Universidad Nacional de Rosario

Secretaría de Ciencia y Tecnología

INFORME FINAL
DE PROYECTOS DE INVESTIGACIÓN
FORMULARIO DE PRESENTACIÓN

1.IDENTIFICACIÓN DEL PROYECTO

1.1.DENOMINACIÓN DEL PROYECTO

1AGR228 MODELIZACIÓN ESTADÍSTICA EN LA CLASIFICACIÓN DE TEXTOS: CIENTÍFICOS Y NO CIENTÍFICOS

AÑO DE INICIO: 2014 **AÑO DE FINALIZACIÓN:** 2017

1.2.DIRECTOR DEL PROYECTO

DIRECTOR
Apellido y Nombre: Beltrán, Celina
CUIL : 27 - 21722712 - 2
Domicilio particular: San Martín 2274 - Casilda
Domicilio laboral: Campo Experimental Villarino cc14- Zavalla
Teléfono: 03464-15683414
FAX:
E-mail: beltranc@dat1.net.ar

1.3.RADICACIÓN DEL PROYECTO

DEPENDENCIA :FAC. DE CS. AGRARIAS
UNIDAD EJECUTORA:Facultad de Ciencias Agrarias

2.LOGROS DEL PROYECTO*

*No es necesario informar sobre todos los items sólo de aquellos en los que se hayan producido logros.

2.1. LOGRO DE LOS OBJETIVOS DEL PROYECTO (contribución al avance del conocimiento científico y tecnológico)

* Reproduzca aquí los objetivos originalmente planteados y luego describa el cumplimiento de los mismos.

OBJETIVOS PLANTEADOS EN EL PROYECTO
-Utilizar las técnicas estadísticas multivariadas para describir y comparar las estructuras de distintos tipos de textos.
- Obtener una regla de clasificación de textos según el género: Científico y No científico.
-Realizar el análisis automático de textos científicos y no científicos y conformar una base de datos. SU CUMPLIMIENTO
Respecto a los datos:

-Se realizó el análisis automático de 150 textos académicos.
-Se logró confeccionar una base de datos formada por 150 textos conteniendo la información procedente del análisis morfológico de éstos.
-La arquitectura de la base de datos fue "ajustada" o modificada cuando así fue requerido de modo de poder aplicar las técnicas estadísticas pertinentes de manera adecuada. En algunos casos fue necesaria la transformación de las variables originales ampliando así la cantidad de campos de la base.
-Durante el proceso de estimación de los modelos se confeccionaron los programas necesarios para ejecutar las técnicas de clasificación en distintos software estadísticos y se comparó la información brindada por cada uno de ellos en cada tipo de análisis.

Respecto a los resultados encontrados en la clasificación según el género de los textos se pueden enumerar los siguientes resultados:

Los resultados del análisis morfológico de los textos se analizaron teniendo en cuenta simultáneamente todas las mediciones realizadas sobre ellos.
El desempeño de las técnicas fue medido con la tasa de mala clasificación (TMC), la precisión (PR) y la cobertura (CO), calculadas sobre una muestra de textos no incluidos en la estimación de los modelos.
En la comparación entre Regresión logística y árboles de clasificación, el árbol presentó una TMC inferior a la del modelo logístico logrando clasificar con mayor precisión los textos científicos. Para el AC la TMC, PR y CO resultaron 4%, 84% y 96% para los textos científicos y 28%, 92% y 72% para los textos no científicos, respectivamente. Para el modelo de RL la TMC, PR y CO resultaron 14%, 83% y 86% para los textos científicos y 26%, 77% y 74% para los textos no científicos, respectivamente. La diferencia en la tasa de mala clasificación sólo se diferenció en el corpus de textos científicos para el cual con el árbol se obtuvo un 4% de mala clasificación versus un 14% para el modelo de regresión logística. En ambos tipos de análisis, las diferencias entre los dos tipos de textos están centradas principalmente en el porcentaje de adverbios, adjetivos, nombres y preposiciones presentes. Sin embargo, en el modelo de regresión logística han intervenido otras variables en la discriminación como los determinantes y conjunciones copulativas; mientras que el árbol de clasificación utiliza el porcentaje de verbos, categoría morfológica no utilizada en la regresión.
Una ventaja observada en el árbol de clasificación es la adaptación para recoger el comportamiento no aditivo de las variables predictoras, de manera que las interacciones se incluyen de manera automática. Sin embargo, en esta técnica se pierde información al tratar a las variables predictoras continuas como variables dicotómicas.
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.
Se compararon los métodos Vector Machine (SVM), Sequential Minimal Optimization (SMO), Regresión Logística, Análisis Discriminante Lineal (ADL) y Cuadrático (ADC). De todos los métodos de clasificación considerados, el que presentó el menor porcentaje de mala clasificación fue el ADC (16.67%). Tanto el ADL como el ADC dieron buenos resultados al clasificar los textos en Científicos y No Científicos, presentando un 18% y 16.67% de mala clasificación respectivamente. En cuanto a los métodos de aprendizaje de máquina, el que presenta mejores resultados es el SVM con kernel lineal y constante de penalización C= 0.1 o 0.2 (19.33%). Del resto de los métodos aplicados, el que presenta peores resultados es SVM con kernel RBF, arrojando valores de porcentaje de error de mala clasificación que van del 34% al 40%. Si bien el método SMO presentó porcentajes bajos de mala clasificación para valores altos de C (18%), no

