

Aplicación de un modelo híbrido de aprendizaje profundo para el Análisis de Sentimiento en Twitter

Application of a hybrid deep learning model for Sentiment Analysis in Twitter

Rosa Montañés, Rocío Aznar, Rafael del Hoyo

Grupo de Big Data y Sistemas Cognitivos
ITAINNOVA (Instituto Tecnológico de Aragón)
C/ María de Luna, nº 7. 50018 Zaragoza
{rmontanes,raznar,rdelhoyo}@itainnova.es

Resumen: En este artículo se describe la participación de ITAINNOVA en la tarea de análisis de sentimiento a nivel de Tweet dentro del taller TASS 2018. Este trabajo pretende explorar modelos presentes en el estado del arte actual del aprendizaje profundo aplicado al modelado y clasificación de texto. Se ha analizado el uso de modelos de redes convolucionales (CNN), Long short Term Memory (LSTM), LSTM bidireccionales (BI-LSTM) y una aproximación híbrida entre CNN y LSTM para su uso en el análisis de sentimiento en Twitter. Se ha optado por la combinación CNN-LSTM ya que integra los beneficios de ambos modelos. Finalmente se presentan los resultados obtenidos y se plantea una posible línea de trabajo futura que combine el uso de esta arquitectura con el algoritmo de representación de texto que presentamos en la anterior edición del TASS.

Palabras clave: Análisis de sentimiento, Twitter, Aprendizaje profundo

Abstract: This paper describes the participation of ITAINNOVA at sentiment analysis at Tweet level task within TASS 2018 workshop. This work explores current state of the art models used in deep learning for modelling and classification tasks over text. It analyzes convolutional neural models (CNN), Long short Term Memory (LSTM), Bidirectional LSTM (BI-LSTM) and an hybrid approach of CNN-LSTM, for its use in sentiment analysis on Twitter data. CNN-LSTM combination has been chosen as it integrates the benefits provided from both models. Finally, obtained results are presented and a possible future work line which combines this architecture with the algorithm presented in the previous TASS edition.

Keywords: Sentiment analysis, Twitter, Deep learning

1 Introducción

El taller de Análisis de Sentimiento celebrado en el marco del congreso internacional de la sociedad española para el procesamiento del lenguaje natural (SEPLN), es un importante punto de encuentro para profesionales e investigadores en el que compartir y discutir nuevos avances en el campo del PLN en general y del análisis de sentimiento en particular (Martínez-Cámara et al., 2018). En esta edición se proponen cuatro tareas distintas, dando un salto cualitativo hacia el análisis semántico, en busca de modelos de comprensión de lenguaje más completos y realistas.

El grupo de Big Data y Sistemas Cogni-

tivos de ITAINNOVA ha participado en la tarea 1: “Análisis de sentimiento a nivel de Tweet”, cuyo objetivo principal es la clasificación de la opinión expresada en estos textos cortos en español. La novedad principal de esta edición viene dada por la expansión del corpus de tweets utilizado en ediciones previas, el corpus InterTASS. Este año se han incorporado publicaciones escritas por usuarios de Perú y Costa Rica, variaciones del español que presentan ciertas diferencias léxicas y gramaticales que han de ser consideradas en el desarrollo de sistemas de procesamiento de lenguaje natural. Nuestro grupo se ha enfocado en el desarrollo de un sistema de clasi-

ficación multiclase basado exclusivamente en la variedad de español hablado en España, utilizando para ello el dataset *InterTASS ES* junto con un subconjunto del corpus general utilizado desde las primeras ediciones, que permitirá predecir la polaridad de los tweets en base a cuatro niveles: P (Positiva), N (Negativa), NEU (Neutra), NONE (sin opinión). Finalmente los resultados obtenidos nos permitirán comprobar la capacidad de generalización del modelo a las otras variantes de español.

