Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 15-22 recibido 28-03-2019 revisado 07-05-2019 aceptado 17-05-2019

Análisis de Sentimiento en el dominio salud: analizando comentarios sobre fármacos

Sentiment analysis on health domain: analyzing patient comments on drugs

Cristóbal Colón-Ruiz, Isabel Segura-Bedmar, Paloma Martínez
Computer Science Department, University Carlos III of Madrid
{ccolon, isegura, pmf}@inf.uc3m.es

Resumen: En el análisis de sentimiento, la mayoría de las investigaciones han sido llevadas a cabo en dominios generales tales como en el análisis de opiniones de películas, restaurantes y otros productos o servicios, con escasa representación en el ámbito médico. Cada vez más, los pacientes buscan información en internet sobre los posibles beneficios, efectos adversos y opiniones que otros pacientes tiene sobre diferentes fármacos. El objetivo de este trabajo es predecir el grado de satisfacción de los pacientes respecto a un determinado fármaco en base a sus comentarios. Para llevar a cabo la tarea, hemos utilizado una colección de comentarios sobre diferentes tipos de fármacos y aplicado una red convolucional para la clasificación de los mismos. Los resultados muestran que este tipo de redes proporciona mejores resultados en términos de precisión, recall y f1-score que empleando algoritmos clásicos de clasificación como las maquinas de soporte vectorial.
Palabras clave: Análisis de Sentimiento, Clasificación de textos multi-clase, Deep Learning, Redes convolucionales

Abstract: Most of the research in sentiment analysis has been conducted in general domains such as the analysis of film reviews, restaurants and other products or services, but without much representation in the medical field. Increasingly, patients are searching the internet for information about the potential benefits, adverse effects and opinions that other patients have about different drugs. The aim of this work is to predict the degree of patient satisfaction with a given drug based on their reviews. To carry out the task, we have used a collection of reviews on different types of drugs and applied a convolutional network for their classification. The results show that this type of network provides better results in terms of precision, recall and f1-score than using classical classification algorithms such as vector support machines.
Keywords: Sentiment Analysis, Multi-Class Text Classifiation, Deep Learning, Convolutional Neural Network

1 Introducción

Actualmente, debido al aumento de redes sociales, blogs, foros de discusión y webs especializadas, existe un creciente volumen de datos de opinión registrados digitalmente y disponibles para su estudio (Liu, 2012). El análisis de sentimientos conforma el campo dedicado al análisis de las emociones, y opiniones de personas a través del Procesamiento del Lenguaje Natural (PLN).

Los sistemas dedicados a la minería de opiniones y al análisis de sentimientos, son ampliamente utilizados en ámbitos sociales y empresariales debido al fuerte papel que representa la opinión de las personas sobre un determinado producto y servicio. En el ámbito de la medicina, se puede extraer información en lo que refiere al estado de salud y condición de los pacientes, así como información sobre los tratamientos que reciben (Denecke y Deng, 2015).

Dentro de este dominio, el análisis y vigilancia de los diferentes fármacos, una vez aprobados en el mercado, es clave en lo que respecta a la mejora de la seguridad. Durante los ensayos clínicos, la probabilidad de detectar todos los efectos adversos a fármacos es muy baja, ya que los ensayos clínicos tie-

ISSN 1135-5948. DOI 10.26342/2019-63-1

nen una duración limitada y se realizan sobre pequeños grupos de la población. El efecto de un fármaco puede variar en función de un elevado número de factores, tales como la edad, condición previa del paciente, enfermedad, tratamiento con otros fármacos, etc. El análisis automático de la opinión que los pacientes tienen sobre determinados fármacos no sólo puede proporcionar una visión sobre el grado de satisfacción que los pacientes tienen sobre un determinado fármaco, sino que también puede dar la pista sobre fármacos con un elevado número de efectos adversos.

