Clasificador de Noticias usando Autoencoders

Gonzalo Farias, Sebastián Vergara, Ernesto Fabregas, Gabriel Hermosilla, Sebastián Dormido-Canto, and Sebastián Dormido

***Abstract*—This article presents a classification system for news with Deep Learning. With this tool the news are classified in the following categories: Sports, Politics, Economics, Show and Police. Also they receives an scope: Local (Valparaíso), National (Chile) and International (Rest of the World). The classifiers were built using a database with 542 news labeled with the previous criteria. The features were extracted with Autoencoders (AE) to train an Artificial Neural Network (ANN) of multiple classes *Softmax* (*Softmax* ANNs). Both stages were stacked following the concept of Deep Learning. The results with the data test (156 news) reach a success rate of 92.3% for the category classifier and 87.2% for the scope classifier. The general success rate for both, category and scope was 83.75%.**

***Keywords*—News Classifier, Autoencoder**

I. Introduction

La clasificacin automtica se puede resumir como el ejercicio de separar algn conjunto de elementos especficos a travs de un sistema artificial (computador), con el fin de otorgarle una categora o clase. Los elementos a clasificar pueden ser una gran variedad de tipos, tales como imgenes, vdeos, sonidos, seales fsicas (temperatura, voltaje, corriente, presin, etc.) y texto.

Respecto a este ltimo tipo de dato se han desarrollado una gran cantidad de algoritmos que hacen la extraccin de caractersticas y clasificacin bajo diferentes criterios [1]. Lo anterior se debe a que los textos generan diariamente grandes volmenes de informacin, los cuales se almacenan en bases de datos de pginas web, redes sociales y de empresas, por lo que discriminarla puede ser muy provechoso.

Existen algunos ejemplos de estudios realizados para la discriminacin de textos, como por ejemplo [2], donde se muestra un sistema clasificador de sentimientos de una base de datos de *tweets* mediante distintos algoritmos de aprendizaje automtico: Mquina de Vectores Soporte (SVM) [3], Clasificador Bayesiano [4] y Clasificador de Mxima Entropa [5]. Por otro lado, en [6] se implementa un sistema capaz de discriminar cuando estamos en presencia de un *tweet* de carcter ofensivo mediante Redes Neuronales Recurrentes (RNNs). Con respecto a la extraccin de caractersticas, en [7] se realiz un estudio comparativo emprico sobre diferentes tcnicas que abordan esta problemtica.

En el rea de los textos noticiosos, se han realizado una serie de investigaciones con la finalidad de discriminarlos bajo diferentes categoras, en [13] se construye un sistema basado en Word2Vec; el cual es un modelo de redes neuronales que busca otorgar a cada palabra un nico vector en un espacio generalmente de cientos de dimensiones y teniendo en cuenta adems, que las palabras que compartan contextos comunes se encuentran a distancia menores. El mtodo de clasificacin en este artculo se basa en algoritmos de Deep Learning. En [14] se presenta un sistema para el anlisis de temas de artculos periodsticos, utilizando un clasificador binario basado en Mquina de Vectores Soporte y Clasificador Bayesiano, el algoritmo es entrenado a partir de una vector generado a travs de la cuenta de palabras claves extradas de manera automtica. Tambin [15] se realiza un trabajo similar utilizando el mtodo de clasificacin *Naive Bayes*.

Particularmente en el presente artculo se busca implementar un sistema de clasificacin de noticias segn las categoras Deporte, Poltica, Economa, Espectculo y Policial, y de acuerdo al mbito que pertenecen, Local (Regin Valparaso), Nacional (Chile) e Internacional. El clasificador diseado se basa en Autoencoders (AE) y una Red Neuronal Artificial con capa de salida *Softmax* (ANNs *Softmax*).

Un sistema clasificador de noticias puede ser un herramienta atrayente para usuarios que busquen informarse solamente de tpicos especficos, ya que de esta manera el sistema se encarga de filtrar toda la informacin que no es inters del usuario y as tener un acceso a las noticias de forma ms amigable.

La estructura de este trabajo inicia con un marco terico, que busca introducir a los algoritmos utilizados, a continuacin, se detalla el funcionamiento del clasificador propuesto y finalmente se exponen los resultados obtenidos.