Nuestro grupo cuenta con experiencia de participación en ediciones anteriores de talleres de análisis de opinión (Montañés Salas et al., 2017)(del Hoyo Alonso et al., 2015). En ellas verificamos y desarrollamos recursos para el procesado del texto previo a su clasificación, tales como la generación de un diccionario afectivo, y experimentamos con el algoritmo FastText, capaz de aprender eficientemente representaciones de palabras partiendo de corpus de tamaño reducido. En el taller de 2017 (Díaz-Galiano et al., 2018) se presentaron una amplia variedad de soluciones basadas en redes neuronales profundas (deep learning) demostrando el gran potencial de las mismas en tareas de análisis de lenguaje. Las conclusiones extraídas en dicha edición han motivado que en nuestra participación en el taller de 2018 hayamos centrado nuestro desarrollo en el estudio e implementación de un algoritmo basado en una combinación de este tipo de redes neuronales. Los resultados obtenidos nos permiten comprobar la capacidad de aprendizaje de este tipo de algoritmos en esta tarea de procesamiento del lenguaje natural.

Este artículo se organiza como sigue. En la sección 2 se presenta el conjunto de trabajos relacionados que han inspirado la aproximación propuesta. En la sección 3 se describe la implementación realizada y en la sección 4 se presentan los resultados experimentales obtenidos. Finalmente, en la sección 5 se recogen las conclusiones de nuestro estudio práctico y se presentan posibles líneas de trabajo futuro a abordar.

2 Trabajos relacionados

La investigación actual en el área de deep learning ha dado lugar al desarrollo de diversos tipos de redes neuronales que presentan unas características óptimas para el aprendizaje de ciertos tipos de información. Por

ejemplo, las redes convolucionales (CNNs) han demostrado su excepcional capacidad de aprendizaje en el área de visión por computador, donde es necesario extraer correlaciones locales en estructuras espacio-temporales para el reconocimiento y clasificación de objetos (Galteri et al., 2017). Del mismo modo, este tipo de redes, aplicadas al área de PLN, son capaces de extraer automáticamente los vectores de características sobre n-gramas mediante filtrado convolucional y *pooling* consiguiendo aprender relaciones de más alto nivel entre los componentes del texto, tal como demuestra en sus experimentos Kim (2014). Existen otro tipo de arquitecturas más enfocadas al modelado de series temporales, éstas son las conocidas LSTM (Long Short-term Memory RNN), que incorporan un novedoso mecanismo de memoria para el modelado de dependencias a largo plazo lo que resuelve algunos de los problemas derivados del cálculo del gradiente en redes recurrentes más simples, convirtiéndose en una de las mejores opciones para el desarrollo de modelos de minería de opinión (Liu, Joty, y Meng, 2015). Como ampliación a este tipo de redes, Zhou et al. (2016) introdujo las LSTM bidireccionales (BiLSTM), que añaden la capacidad de preservar información pasada y también futura. Estos modelos basan su funcionamiento en el uso de los populares Word Embeddings (Mikolov et al., 2013), siendo uno de los métodos de representación de la información textual en vectores n-dimensionales más eficiente y ampliamente utilizado en la actualidad.

Adicionalmente, en el campo del procesado de texto, la necesidad de realizar un correcto tratamiento de este tipo de datos para su posterior uso en algoritmos de Inteligencia Artificial, ha sido estudiada extensamente, desde el punto de vista del análisis de sentimiento en general (Haddi, Liu, y Shi, 2013) como en el caso concreto de análisis de sentimientos sobre textos cortos informales como son los tweets (Sing y Kumari, 2016). En ambos artículos se ofrecen conclusiones que demuestran la importancia y el papel que juega el preprocesamiento de textos en tareas como el análisis de sentimientos, con el objetivo de eliminar ruido y reducir la dimensionalidad del problema para mejorar la clasificación.

La solución propuesta en el taller de 2017 (Montañés Salas et al., 2017) tenía su fundamento en dos de las anteriores aproximacio-

nes, por un lado la aplicación de un algoritmo basado en el uso de representaciones de palabras (word embeddings enriquecidos con información de n-gramas) y por otro lado la aplicación de una fase previa de procesamiento de texto que mejoró ligeramente los resultados obtenidos, por lo que, siguiendo un planteamiento similar, se ha decidido continuar trabajando con dichas aproximaciones e ir un paso más lejos con el uso de arquitecturas de deep learning.