El objetivo de este trabajo es predecir el grado de satisfacción que los pacientes tienen sobre determinados fármacos, en base a sus comentarios. Concretamente, realizamos una serie de experimentos sobre una colección de comentarios extraídos de la web Drugs.com (Gräßer et al., 2018). El corpus contiene una serie de comentarios y opiniones escritas en inglés, por usuarios no expertos, sobre diferentes fármacos. Además de los comentarios, los usuarios dan una valoración relativa a la efectividad y los posibles efectos adversos del fármaco. La valoración puede tomar valores entre 1 (valoración más negativa) y 10 (valoración más positiva). La representación de las diferentes clases suponen un problema de multiclase desbalanceado donde aquellas más representadas son las más polarizadas, siendo las clases 10, 9 y 1 las que presentan un mayor número de instancias.

Para esta tarea, proponemos un enfoque de aprendizaje profundo supervisado, en concreto, una red neuronal convencional (CNN) (Collobert y Weston, 2008), cuya entrada es un modelo pre-entrenado de word embeddings (Pyysalo et al., 2013), que nos permite representar los comentarios. Además, realizamos una comparativa entre dicho enfoque con un modelo Máquinas de Soporte Vectorial (SVM) (Joachims, 1998). En este caso, los textos son representados utilizando un modelo de bolsa de palabras con la frecuencia inversa de las palabras.

El documento está estructurado de la siguiente manera: la sección 2 presenta brevemente una serie de estudios relacionados con en análisis sentimental en comentarios sobre fármacos y medicamentos. En la sección 3, se describen la colección de textos utilizada para llevar a cabo el estudio y la metodología empleada. En las secciones 4 y 5, se muestran los resultado obtenidos y una discusión sobre los mismos. Finalmente, en la sección 6 se presentan las conclusiones y trabajos futuros.

2 Estado de la cuestión

Los principales trabajos en análisis de sentimiento aplicado al dominio de salud en fármaco-vigilancia resultan útiles a la hora de seleccionar que fármacos deben ser vigilados para identificar posibles efectos adversos. Dichos trabajos se han centrado en dos grandes tareas: (I) clasificación de comentarios y (II) extracción de aspectos de opiniones. En la clasificación de los comentarios se extrae el sentimiento u opinión global sobre el texto, en este caso, sobre los comentarios de fármacos (ejemplo: positivo, negativo, neutro). En la extracción de aspectos, se identifican las opiniones o sentimientos referidas a determinadas características o aspectos de la entidad (fármaco) sobre el que se realiza el comentario.

Los diferentes enfoques empleados para la clasificación de los comentarios pueden agruparse en enfoques de aprendizaje automático (supervisado o no supervisado), enfoques basados en el léxico y reglas o en enfoques híbridos (Pang y Lee, 2008).

En (Na et al., 2012), realizan una clasificación binaria de sentimientos (positivos y negativos) en comentarios de medicamentos extraídos de la pagina web DrugLib.com. En este trabajo, se aplicó un lexicón construido a partir del Subjetivity Lexicon (SL) (Wilson, Wiebe, y Hoffmann, 2005) y de SentiWordNet (SWN) (Baccianella, Esuli, y Sebastiani, 2010). Además, los autores también exploraron un enfoque basado en SVM, donde los textos habían sido representados utilizando el modelo de bolsa de palabras. Los experimentos muestran que el enfoque lingüístico obtiene mejores resultados (accuracy de 78 %), frente a un 73 % de accuracy obtenido por el modelo SVM.

En (Na y Kyaing, 2015), se emplea otro enfoque puramente lingüístico en clasificación multiclase de sentimientos (positivo, negativo y neutral) en comentarios extraídos de webmd.com. Los autores también utilizaron un lexicón extraído de SL y SWN, además de un árbol de dependencias, obtenido con la herramienta Stanford NLP library (De Marneffe, MacCartney, y Manning, 2006). En este artículo también se realiza una comparativa con modelos SVM y bolsas de palabras, obte-

niendo unos resultados de 69 % de accuracy en el enfoque lingüístico y 66 % de accuracy usando SVM.