II. Marco Terico

El *deep learning* (aprendizaje profundo) ha surgido como uno de los enfoques ms utilizados en los ltimos años ya que es capaz de resolver problemas en muchas reas, tanto cientficas como cotidianas donde otras tcnicas de *machine learning* tienen limitaciones. En este apartado se explican de manera general los algoritmos de *deep learning* utilizados en el sistema clasificador propuesto [8].

A. Red Neuronal Artificial Como multiclasificador

El aprendizaje se define como el proceso de adquisicin de conocimiento a travs de la experiencia, estudio o prctica. Una vez terminado el proceso, el sujeto es capaz de distinguir de acuerdo a lo aprendido. Esta capacidad est asociada directamente con los humanos, pero gracias a los avances tecnolgicos, se han logrado encontrar algoritmos capaces de realizar un proceso anlogo, pero mediante computadores (mquinas). En otras palabras, el aprendizaje automtico busca la capacidad de clasificar de la forma ms general y robusta posible diferentes clases y a travs de variados tipos de algoritmos como SVM,

G. Farias, S. Vergara and G. Hermosilla are with the Escuela de Ingeniería Eléctrica. Pontificia Universidad Católica de Valparaíso. Valaraíso, Chile e-mail: gfarias,gabriel.hermosilla@pucv.cl, sebantoniovi@gmail.com

E. Fabregas, S. Dormido-Canto and S. Dormido are with Departamento de Informática y Automática. Universidad Nacional de Educación a Distancia. Madrid, Spain email: efabregas@bec.uned.es, sebas,sdormido@dia.uned.es

978-1-5386-5586-3/18/$31.00

rboles de decisin, reglas de asociacin, redes bayesianas, algoritmos de agrupamiento, ANNs, entre otros.

Las redes neuronales artificiales en particular, son modelos matemticos que intentan reproducir el comportamiento del sistema nervioso humano. Se conforman a partir de la interconexin de unidades ms bsicas denominadas perceptrn (o neurona artificial), las cuales son una simplificacin matemtica de una neurona biolgica. Cuando un perceptrn recibe un dato de entrada, ste lo pondera (atena, amplifica o inhibe) a travs de un parmetro llamado peso, asociado a la sinopsis biolgica y a su salida se evala una funcin de activacin que determina si la neurona est excitada o no [9]. La figura 1 muestra la arquitectura de una red neuronal de 4 capas: la capa de entrada formada por 4 neuronas, las dos capas ocultas formadas por 3 neuronas cada una y la capa de salida formada por una nica neurona.

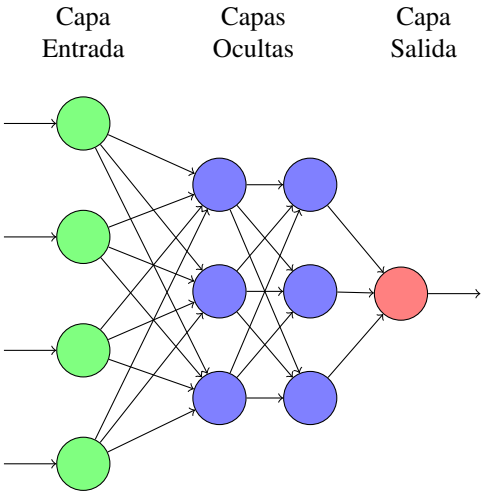

Fig. 1. Red Neuronal Artificial (ANN).

El apilamiento de capas de perceptrones interconectados se entrenan a partir de datos conocidos, es decir, se ingresa un conjunto de datos y se obtiene un resultado a la salida de la red, dicho resultado se compara con el deseado y segn el error obtenido se modifican los pesos de cada perceptrn. El algoritmo utilizado para el entrenamiento de la red se denomina *Backpropagation*, el cual funciona en dos etapas. Primeramente, al excitar las neuronas de entrada con un conjunto de datos, estos se propagan a travs de la red hasta generar una salida, lo obtenido se compara con la salida deseada y se calcula el error de forma individual. En una segunda etapa el error obtenido se propaga haca atrs, partiendo por las capas superiores hasta las iniciales, modificando los pesos de la manera ms equilibrada posible, con la finalidad de que cada neurona aprenda una caracterstica especfica de los datos de entrada. Al realizar el proceso anterior de manera repetitiva y con diferentes muestras, la red es capaz de aprender a clasificar los datos. Cabe destacar el aumento de las capas ocultas en una red neuronal le dan el carcter de profundo (deep).