3 Solución propuesta

Previamente a la elección del algoritmo final utilizado, se han realizado varias pruebas de concepto con la intención de verificar la benevolencia de la hibridación de técnicas convolucionales y recurrentes en el procesamiento de lenguaje natural. Para ello, por un lado, se han estudiado varios algoritmos convolucionales y recurrentes, en particular, una red neuronal convolucional (CNN), una red neuronal recurrente LSTM y una red recurrente LSTM bidireccional (BLSTM) y, por otro lado, el algoritmo híbrido que combina la red convolucional con la red recurrente LSTM. Los conjuntos de entrenamiento y desarrollo usados son los utilizados en la tarea (InterTASS ES). Las métricas de accuracy se muestran en la tabla 1.

Red	Training	Development
CNN	0.695114	0.463303
LSTM	0.654723	0.441964
CLSTM	0.715961	0.471964
BLSTM	0.682736	0.426339

Tabla 1: Pruebas previas

Como se ha adelantado en la sección anterior, las redes neuronales CNN y LSTM son algoritmos que ofrecen resultados satisfactorios en el estado del arte actual del análisis de opinión. La solución propuesta permite combinar estos dos algoritmos, beneficiándose de las ventajas de ambos: la extracción de términos relevantes y correlaciones a nivel local de las redes convolucionales junto con la capacidad de aprendizaje sobre secuencias de datos temporales y la estabilidad de las LSTMs. El algoritmo propuesto se basa en la publicación (Zhou et al., 2015) y se explica con más detalle en la siguiente sección.

Los resultados obtenidos que se muestran

en la tabla 1 ponen de manifiesto que la solución híbrida (CLSTM) es la mejor opción.

La implementación del sistema se ha llevado a cabo en Python, haciendo uso de la librería Tensorflow con soporte para GPU.

3.1 Modelo C-LSTM

A la vista de los resultados expuestos, el sistema elegido para la participación en el taller de este año es una implementación del modelo C-LSTM descrito por Zhou et al. (2015). Éste se compone principalmente de 2 redes neuronales: una red neuronal convolucional (CNN) y una red neuronal recurrente de memoria a corto y largo plazo (LSTM). En la Figura 1 se muestra un diagrama de la arquitectura CLSTM.

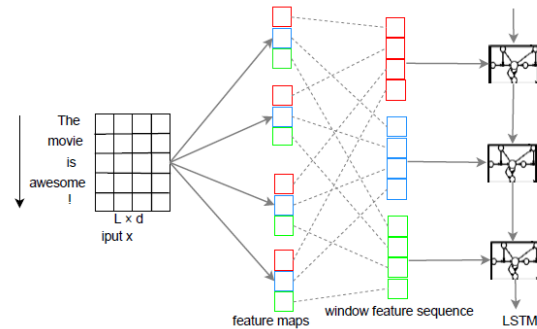


Figura 1: Arquitectura CLSTM.

Como paso previo al entrenamiento del modelo de análisis de sentimiento, el sistema calcula de forma no supervisada la representación vectorial del vocabulario presente en el corpus, creando un conjunto de word embeddings con los que se generará la representación matricial de las sentencias que alimentan el sistema. La red convolucional que recibe dicho input, aplica un filtrado convolucional según los tamaños de filtros definidos, generándose varios mapas de *features* que son reorganizados secuencialmente obteniendo una representación del texto a un nivel conceptual más alto que los embeddings. La red LSTM, compuesta por un conjunto de celdas que se definirá en la configuración, recibe estas *features* como entrada, y se encarga de determinar qué información almacena en la memoria de cada celda o descarta en cada paso temporal de la secuencia, capturando de este modo dependencias a corto y largo plazo en las sentencias. La salida del modelo viene dada por la última capa oculta de la red LSTM.

3.2 Preprocesado de texto

Adicionalmente, y del mismo modo que en la edición TASS 2017, se ha realizado un preprocesado de los textos, aplicando las siguientes técnicas:

- Eliminación de patrones como URLs, emails, menciones, etc., típicos en publicaciones de redes sociales.
- Conversión a minúsculas.
- Lematización en español.
- Eliminación de stopwords.
- Sinónimos basados en diccionarios emocionales.