En (Gopalakrishnan y Ramaswamy, 2017), se realiza una clasificación polarizada (positivo, negativo y neutral) en comentarios extraídos de la web askapatient.com. Los comentarios son preprocesados y representados con vectores de frecuencia inversa (Tf-idf), donde las diferentes características son unigramas, bigramas y trigramas. Los autores exploraron diferentes algoritmos de aprendizaje supervisado tales como SVM, Redes Neuronales Probabilísticas (PNN) y Redes Neuronales de Base Radial (RBFN), obteniendo los mejores resultados con los modelos RBFN, con una F1 en torno al 84-94 %, para cada una de las clases.

En (Yadav et al., 2018), abordan el problema de multiclasificación de comentarios sobre medicamentos. Las posibles clases son: medicamentos efectivos, ineficaces o con efectos adversos graves. El corpus es extraído de la web patient.info y los textos son preprocesados obteniendo unigramas, bigramas y trigramas. Además, cada comentario es asignado con una posible puntuación (positiva, negativa y neutral), obtenida empleando el lexicón SWN. En este artículo, se evalúan diferentes modelos tales como SVM, Random Forest, Perceptrón Multicapa (MLP) y CNN. El mejor resultado es obtenido por el modelo de CNN obteniendo un 82 % de macro-F1. En este modelo, los textos fueron representados utilizando el modelo pre-entrenado de word embeddings de Google news.

En el trabajo (Gräßer et al., 2018), realizan una predicción de la satisfacción de los pacientes (positivo, negativo y neutral) sobre una colección de comentarios extraídos de Drugs.com y Druglib.com. Los textos son representados utilizando un modelo de bolsa de bigramas y trigramas. Los autores aplican regresión logística y obtienen resultados del 92.24 % de accuracy sobre los comentarios de Drugs.com y del 69.88 % sobre Druglib.com.

3 Metodología

3.1 Dataset

El corpus consiste en una colección de comentarios escritos en inglés, extraídos de forma automática por (Gräßer et al., 2018) de la web Drugs.com, que proporciona información sobre medicamentos tanto a pacientes como a profesionales de la salud.

El contenido del corpus consiste en comentarios de pacientes sobre un fármaco específico, así como una puntación del 1 al 10, que refleja el grado de satisfacción del paciente con el fármaco. A continuación, se describe con más detalle la información que proporciona el corpus para cada uno de los comentarios:

- Nombre del fármaco.
- Condición o estado en el que se encuentra el paciente; por ejemplo, *insomnio, depresión, trastorno bipolar, etc.*
- Texto del comentario del paciente sobre el fármaco.
- Puntuación del 1 al 10 del grado de satisfacción del paciente sobre el fármaco.
- Fecha de en la que se registró el comentario.
- Número de usuarios, consumidores y profesionales, que marcaron el comentario como útil.

Por ejemplo, un paciente con fibrilación auricular publicó el siguiente comentario: "*Only on it for 8 days. After 5 days started having shortness of breath, muscle spasms in upper back, pounding heart rate, fatigue, stiffness in neck and face*". Dicho comentario hace referencia al fármaco Flecainide y describe una serie de efectos adversos que éste le ha producido a los pocos días. La puntuación del fármaco proporcionada por el paciente es negativa con un valor de 1.

El corpus empleado está formado por una total de 215.063 comentarios. El corpus ha sido dividido en datasets de entrenamiento y test, manteniendo en ambos la proporción de las clases mediante un muestreo aleatorio estratificado. Los datasets de entrenamiento y test se encuentran en una proporción de 75 % y 25 %, respectivamente. Adicionalmente, un 15 % del total del dataset de entrenamiento ha sido reservado como dataset de validación, que nos permite aprender los mejores hiperparámetros de los distintos modelos.

La distribución de los diferentes comentarios en función de su clase (grado de satisfacción) puede ser observada en la figura 1. Exista una fuerte desproporción del número de instancias entre las distintas clases, predominando aquellas con valoraciones más polarizadas.