B. Clasificador

Existen dos enfoques respecto a la cantidad de clases que se desean clasificar, por un lado, un clasificador busca separar en dos categoras y un multiclasificador en tres o ms. Este ltimo puede ser ms complicado de entrenar ya que necesita diferenciar ms tipos de clases (problema complejo) en comparacin al primero. Es importante destacar que a partir de varios clasificadores se puede obtener los resultados de uno mltiple. En particular para el sistema propuesto se utiliza una capa de salida tipo *Softmax* multiclase.

Softmax, es una funcin exponencial normalizada utilizada para otorgar una probabilidad sobre un nmero determinado de salidas que se emplean como capa final en clasificadores basados en redes neuronales. A cada salida de las neuronas de la ltima capa se le asignan valores dentro del intervalo [0,1] y segn la neurona que entrega el valor mximo se les otorga una clase a los datos de entrada.

C. Autoencoder

Las caractersticas de un elemento aluden a las cualidades (informacin) que lo definen. Mientras que el agrupamiento de stas conforman el vector de caractersticas y segn el tamao de este ltimo se define la dimensionalidad del problema. Una excesiva dimensionalidad puede generar problemas de sobreajuste y un perjuicio en la velocidad de procesamiento, por lo que reducirla es una etapa fundamental en el proceso de clasificacin [10].

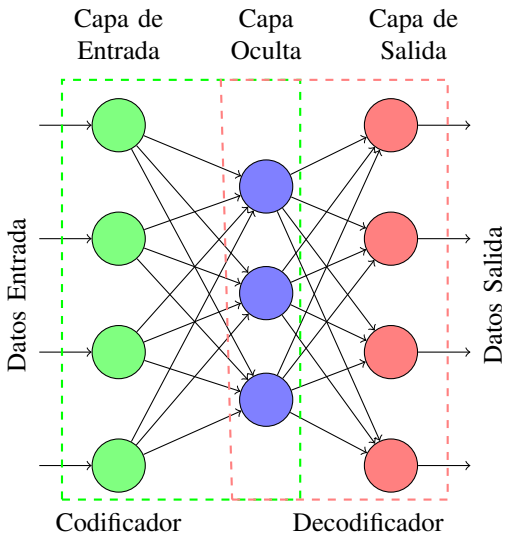

Fig. 2. Esquema de un Autoencoder.

Un *Autoencoder* (AE) tiene como objetivo extraer caractersticas de manera automtica y sin supervisin mediante una red neuronal artificial capaz de aprender una representacin codificada para un conjunto de datos. Se conforma tpicamente de tres capas de neuronas artificiales, una de entrada, una de salida y una capa oculta que las conecta. Se buscan los pesos de la red con la finalidad de minimizar la diferencia entre la salida y entrada del AE [11]. La figura 2 muestra grficamente la arquitectura de un AE, con sus capas y las etapas de codificacin y decodificacin.

III. Clasificador propuesto

En esta seccin se detalla el funcionamiento del sistema clasificador de noticias de acuerdo a sus diferentes etapas.

A. Matriz de Frecuencia

Para el entrenamiento del sistema propuesto como primera etapa se construye la matriz de frecuencia. Esta matriz se genera a partir de un listado que contiene palabras clave respecto a cada **Categora** y **mbito** de una noticia, las longitudes de las listas son de 1351 y 894 palabras respectivamente y se componen como se indica en la tabla I.

TABLE I
CANTIDAD DE PALABRAS POR CATEGORA Y MBITO*

Categora o mbito	Cantidad de Palabras
Deporte	290
Poltica	314
Economa	247
Policial	199
Espectculo	301
Local*	135
Nacional*	414
Internacional*	345

Cabe destacar que la eleccin de las palabras claves se realiza de manera manual, es decir, que se inspecciona un conjunto de noticias y se seleccionan aquellas palabras que tengan ms relacin con el tema, por ejemplo para la categora de Poltica, palabras como presidente, ministro y diputado forman parte de la lista. Conjuntamente se agregan palabras directamente sin la necesidad de ser extradas de noticias.