Las etapas de lematización, eliminación de stopwords y aplicación de sinónimos basados en diccionarios emocionales, utilizan recursos propios desarrollados para el español hablado en España, por lo que se ha llevado a cabo esta fase en el experimento monolingüe con el objetivo de comprobar su efectividad en el contexto de las redes neuronales profundas.

4 Resultados experimentales

En la Tabla 2 se muestran los resultados obtenidos en las subtareas evaluadas por ITAINNOVA.

Los dos modelos (cl-base y cl-proc) se han entrenado con el mismo conjunto de hiperparámetros, elegidos a partir de las recomendaciones expuestas en (Zhou et al., 2015) y realizando una búsqueda empírica con pequeñas variaciones en los valores:

- Tamaño del embedding: 256
- Mínima frecuencia de palabra: 0
- Número de filtros (CNN): 128
- Tamaño de los filtros (CNN): 2,3,4
- Número de capas: 2 (1 CNN + 1 LSTM)
- Learning rate: 0.001
- Tamaño del batch: 32
- Número de épocas: 100

Si bien nuestro sistema no ha quedado dentro del ranking de los 10 mejores respecto al resto de sistemas presentados en esta edición, los resultados obtenidos se aproximan a lo resultados publicados en el paper

de referencia para la clasificación de polaridad en 5 niveles. Tras analizar los resultados de la evaluación observamos que el modelo presenta ciertas deficiencias en la clasificación de publicaciones como NEU y NONE, clases semánticamente y conceptualmente muy próximas y por tanto difíciles de clasificar. Aunque pensamos que su diferenciación podría ser factible usando un modelo de representación de palabras que disponga de un vocabulario más extenso modelizando de forma más precisa el lenguaje español.

Asimismo, el experimento monolingüe realizado, nos indica que en este caso, el preprocesado de los textos no mejora la predicción del sentimiento. Idea que ya apuntaba Zhou et al. (2015) en su publicación mencionando la independencia del modelo de conocimiento del lenguaje a través de recursos externos ni necesidad de un complejo procesado que reduzca excesivamente la dimensionalidad del sistema.

En cuanto a los experimentos entre variantes del español, aplicando el modelo de España a los textos de Perú (PE) y Costa Rica (CR), se observa que el sistema es bastante robusto a las diferencias léxicas y gramaticales que se pueden presentar en estos textos cortos.

5 Conclusiones y trabajo futuro

En este artículo se ha presentado la aproximación llevada a cabo por el equipo de ITAINNOVA en el TASS 2018. Dicha aproximación tenía como objetivo principal la aplicación de técnicas basadas en redes neuronales profundas para la clasificación de la polaridad en tweets. Los resultados obtenidos de la aplicación de algoritmos de CNN y LSTM y un algoritmo híbrido combina ambos algoritmos mostraron obtener una mayor métrica de accuracy en el algoritmo híbrido, razón por la que dicho algoritmo fue el elegido como solución propuesta a la tarea.

El conjunto de entrenamiento sobre el que se ha trabajado es el INTERTASS ES unificado con un subconjunto de las publicaciones del corpus general. Los resultados obtenidos en la tarea 'Monolingual ES' muestran que el preprocesado de textos considerado no mejora la predicción del sentimiento. Por otro lado, se observa que el algoritmo entrenado solamente con el conjunto de datos en español de España produce resultados similares en el conjunto de test de textos de Perú y Costa Ri-

	<i>Algoritmo</i>	Macro-F1	Accuracy
Monolingual ES	cl-base	0.383	0.433
	cl-proc	0.320	0.395
Cross-lingual PE	cl-base	0.367	0.382
Cross-lingual CR	cl-base	0.409	0.440

Tabla 2: Resultados experimentales

ca. No obstante, la solución propuesta no ha conseguido quedarse en el ranking de las 10 propuestas con mejores valores en las métricas, obteniendo incluso métricas inferiores a las obtenidas en la solución que propusimos en la tarea de 2017 (Montañés Salas et al., 2017).