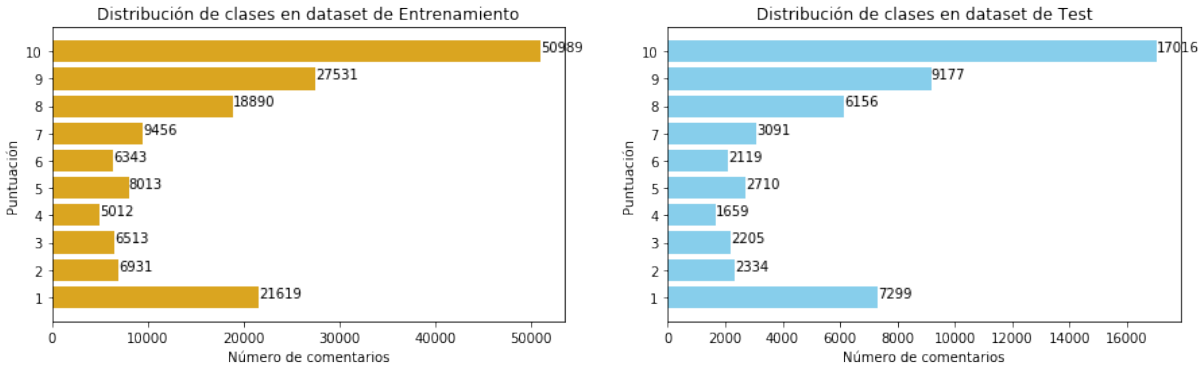

Figura 1: Distribución de las clases en los datasets de Entrenamiento y Test

3.2 Sistema

La finalidad de esta tarea consiste en resolver un problema de multiclasificación de textos. En particular, el objetivo es inferir un modelo capaz de predecir la puntuación (clase) o grado de satisfacción para un comentario concreto.

Los enfoques de aprendizaje profundo han demostrado su capacidad de aprender las características más relevantes para la tarea de PLN a resolver. En concreto, las redes convolucionales (CNN) han demostrado ser muy efectivas a la hora de realizar tareas de clasificación de textos. Dicha eficacia reside en su capacidad para extraer patrones representativos que describen el texto en forma de n-gramas (Kim, 2014), (Wang et al., 2017). Por este motivo, en este trabajo proponemos una arquitectura basada en CNN.

Para el preprocesado de los comentarios, el texto es tokenizado y las expresiones numéricas son cegadas, obteniendo un vocabulario de 49.339 tokens. Las CNN requieren que su entrada siempre tenga la misma longitud. Esto último es debido a la necesidad de definir unos filtros convolucionales donde los pesos se modificarán teniendo en cuenta un tamaño de ventana concreto para todos y cada uno de los textos (comentarios). Sin embargo, debido a la longitud variable de los comentarios, es necesario realizar un padding o rellenado de tokens, que no aporten información (zero-padding) al final de cada uno de los textos para igualar su longitud.

Basándonos en la función de distribución acumulada sobre la longitud de los comentarios, el 90 % de los textos preprocesados tienen una longitud menor o igual a 250 tokens. Sin embargo, como el tamaño máximo es relativamente pequeño (2.006 tokens), hemos decidido definir el tamaño de la entrada como este tamaño máximo, y evitando así realizar truncado.

La arquitectura propuesta consta de una capa de entrada donde cada uno de los textos, ya procesados, son representados como una matriz de embeddings gracias a un modelo pre-entrenado de word embeddings (Pyysalo et al., 2013). Este modelo de word embeddings fue entrenado sobre una colección de textos de PubMed, PMC y de la Wikipedia en inglés (versión de 2013), utilizando word2vec (Mikolov et al., 2013). El modelo contiene un vocabulario de 5.443.656 de palabras. La dimensión de los word embeddings es 200.

A continuación, la arquitectura cuenta con una capa de convolución, donde diferentes filtros o sub-matrices operan deslizándose a lo largo de la matriz de entrada, produciendo como salida un mapeo de características de los textos. Para obtener los mejores hiperparámetros de la red, aplicamos grid-search, obteniendo 64 filtros con un tamaño de ventana de 2, 3 y 5 vectores de palabras. Como función de activación empleamos una unidad de rectificación linear (ReLU) (Nair y Hinton, 2010) para evitar problemas de desvanecimiento de gradiente. (Ide y Kurita, 2017).