En este punto finalmente se cuenta la cantidad de veces que aparece una palabra especfica en una noticia. El resultado debe ser que para cada tipo de noticia, la frecuencia de palabras de igual clase, debe ser mayor. En la tabla II se muestra de manera ilustrativa la matriz de frecuencia para el listado de **Categora** para noticias de diferentes clases. En nuestro caso, la matriz tiene una dimensionalidad de 542x1351. La matriz de frecuencia para el listado de **mbito** es anloga.

TABLE II
EJEMPLO DE MATRIZ DE FRECUENCIA

Noticia	D	P	E	Po	Es
Deporte(D)	25	2	1	3	0
Economa(E)	2	0	32	12	5
Policial (Po)	0	1	1	15	3
Espectculo (Es)	5	3	2	7	24
Poltica (P)	1	33	15	2	0

Cabe destacar que el proceso de eleccin de palabras es sensible al sistema y debe funcionar de la manera ms robusta posible. Adems su determinacin debe ser actualizado empricamente segn los resultados obtenidos en la clasificacin. Adems es fundamental contar con una base de datos de noticias lo ms amplia posible. Siguiendo estas premisas, se construyeron dos conjuntos de noticias conformados por 542 para el clasificador de categoras y 395 para el de mbito, subdivididas de acuerdo a lo que se indica en la tabla III.

Una vez construida la matriz de frecuencia con los datos de entrenamiento, se continua con el proceso de extraccin de caractersticas.

B. Extraccin de Caractersticas

Como se coment anteriormente, el proceso de extraccin de caractersticas busca disminuir la dimensionalidad del problema para generar una clasificacin ms robusta y con menor exigencia en el procesamiento. La dimensionalidad del vector de caractersticas de cada noticia respecto a la categora es de 1351. Mediante la implementacin del autoencoder se redujo a 100 caractersticas, disminuyendo la dimensionalidad en un factor de 13.51. Con respecto al mbito de la noticia, el nmero de caractersticas es de 894 y a travs de otro AE se redujo la dimensionalidad a 45, por lo que el factor de reduccin es de 19.86.

TABLE III
CANTIDAD DE NOTICIAS POR CATEGORA Y MBITO*

Categora o mbito	Cantidad de Noticias
Deporte	142
Poltica	109
Economa	93
Policial	99
Espectculo	99
Local*	107
Nacional*	139
Internacional*	149

C. Clasificadores Implementados

Los clasificadores implementados se conforman a partir de una nica capa tipo *Softmax* de 5 neuronas para el clasificador de categoras (una neurona por categora) y 3 para el de mbito (una neurona para cada mbito), sus entrenamientos se realizaron a partir de las caractersticas extradas en el proceso anterior.

Como ltima etapa se apilan los autoencoders de categora y mbito con la capa *Softmax* respectiva y se realiza un ltimo entrenamiento de refinamiento. En la siguiente figura se ilustra el sistema de clasificacin construido segn la categora (anlogo para el clasificador de mbito). La figura 3 muestra la implementacin del sistema clasificador con el *Neural Network Toolbox* de MATLAB.

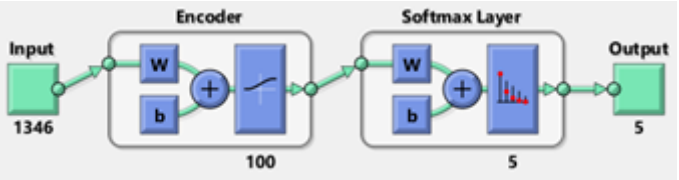

Fig. 3. ANNs diseado con el *Neural Network Toolbox* de MATLAB.

Es fundamental sealar que la entrada al sistema clasificador es el cdigo HTML que posee la noticia en su totalidad, en la figura 4 se muestra el diagrama de bloques.

IV. RESULTADOS

Para evaluar el sistema construido se realizaron tres pruebas, en primer lugar, se evalu el clasificador de Categora, seguido del de mbito para finalmente realizar un test general del clasificador de noticias. Para representar los resultados obtenidos de cada clasificador se utiliza la matriz de confusin [12]. En la tabla IV se muestra la cantidad de noticias utilizadas en el test. Las filas representan las categoras y las columnas el mbito.