En base a los resultados obtenidos en dicha tarea y en la tarea de 2017 en la que participamos, vemos la necesidad de aumentar el corpus en este tipo de redes o la utilización de modelos de word embedding más extensos para introducir más información al sistema. Consideramos explorar una posible línea de trabajo futuro que podría conducir a la participación en esta tarea de análisis de sentimiento de próximas ediciones con un algoritmo que integre las aspectos más valiosos de las aproximaciones consideradas. En concreto, se aplicaría el algoritmo de FastText utilizado en la tarea de 2017 para la representación vectorial de los tweets que además ofrece una eficiencia computacional bastante elevada. Posteriormente se aplicaría el algoritmo híbrido de CNN y LSTM para la clasificación, redes neuronales profundas con gran potencial en tareas de procesamiento de lenguaje natural y que, como se ha demostrado en nuestros resultados y en (Zhou et al., 2015) mejora a los algoritmos simples (no-híbridos) de redes CNN y LSTM. Se piensa que los resultados obtenidos en esta nueva línea podrían mejorar considerablemente los conseguidos hasta ahora.

Bibliografía

- del Hoyo Alonso, R., V. Rodrigálvarez Chamarro, J. Veá-Murguía Merck, y R. M. Montañés Salas. 2015. Ensemble algorithm with syntactical tree features to improve the opinion analysis. En *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, volumen 1397 de *CEUR Workshop Proceedings*, páginas 53–58, Alicante, Spain, September. CEUR-WS.
- Díaz-Galiano, M. C., E. Martínez-Cámara, M. Á. G. Cumberras, M. G. Vega, y J. Villena-Román. 2018. The democratization of deep learning in tass 2017. *Procesamiento del Lenguaje Natural*, 60(0):37–44.
- Galteri, L., L. Seidenari, M. Bertini, y A. Del Bimbo. 2017. Spatio-temporal closed-loop object detection. *IEEE Transactions on Image Processing*, PP:1–1, 01.
- Haddi, E., X. Liu, y Y. Shi. 2013. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26 – 32. First International Conference on Information Technology and Quantitative Management.
- Kim, Y. 2014. Convolutional neural networks for sentence classification. En *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, páginas 1746–1751, September.
- Liu, P., S. R. Joty, y H. M. Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. En *EMNLP*, páginas 1433–1443. The Association for Computational Linguistics.
- Martínez-Cámara, E., Y. Almeida Cruz, M. C. Díaz-Galiano, S. Estévez Velarde, M. A. García-Cumberras, M. García-Vega, Y. Gutiérrez Vázquez, A. Montejó Ráez, A. Montoyo Guijarro, R. Muñoz Guillena, A. Piad Morffis, y J. Villena-Román. 2018. Overview of TASS 2018: Opinions, health and emotions. En E. Martínez-Cámara Y. Almeida Cruz

- M. C. Díaz-Galiano S. Estévez Velarde M. A. García-Cumbreras M. García-Vega Y. Gutiérrez Vázquez A. Montejó Ráez A. Montoyo Guijarro R. Muñoz Guillena A. Piad Morffis, y J. Villena-Román, editores, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volumen 2172 de *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.
- Mikolov, T., K. Chen, G. Corrado, y J. Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- Montañés Salas, R. M., R. del Hoyo Alonso, J. Veá-Murguía Merck, R. Aznar Gimeno, y F. J. Lacueva-Pérez. 2017. Fasttext como alternativa a la utilización de deep learning en corpus pequeños. En *Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN*, volumen 1896 de *CEUR Workshop Proceedings*, páginas 65–69, Murcia, Spain, September. CEUR-WS.
- Sing, T. y M. Kumari. 2016. Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science*, 89:549–554. Twelfth International Multi-Conference on Information Processing (IMCIP-2016).
- Zhou, C., C. Sun, Z. Liu, y F. Lau. 2015. A c-lstm neural network for text classification. *CoRR*, abs/1511.08630.
- Zhou, P., W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, y B. Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, páginas 207–212. Association for Computational Linguistics.