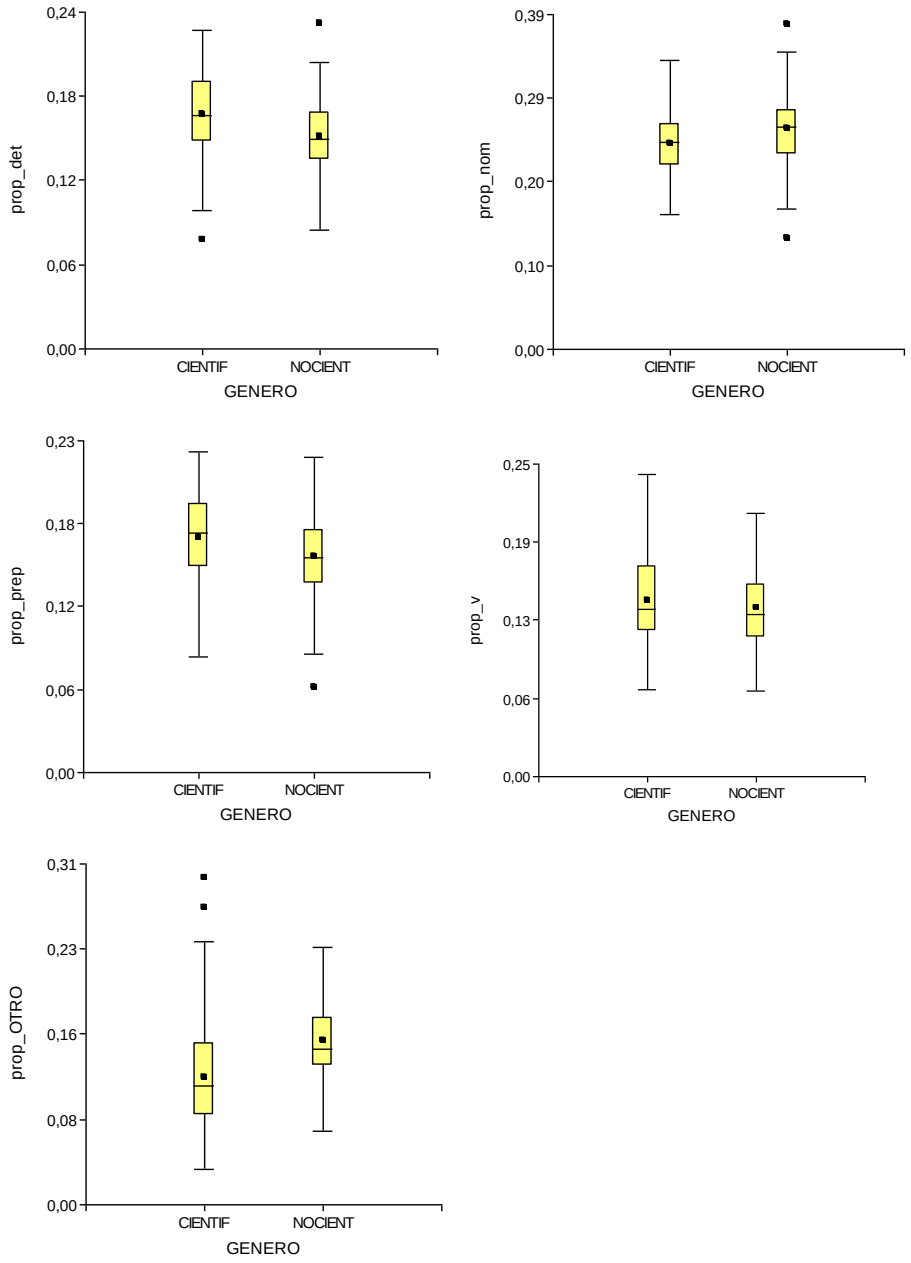


Gráfico 1: Diagrama de caja (Box-plot) para la proporción de palabras por categoría morfológica según género

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 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	Corpus predicho		
	CIENTIFICO	NO CIENTIFICO	Total
CIENTIFICO	89	1	90
NO CIENTIFICO	3	57	60
Total	92	58	150

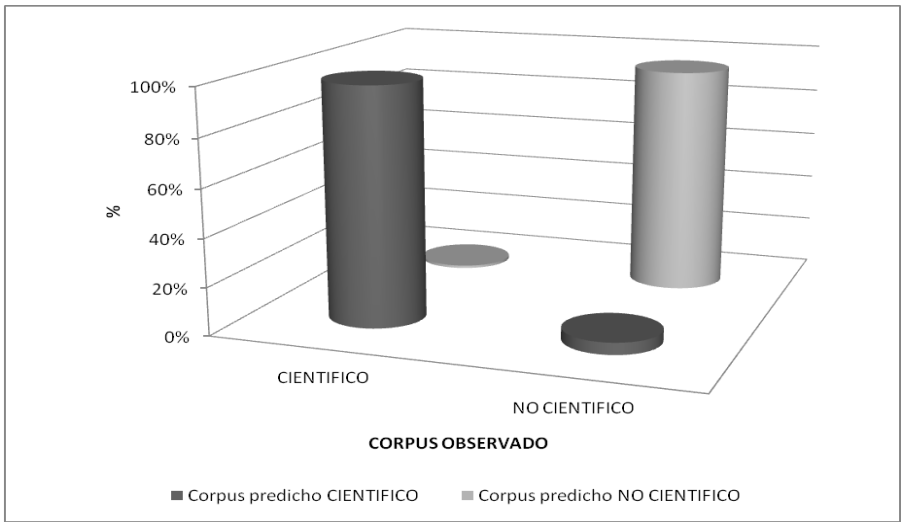


Gráfico 2: Clasificación según género mediante Redes Neuronales

Tabla 5: Tasa de error estimada, Precisión y Cobertura

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	1,1%	5,0%
Precisión	96,7%	98,3%
Cobertura	98,9%	95,0%

La red fue evaluada utilizando la muestra que no fue utilizada en el entrenamiento, hallando una tasa de mala clasificación del 2,7%, siendo 1% para los textos científicos y 5% para los no científicos. Respecto a la precisión y cobertura fueron de 97% y 99% para el género CIENTÍFICO y de 98% y 95% para los textos NO CIENTÍFICOS, respectivamente.

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 género: Científico / No científico.

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 género correspondiente a un texto. Las clasificaciones realizadas evidencian que la aplicación de este modelo es adecuada para predecir el género.

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 2 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. Esto evidencia un buen desempeño de la red para discriminar los textos por su género.

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.

Beltrán, C. 2012 Aplicación de redes neuronales artificiales en la clasificación de textos académicos según disciplina: Biometría, Filosofía y Lingüística informática. Revista de Epistemología y Ciencias Humanas. Grupo IANUS. 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).