Una vez obtenido el mapeo de características, se emplea una capa de agrupación (pooling) para obtener aquellas características más relevantes a lo largo de las salidas de los diferentes filtros. Existen diferentes métodos de agrupamiento, tales como la media, el máximo de las entradas de la capa o modelos de atención (Boureau et al., 2010), (Er et al., 2016). Para nuestra experimentación, empleamos un agrupamiento basado en el máximo de los valores de cada convolución. Dicho método es empleado con el fin evitar que

la operación de padding pueda afectar negativamente a la representación de los textos (Suárez-Paniagua y Segura-Bedmar, 2018).

Finalmente, las salidas de las capas de agrupamiento son concatenadas y enviadas a una capa de perceptrones totalmente conectados. En esta capa, se utiliza la función de activación ReLU con el objetivo de mejorar la predicción de las clases. En este punto, los diferentes vectores que representan cada uno de los textos, son conectados con una capa Softmax, que finalmente predice la puntuación del 1 al 10 para cada comentario. En la figura 2 podemos observar la arquitectura descrita junto con las diferentes capas de regulación (dropout) empleadas para evitar sobreajuste durante el entrenamiento.

4 Resultados

En la evaluación de nuestros sistemas, hemos utilizado las métricas estándar para las tareas de clasificación de textos: precisión, recall y F1 (Frakes y Baeza-Yates, 1992). Dado que nos encontramos ante un problema de multiclasificación, donde las diferentes clases están desbalanceadas, empleamos las métricas de micro y macro-averages. Con macro-average, mostramos la media de los valores obtenidos para cada clase de forma independiente, mientras que en micro-average se agrega la contribución de cada una de las clases.

Como sistema baseline, consideramos un modelo de Support Vector Machines (SVM) con kernel lineal. En este sistema, los comentarios son representado el modelo extendido de bolsa de palabras con tf-idf. Los clasificadores SVM han demostrado buenos resultados en problemas de clasificación de textos, ya que suelen ser linealmente separables (Joachims, 1998).

Los experimentos con arquitecturas de CNN han sido realizados empleando la librería Keras con tensorflow. Los experimentos realizados SVM han sido realizados empleando la librería scikit-learn de Python.

En el caso de la CNN, además se utilizó un optimizador ADAM (Kingma y Ba, 2014), con un learning rate de 0.001, un batch size de 500 comentarios durante 200 epoch y categorical-crossentropy como función de pérdida.

En el caso del clasificador SVM se empleó un parametro de penalización C=10 con squared hinge como función de pérdida.

Para la configuración de los hiper-parámetros de los diferentes clasificadores empleamos grid search. Los diferentes modelos se han evaluado empleando el dataset de validación.

Label	**Precision**	**Recall**	**F1-score**
1	0.5987	0.7282	0.6571
2	0.4562	0.3303	0.3832
3	0.4211	0.3156	0.3608
4	0.4135	0.3140	0.3570
5	0.3964	0.3030	0.3434
6	0.4017	0.2728	0.3249
7	0.3847	0.2705	0.3176
8	0.3915	0.3236	0.3543
9	0.4302	0.3655	0.3952
10	0.6014	0.7675	0.6744
Micro	0.5197	0.5197	0.5197
Macro	0.4495	0.3991	0.4168

Tabla 1: Baseline. Resultados empleando tf-idf y LinearSVM

La tabla 1 muestra los resultados del sistema baseline. los mejores resultados se obtienen en aquellas clases con mayor representación y más polarizadas, obteniendo un 67.44 % y 65.71 % de F1 para las clases 10 y 1 respectivamente. De igual forma, los peores resultados son aquellos obtenidos para las clases menos representadas, llegando a obtener un 32.49 % de F1 para la clase 6 y un 31.76 % para la clase número 7.