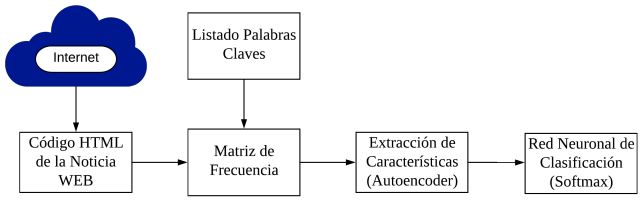

Fig. 4. Diagrama de bloque Sistema Clasificador de Noticias.

TABLE IV
CONJUNTO DE NOTICIAS DE TEST POR CATEGORA Y MBITO

Categora	mbito			
	Local	Nacional	Internacional	Total
Deporte	11	11	11	33
Poltica	10	10	12	32
Economa	7	10	11	28
Policial	12	10	10	32
Espectculo	11	10	10	31
Total	51	51	54	156

A. *Matriz de confusin*

La matriz de confusin es muy utilizada en sistemas de clasificacin, en ella se muestra ms detalladamente cmo se predijeron las muestras. La figura 5 muestra un ejemplo de una matriz de confusin.

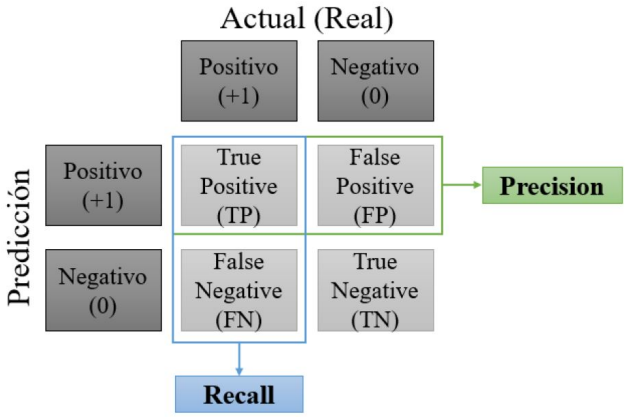

Fig. 5. Matriz de confusin, en la seccin superior o inferior los valores reales de las noticias. En la seccin izquierda, las predicciones realizadas por el algoritmo de clasificacin.

Dicha matriz ilustra cmo se proporciona informacin relevante bajo los siguientes cuatro conceptos:

- **TP**: *True Positive* (Verdadero Positivo), cuando la clase predicha y la real son iguales.
- **FP**: *False Positive* (Falso Positivo), cuando se predice como como +1 y corresponde a 0.
- **FN**: *False Negative* (Falso Negativo), se predice como clase 0 pero pertenece a +1.
- **TN**: *True Negative* (Verdadero Negativo), cuando el clasificador predice como clase 0 y efectivamente pertenece a ella.

De esta matriz, se puede extraer informacin realizando los siguientes dos anlisis:

1) Horizontal: Cuntas muestras predichas en la clase i (fila) pertenece a la clase j (columna).
2) Vertical: Cuntas muestras que efectivamente pertenecen a la clase j (columna) fueron catalogadas como i (fila).

En las figuras 6 y 7 se muestran las matrices de confusin obtenidas a partir del clasificador de categora y mbito respectivamente.

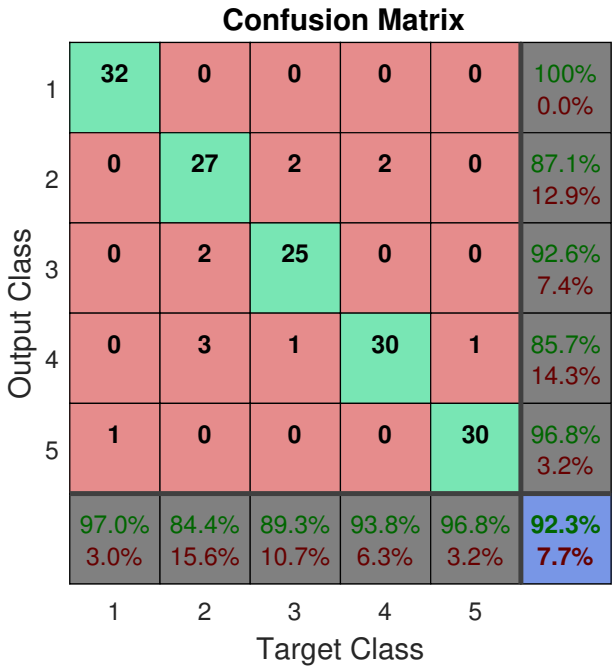

Fig. 6. Matriz de Confusin clasificador de categora (1 Deporte, 2 Poltica, 3 Economa, 4 Policial, 5 Espectculo).