es considerado uno de los mejores debido a la variabilidad que presenta en sus resultados al considerar distintos valores de la constante C, dando indicios de cierta inestabilidad del método para clasificar bien.
De los métodos aplicados el del Vecino más Cercano presenta el mejor desempeño (13% de mala clasificación) teniendo como principales ventajas la simpleza de su aplicación y la estabilidad de su comportamiento. También presentaron desempeños aceptables los métodos Árboles de Clasificación (14% de mala clasificación) y Análisis Discriminante Cuadrático (17 % de mala clasificación). Cabe destacar, que debido que los grupos presentan estructuras de covariancias distintas, es de esperar que el Análisis Discriminante Cuadrático clasifique mejor que el Análisis Discriminante Lineal (18% de mala clasificación). Por otro lado, no es posible conocer en de qué manera afecta la presencia de estructuras de covariancias distintas entre los grupos para los métodos restantes.
Se utilizó el algoritmo AdaBoost, y se evaluó su desempeño en los métodos de clasificación Regresión Logística y SMO (Sequential minimal optimization. Se observó que al aplicar AdaBoost teniendo en cuenta como algoritmo de base el método de Regresión Logística no se presentó una reducción en el porcentaje de mala clasificación. En cambio, para el caso del método SMO como algoritmo de base, el porcentaje de mala clasificación bajó un 8,67%.

2.2. LOGROS METODOLÓGICOS

La utilización que se ha hecho de estas técnicas multivariadas en la clasificación de textos académicos considerando la distribución de las categorías gramaticales de los mismos, es una aplicación novedosa en la lingüística computacional. Esto se debe principalmente que se le ha dado la importancia adecuada a la interpretación de los coeficientes de los modelos estadísticos logrando una descripción muy clara de los textos y de las categorías consideradas en función de la información proveniente del análisis morfológico automático.

2.3.CONTRIBUCIÓN A LA FORMACIÓN DE RECURSOS HUMANOS

Un miembro del equipo ha cursado durante la ejecución del proyecto la Especialización en Bioinformática en la Facultad de Ciencias Agrarias de la Universidad Nacional de Rosario y se encuentra en la última etapa de su trabajo final de especialización.

2.4.CONTRIBUCIÓN AL DESARROLLO ECONÓMICO Y SOCIAL

La clasificación automática de textos ha recibido más atención en los últimos tiempos debido a la gran cantidad de información disponible en formato electrónico. Esta nueva perspectiva en el estudio de las técnicas estadísticas multivariadas en la clasificación de textos académicos significa una contribución de importancia y una contribución novedosa.