Las tablas 2 y 3 muestran los resultados obtenidos por el modelo CNN. En la tabla 2 se muestran los resultados obtenidos al mantener los word embedding estáticos en el entrenamiento del modelo CNN. Mientras, en la tabla 3, podemos observar los resultados obtenidos al permitir que los valores de los word embeddings sean ajustados durante el entrenamiento de la red.

Como podemos comprobar en la tabla 2, los resultados siguen el mismo patrón que los obtenidos con el modelo SVM. Los mejores resultados son aquellos obtenidos para las clases mayoritarias (68.07 % y 64.74 %). No obstante, los resultados de las clases minoritarias son incluso más bajos que los proporcionados por el modelo SVM.

Si comparamos el modelo SVM y el modelo CNN con word embeddings estáticos descritos hasta el momento, el modelo SVM obtiene un 51.97 % de micro-F1, mientras que el modelo CNN obtiene un 47.50 % de micro-

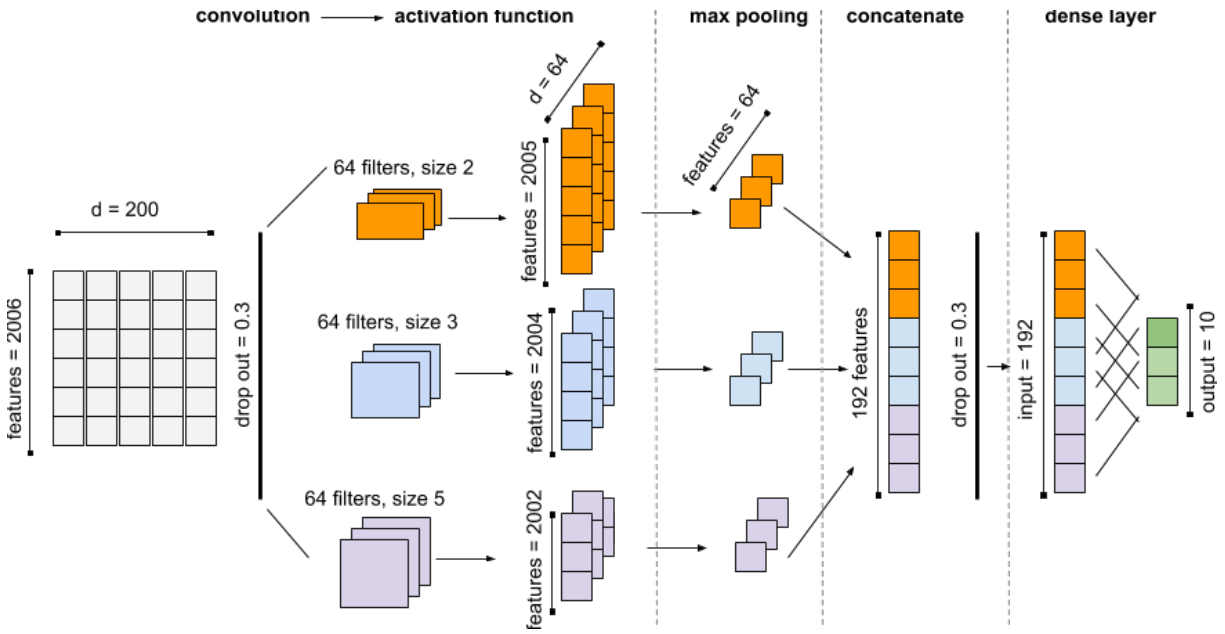

Figura 2: Arquitectura de nuestro sistema

Label	Precision	Recall	F1-score
1	0.5339	0.8223	0.6474
2	0.2690	0.1195	0.1655
3	0.2592	0.1247	0.1684
4	0.2416	0.0820	0.1224
5	0.2473	0.1524	0.1886
6	0.2047	0.0897	0.1247
7	0.2625	0.0984	0.1431
8	0.2815	0.2695	0.2754
9	0.3737	0.2528	0.3016
10	0.5816	0.8205	0.6807
Micro	0.4750	0.4750	0.4750
Macro	0.3255	0.2832	0.2818

Tabla 2: Resultados con CNN utilizando word embeddings estáticos (no entrenables)

F1. Estos resultados pueden deberse a que las características extraídas de la matriz de embeddings por los filtros convolucionales no fueran suficientes a la hora de discriminar entre una clase u otra.