De la figura anterior se puede extraer que la tasa de acierto para este clasificador de Categora alcanz un porcentaje del 92.3%. Los errores generados en la clasificacin se deben a diferentes fenmenos, como por ejemplo:

- Ambigedad en la clasificacin: esto se ve reflejado en los errores cometidos en noticias polticas y econmicas, ya que stas suelen estar muy relacionadas entre s.
- Ruido en la informacin: las noticias en sitios web muestran generalmente la informacin principal y en su alrededor otros tipos de noticias, esto genera problemas al ahora de clasificar.
- Ruido en el cdigo fuente de la pgina: hay palabras, generalmente con tilde, que al buscarlas en la noticias no se detectan debido que en el cdigo de la pgina estn escritas de diferente manera (poltica y *política*).

Para combatir los tres puntos anteriores, realizar un preprocesamiento con la finalidad de obtener el cuerpo de la noticia sin otra informacin sera una buena manera de perfeccionar el sistema. En trabajos posteriores se evaluar la inclusin de una etapa de preprocesamiento.

De acuerdo a la matriz de confusin tambin se pueden extraer otros resultados. Por ejemplo, de las 5 categoras clasificadas, con las que se obtiene un mejor desempeo son: Deporte y Espectculo (clases 1 y 5), con un 100% y 96.8% respectivamente. Lo anterior concuerda con lo esperado ya

que este tipo de noticias suelen ser muy diferentes a las otras clases, en cambio para Poltica, Economa y Policial existe un grado de cercana, por ejemplo: *son muy comunes noticias donde un Presidente comenta sobre proyectos de ley de carcter policial o econmico*.

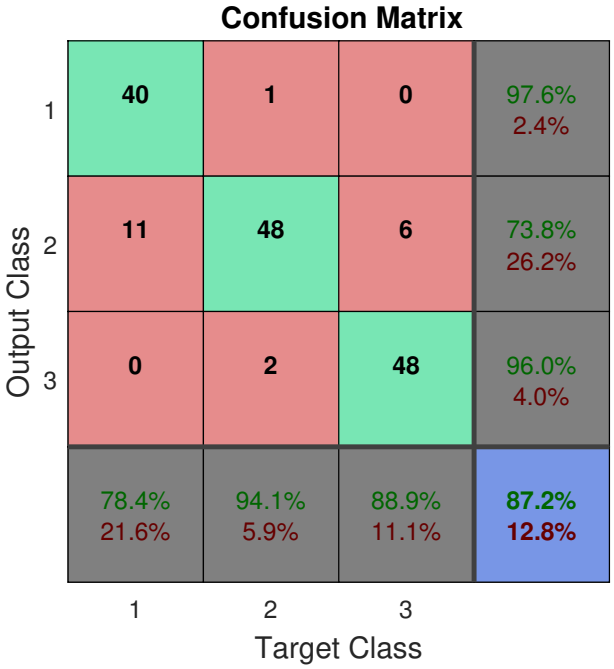

Fig. 7. Matriz de Confusin clasificador de mbito (1 Local, 2 Nacional, 3 Internacional).

Para el clasificador de mbito se obtuvo una tasa de acierto de 87.2%, logrando resultados inferiores, aunque cercanos al clasificador anterior y los fenmenos que generan el error son bsicamente los mismos.

En este caso el mayor error en la clasificacin se gener en las noticias de mbito nacional, obteniendo una tasa de Falsos Positivos de $26,2\%$, esto se puede deber a que la gran mayora de las noticias con que se construy el clasificador son de sitios web nacionales y por ende aparecen constantemente palabras claves relacionadas con este mbito. Por otro lado, para los mbitos Local e Internacional se alcanzan aciertos en torno al 96%, ambas tienen palabras muy caractersticas, como son el nombre de ciudades locales y de pases o ciudades extranjeras.