2.5.TRANSFERENCIA DE RESULTADOS REALIZADAS

Se han realizado publicaciones, seminarios y presentaciones a congresos.
Los resultados que se refieren a la aplicación de la metodología estadística y a la confección de la base de datos han sido transferidos a los cursos de posgrado que he dictado en el período correspondiente al informe:
-Metodología de la Investigación, Doctorado de Humanidades y Artes, Mención Lingüística. Universidad Nacional de Rosario. Docentes: Dra. Celina Beltrán- Dr. Diego Beltrán. Año 2017. Duración: 30 hs.
-Estadística I. Doctorado en Ciencias Agrarias. Facultad de Ciencias Agrarias. Universidad Nacional de Rosario. Docentes: Dra. Celina Beltrán. Años 2014, 2015, 2016 y 2017. Duración: 50 hs.
-Estadística II. Doctorado en Ciencias Agrarias. Facultad de Ciencias Agrarias. Universidad Nacional de Rosario. Años 2014, 2015, 2016 y 2017. Docentes: Dra. Celina Beltrán. Duración: 50 hs.
-Estadística Multivariada. Especialización en Bioinformática. Facultad de Ciencias Agrarias. Universidad nacional de Rosario. Años 2015 y 2017. Duración 30 hs.
-Estadística Univariada. Especialización en Bioinformática. Facultad de Ciencias Agrarias. Universidad nacional de Rosario. Años 2015 y 2017. Duración 30 hs.
-Introducción al análisis estadístico en la investigación. Doctorado en Educación. Facultad de Humanidades y Artes. Universidad Nacional de Rosario. Años 2014, 2015,2 016 y 2017. Duración 30 hs.
- Taller de análisis de datos. Facultad de Ciencias Agrarias. Universidad nacional de Rosario. Docentes: Dra. Celina Beltrán- Lic. Ivana Barbona. Año 2016. Duración 30 hs.

2.6.PERSPECTIVAS DE FUTURA TRANSFERENCIA

Se espera profundizar la transferencia iniciada.

2.7. DIVULGACIÓN REALIZADA (Publicaciones, comunicaciones, etc)

Las siguientes publicaciones han surgido de esta investigación:

Libros:

"La estadística en la investigación. Elementos básicos y aplicaciones." Autores: Celina Beltrán; Ivana Barbona. Ediciones Del Revés. Rosario. Año 2016. I.S.B.N. 978-987-3852-13-8.

Artículos en revistas:

-Redes neuronales artificiales. Una aplicación a la clasificación de textos según el género: Científicos – No científicos. Autor: Celina Beltrán. Revista de Epistemología y Ciencias Humanas. Nro. 6 Año 2014. ISSN 1852-625X. http://www.revistaepistemologi.com.ar/

- La lectura de los textos electrónicos empleados en evaluaciones internacionales para estudiantes de nivel medio. Autor: Carolina P. Tramallino. TECNOLOGÍA EDUCATIVA REVISTA CONAIC – ISSN: 2395-9061

-Técnicas estadísticas de clasificación. Una aplicación en la clasificación de textos según el género: Textos Científicos y Textos No Científicos. Autor: Celina Beltrán. http://www.infosurrevista.com.ar/ Revista INFOSUR. ISSN: 1851-1996. Número 7. Año 2015

-Comparación de métodos de clasificación aplicados a textos Científicos y No Científicos. Autor: Ivana Barbona. http://www.infosurrevista.com.ar/ Revista INFOSUR. ISSN: 1851-1996. Número 7. Año 2015

-Comparación de dos técnicas multivariadas en la categorización de textos: Sistema de clasificación Bagging y Método del vecino más cercano. Autor: Celina Beltrán. Revista de Epistemología y Ciencias Humanas. Nro. 7 Año 2015. ISSN 1852-625X. http://www.revistaepistemologi.com.ar/

-Método de clasificación supervisada Support Vector Machine: Una aplicación a la clasificación automática de textos. Autor: Ivana Barbona, Celina Beltrán. Revista de Epistemología y Ciencias Humanas. Nro. 8 Año 2016. ISSN 1852-625X. http://www.revistaepistemologi.com.ar/

- La lectura de hipertextos para medir la comprensión lectora en evaluaciones destinadas a estudiantes de nivel medio. Autor: Bárbara Méndez, Carolina Tramallino. Reflexión Académica en Diseño y Comunicación (ISSN 1668-1673). Año XVII. Vol 29. Noviembre 2016. Buenos Aires. Argentina (Pag 202-206)

-Una revisión de las técnicas de clasificación supervisada en la clasificación automática de textos. Autor: Celina Beltrán, Ivana Barbona. Revista de Epistemología y Ciencias Humanas. Nro. 9 Año 2017. ISSN 1852-625X. http://www.revistaepistemologi.com.ar/

Trabajos presentados a congresos:

-Clasificación supervizada Support Vector Machine: Una aplicación a la clasificación automática de textos. Barbona, Ivana; Beltrán, Celina. 1º Congreso Argentino de Estadística (CAE I). Ciudad Autónoma de Buenos Aires. 6 al 9 de octubre 2015.