Sin embargo, el modelo CNN con word embeddings no estáticos (tabla 3) mejora significativamente los resultados de los modelos anteriores. En este caso, los resultados obtenidos por las clases mayoritarias continúan siendo superiores al resto clases. No obstante, existe una notable mejoría en lo que respecta a las clases minoritarias, obteniendo finalmente una micro-F1 de 66.72 % y una macro-F1 de 61.81 % sobre el dataset de test.

Label	Precision	Recall	F1-score
1	0.7119	0.7556	0.7331
2	0.6227	0.5771	0.5991
3	0.6254	0.5664	0.5945
4	0.6451	0.5413	0.5887
5	0.6035	0.5572	0.5794
6	0.6444	0.5139	0.5718
7	0.5903	0.5290	0.5579
8	0.5822	0.5770	0.5796
9	0.6026	0.6214	0.6119
10	0.7454	0.7858	0.7651
Micro	0.6672	0.6672	0.6672
Macro	0.6374	0.6025	0.6181

Tabla 3: Resultados con CNN utilizando word embeddings entrenables

	Micro	Macro
SVM(Tf-idf)	0.5197	0.4168
CNN(w2v-estático)	0.4750	0.2818
CNN(w2v-entrenable)	**0.6672**	**0.6181**

Tabla 4: Resultados de micro y macro-F1 de los diferentes modelos. Los valores más altos están resaltados en negrita

5 Discusión

Como se ha mostrado anteriormente, nos encontramos ante un problema de multiclasificación con las clases desbalanceadas. Esto resulta problemático a la hora de clasificar correctamente aquellos comentarios cuya clase está escasamente representada en el dataset de entrenamiento.

La representación de los textos utilizando una bolsa de palabras o tf-idf produce vectores de elevadas dimensiones donde los datos están muy dispersos. Teniendo todo esto en cuenta, emplear un clasificador SVM podría proporcionar buenos resultados (Joachims, 1998).

Cuando comparamos el modelo generado con embeddings estáticos, nos encontramos con que los resultados son incluso inferiores a los proporcionados por el clasificador SVM. A pesar de que el modelo de word embeddings incluya un 82.04 % de las palabras del corpus, las características extraídas por los filtros convolucionales no parecen ser lo suficientemente discriminatorias como para alcanzar los resultados del modelo SVM.

No obstante, al permitir ajustar los word embeddings durante el entrenamiento de la red, podemos comprobar que una arquitectura sencilla de CNN es capaz de alcanzar una mejoría significativa, incluso en las clases menos representadas. En la tabla 4, vemos como el clasificador CNN nos proporciona un 66.72 % de micro-F1 frente a un 51.97 % obtenido por el clasificador SVM. Al realizar la media entre sus F1-score por igual sin ninguna ponderación, obtenemos un 61.81 % de macro-F1 frente al 41.68 % del clasificador SVM, corroborando que el modelo CNN mejora el desempeño a la hora de predecir las clases menos representativas.

6 Conclusión

Debido al creciente volumen de datos disponibles en internet, cada vez un mayor número de pacientes buscan opiniones sobre la eficacia y los posibles efectos de sus tratamientos. Por otro lado, el seguimiento y análisis de este tipo de opiniones sobre los fármacos, aporta una valiosa información complementaria a la hora de detectar posibles efectos adversos, que no fueron detectados durante los ensayos clínicos.

Mientras que la mayoría de los trabajos que han investigado en análisis de sentimiento en el dominio de salud, se han centrado en la clasificación de comentarios en dos o tres clases (positiva, negativa y neutra), en nuestro trabajo abordamos una tarea más complicada al tratar de clasificar los comentarios en 10 clases distintas (cada una de ella refleja el grado de satisfacción del paciente con respecto al fármaco). Además, otro reto importante es que debido al elevado número de clases, muchas clases están poco representadas. Nuestro mejores resultados son obtenidos con el modelo CNN (61.81 % de macro-F1), frente al enfoque tradicional de SVM con bolsa de palabras tf-idf (41.68 % de macro-F1).