El comportamiento del sistema general alcanz un porcentaje de acierto de 83.75%, es decir, de las 156 noticias de test, 130 fueron bien predichas tanto en su categora como mbito. Cabe destacar que la tasa de acierto total del sistema no corresponde a la multiplicacin entre las tasas de acierto de cada clasificador, ya que hay casos donde el error se produce en una misma noticia, de manera contraria se estara contando dos veces un mismo error.

V. Conclusin

Diariamente se generan una gran cantidad de datos de diferente ndole, por lo que categorizarlos y extraer informacin de ellos puede ser muy provechoso. En particular, la clasificacin de noticias puede funcionar para generar aplicaciones que las recopile de diferentes sitios, las ordene en categoras. De esta forma el usuario puede tener un mejor acceso a la actualidad local, nacional o internacional.

En este trabajo se presenta la implementacin de dos clasificadores basados en *deep learning*, para determinar la categora y el mbito de una noticia extrada de Internet. El proceso de clasificacin se genera a travs de la construccin de matrices de frecuencias mediante de una lista de palabras clave. Luego se extraen caractersticas a travs de *Autoencoders* para finalmente realizar la clasificacin con una capa final de neuronas artificiales tipo *Softmax*.

Las noticias se clasifican segn su Categora: Deporte, Poltica, Economa, Espectculo y Policial; y segn el mbito: Local (Regin Valparaso), Nacional (Chile) o Internacional (resto del mundo). Los resultados obtenidos alcanzaron tasas de acierto del 92.3% para el clasificador de Categora y 87.2% para el de mbito. Las noticias bien predichas para ambos casos lograron un porcentaje de acierto de 83.75%.

Para mejorar las tasas del sistema es necesario refinar el listado de palabras claves y ampliar la base de datos de noticias. Adems para ste ltimo se requiere abarcar la mayor cantidad de sitios web y as generar un sistema ms robusto. Tambin se puede filtrar el ruido obtenido al extraer el texto de la noticia, lo cual puede generar una considerable en el comportamiento del sistema.

Acknowledgment

This work has been funded by the Chilean Ministry of Education under the Project FONDECYT 1161584, and the Spanish Ministry of Economy and Competitiveness under the Project No. ENE2015-64914-C3-2-R.

References

[1] G.E. Hinton, R.R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504507, 2006.
[2] A. Go, R. Bhayani, L. Huang, “Twitter sentiment classification using distant supervision,” CS224N Project Report, Stanford, vol. 1, no. 12, 2009.
[3] C. Schuldt, I. Laptev, B. Caputo. Recognizing human actions: a local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 3, pp. 32-36.
[4] P. Cheeseman, M. J. Kelly, M. Self, J. Stutz, W. Taylor, D. Freeman. “Autoclass: A Bayesian classification system.” In Machine Learning Proceedings, pp. 54-64. 1988.
[5] A.L. Berger, J. Vincent, P. Della, A. Stephen. “A maximum entropy approach to natural language processing.” Computational linguistics, vol. 22, no. 1, pp. 39-71, 1996.
[6] G. K. Pitsilis, H. Ramampiaro, H. Langseth, “Detecting of- fensive language in tweets using deep learning,” arXiv preprint arXiv:1801.04433, 2018.
[7] G. Forman, “An extensive empirical study of feature selection metrics for text classification,” Journal of machine learning research, vol. 3, no. Mar, pp. 12891305, 2003.
[8] F. Chollet. Deep learning with python. Manning Publications, 2017.
[9] S. Haykin. “Neural Networks: A comprehensive foundation, ” Prentice Hall, 2004.
[10] I. Guyon, A. Elisseeff. “An introduction to feature extraction.” In Feature extraction, pp. 1-25. Springer, Berlin, Heidelberg, 2006.
[11] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, pp.3371-3408, 2010.
[12] T. Fawcett. “An Introduction to ROC Analysis”, Pattern Recognition Letters. no. 27, vol 8, pp. 861874, 2006.

[13] Kato, Ryoma, and Hiroyuki Goto. "Categorization of web news documents using word2vec and deep learning." Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia. 2016.'

[14] Bracewell, David B., et al. "Category classification and topic discovery of japanese and english news articles." Electronic Notes in Theoretical Computer Science 225 (2009): 51-65.

[15] Asy'arie, Arni Darliani, and Adi Wahyu Pribadi. "Automatic news articles classification in indonesian language by using naive bayes classifier method." Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services. ACM, 2009.