-Comparación del desempeño de técnicas de clasificación mediante una aplicación a textos científicos y no científicos. Beltrán, Celina; Barbona, Ivana. 1º Congreso Argentino de Estadística (CAE I). Ciudad Autónoma de Buenos Aires. 6 al 9 de octubre 2015.

-Aplicación del algoritmo Boosting adaptativo (adaboost) a un problema de clasificación automática de textos. Beltrán, Celina; Barbona, Ivana. Congreso Interamericano de Estadística. Rosario, octubre 2017

3. PRESUPUESTO EJECUTADO

(**)	ORIGEN DEL FINANCIAMIENTO(*)	PRESUPUESTO EJECUTADO EN EL AÑO 2014 (en $)	PRESUPUESTO EJECUTADO EN EL AÑO 2015 (en $)	PRESUPUESTO EJECUTADO EN EL AÑO 2016 (en $)	PRESUPUESTO EJECUTADO EN EL AÑO 2017 (en $)
5.580	U.N.R	2.790	2.790	0	0

(*): Excluidos los salarios y/o becas de los docentes-investigadores
(**): Puede consignarse monto global o discriminado en rubros
Nota: Si los años de duración del proyecto son mas de 3, repetir el cuadro precedente.

4. RECURSOS HUMANOS

4.1. INTEGRACIÓN DEL EQUIPO DE TRABAJO

APELLIDO Y NOMBRE	PERIODO EN EL QUE PARTICIPO EN LA EJECUCIÓN DEL PROYECTO	FUNCIÓN DENTRO DEL PROYECTO	MÁXIMO TÍTULO ACADÉMICO ALCANZADO	CARGO DOCENTE	DEDICACION	CARGO CIUNR	CARGO CONICET	HS. SEM. DED. AL PROY.	CATEGORÍA FIRME EN EL PROG. DE INCENTIVOS	CATEGORÍA EN TRÁMITE EN EL PROG. DE INCENTIVOS	UNIVERSIDAD	FIRMA
BELTRAN, CELINA	2014-2017	DIRECTOR	pos doctorado	J.T.P.	Exclusiva	Ninguno	Ninguno	20	II	Ninguna	UNR	
Méndez, Bárbara	2014-2017	INTEGRANTE	doctorado	Aux. 1era.	Simple	Ninguno	Ninguno	5	V	Ninguna	UNR	
Carolina Trammallino	2014-2017	INTEGRANTE	doctorado	Aux. 1era.	Simple	Ninguno	Ninguno	5	V	Ninguna	UNR	
Ivana Barbona	2014-2017	INTEGRANTE	Licenciada	Aux. 1era.	Semi	Ninguno	Ninguno	10	V	Ninguna	UNR	
		DIRECTOR		Ninguno	Ninguna	Ninguno	Ninguno		Ninguna	Ninguna	UNR	
		DIRECTOR		Ninguno	Ninguna	Ninguno	Ninguno		Ninguna	Ninguna	UNR	
		DIRECTOR		Ninguno	Ninguna	Ninguno	Ninguno		Ninguna	Ninguna	UNR	
		DIRECTOR		Ninguno	Ninguna	Ninguno	Ninguno		Ninguna	Ninguna	UNR	
		DIRECTOR		Ninguno	Ninguna	Ninguno	Ninguno		Ninguna	Ninguna	UNR	
		DIRECTOR		Ninguno	Ninguna	Ninguno	Ninguno		Ninguna	Ninguna	UNR	
		DIRECTOR		Ninguno	Ninguna	Ninguno	Ninguno		Ninguna	Ninguna	UNR	
		DIRECTOR		Ninguno	Ninguna	Ninguno	Ninguno		Ninguna	Ninguna	UNR	

4.2. OTROS RECURSOS HUMANOS QUE PARTICIPARON EN EL EJECUCIÓN DEL PROYECTO

APELLIDO Y NOMBRE	DNI	FUNCIÓN DENTRO DEL PROYECTO	PERÍODO EN EL QUE PARTICIPO	MÁXIMO TÍTULO ACADÉMICO ALCANZADO	HS. SEM. DED. AL PROY.	FIRMA

Nota: Completar el cuadro precedente con los datos de alumnos, becarios, pasantes y graduados que hayan integrado el equipo, de acuerdo al marco regulatorio de cada facultad.

5. REGISTRO INSTITUCIONAL

La Unidad Académica donde se radicó el Proyecto hace constar el registro del presente informe.

Firma : Aclaración : Cargo :