Como trabajo futuro, exploraremos otras arquitecturas de aprendizaje profundo, tales como CNN con un mayor número de capas convolucionales, redes recurrentes y arquitecturas híbridas. Además, integraremos información semántica mediante el uso de modelos de embeddings de conceptos. Para abordar el problema de desbalanceo de datos, aplicaremos técnicas de sampling. También nos planteamos reducir el número de clases de 10 a 3 (positivo, negativo y neutro).

Agradecimientos

Este trabajo ha sido financiado por el Programa de Investigación del Ministerio de Economía y Competitividad del Gobierno de España (proyecto DeepEMR TIN2017-87548-C2-1-R).

Bibliografía

Baccianella, S., A. Esuli, y F. Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. En *Lrec*, volumen 10, páginas 2200–2204.

Boureau, Y.-L., F. Bach, Y. LeCun, y J. Ponce. 2010. Learning mid-level features for recognition.

Collobert, R. y J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. En *Proceedings of the 25th international conference on Machine learning*, páginas 160–167. ACM.

De Marneffe, M.-C., B. MacCartney, y C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy,

Mayo. European Language Resources Association (ELRA).

Denecke, K. y Y. Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 64(1):17 – 27.

Er, M. J., Y. Zhang, N. Wang, y M. Pratama. 2016. Attention pooling-based convolutional neural network for sentence modelling. *Information Sciences*, 373:388–403.

Frakes, W. B. y R. Baeza-Yates. 1992. *Information retrieval: Data structures & algorithms*, volumen 331. Prentice Hall Englewood Cliffs, NJ.

Gopalakrishnan, V. y C. Ramaswamy. 2017. Patient opinion mining to analyze drugs satisfaction using supervised learning. *Journal of applied research and technology*, 15(4):311–319.

Gräßer, F., S. Kallumadi, H. Malberg, y S. Zaunseder. 2018. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. En *Proceedings of the 2018 International Conference on Digital Health*, páginas 121–125. ACM.

Ide, H. y T. Kurita. 2017. Improvement of learning for cnn with relu activation by sparse regularization. En *2017 International Joint Conference on Neural Networks (IJCNN)*, páginas 2684–2691. IEEE.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. En *European conference on machine learning*, páginas 137–142. Springer.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P. y J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, y J. Dean. 2013. Distributed representations of words and phrases and their compositionality.

Na, J. y W. Y. M. Kyaing. 2015. Sentiment analysis of user-generated content on drug review websites. *Journal of Information Science Theory and Practice*, 3(1):6–23.

Na, J.-C., W. Y. M. Kyaing, C. S. Khoo, S. Foo, Y.-K. Chang, y Y.-L. Theng. 2012. Sentiment classification of drug reviews using a rule-based linguistic approach. En *International conference on asian digital libraries*, páginas 189–198. Springer.

Nair, V. y G. E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. En *Proceedings of the 27th international conference on machine learning (ICML-10)*, páginas 807–814.

Pang, B. y L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pyysalo, S., F. Ginter, H. Moen, T. Salakoski, y S. Ananiadou. 2013. Distributional semantics resources for biomedical text processing.

Suárez-Paniagua, V. y I. Segura-Bedmar. 2018. Evaluation of pooling operations in convolutional architectures for drug-drug interaction extraction. *BMC bioinformatics*, 19(8):209.

Wang, J., Z. Wang, D. Zhang, y J. Yan. 2017. Combining knowledge with deep convolutional neural networks for short text classification. En *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, páginas 2915–2921.

Wilson, T., J. Wiebe, y P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. En *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Yadav, S., A. Ekbal, S. Saha, y P. Bhattacharyya. 2018. Medical sentiment analysis using social media: Towards building a patient assisted system. En *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.