Evaluación de los nuevos enfoques de aprendizaje automático aplicados a la clasificación de correos electrónicos en español

Juan M. Fernandez[1,2], Nicolás Cavasín[3], and Marcelo Errecalde[4]

[1] Master Student at Computer Science School, La Plata National University
[2] Professor and Researcher at Luján National University
[3] Luján National University
[4] Professor and Researcher at LIDIC, San Luis National University
{jmfernandez, ncavasin}@unlu.edu.ar, merreca@unls.edu.ar

Abstract. En la actualidad se generan millones de datos cada día y su aprovechamiento e interpretación se han vuelto fundamentales en todos los ámbitos. Sin embargo, la mayor parte de esta información posee un formato textual, sin la estructura ni la organización de las bases de datos tradicionales, lo cual representa un enorme desafío.
A lo largo del tiempo, se han desarrollado diversos enfoques para la representación de texto y la generación de conocimiento a partir de esta fuente de datos, desde la aparición del modelo de representación vectorial hasta los enfoques basados en aprendizaje profundo y transformadores, pasando por las técnicas basadas en redes neuronales y *embbedings*. No obstante, la mayoría de los recursos disponibles se centran en el idioma inglés, existiendo alternativas reducidas para el resto de los idiomas. En este trabajo, a partir de un caso de estudio basado en correos electrónicos, se estudia la performance de tres de los enfoques de aprendizaje automático aplicado a clasificación de textos más relevantes de la disciplina a efectos de verificar si la aparición de nuevos enfoques trae aparejado un mejor rendimiento de los modelos de clasificación para el idioma español.

Keywords: Text Classification · BERT · LSTM · Word2Vec · SVM

1 Introducción

Producto de la masificación del acceso internet, se generan millones y millones de datos cada día y su aprovechamiento e interpretación se han vuelto fundamentales en todos los ámbitos. La recuperación de información y la minería de textos se han convertido a lo largo de los años en campos de investigación muy populares, especialmente en el campo de la clasificación de textos [5]. En este sentido, es posible encontrar estudios sobre clasificación de textos incluso en el año 1957, donde se realizaron trabajos de investigación que proponían la clasificación de textos utilizando el método de frecuencia de palabras [9]. Desde entonces, se han desarrollado diversos enfoques para la representación de texto y la generación

de conocimiento a partir de esta fuente de datos. No obstante, la mayoría de los recursos disponibles se centran en el idioma inglés, existiendo alternativas reducidas para el resto de los idiomas. Al mismo tiempo, no abundan trabajos respecto a comparaciones empíricas del redimiento de estos nuevos enfoques en idiomas como el español, donde muchas veces no existen recursos disponibles fiables para la implementación de los nuevos abordajes de clasificación de texto. En este trabajo, se presentan experimentos que comparan la performance de tres de los enfoques de aprendizaje automático aplicado a clasificación de textos más relevantes de la disciplina a efectos de indagar respecto a los aportes reales que estos introducen para idiomas diferentes del inglés.

2 Trabajos relacionados

Como se plantea antes, si bien en los últimos 60 años proliferaron diversos abordajes para la clasificación automática de textos, no abundan los trabajos de investigación que indaguen sobre la performance de las diferentes estrategias en idiomas diferentes al inglés. Al mismo tiempo, resulta abrumadora la cantidad de investigaciones que proponen diferentes abordajes para la clasificación automática de textos. A continuación se realiza una breve reseña de las tres estrategias que se utilizan en el marco de esta investigación para el estudio comparado.

#1: BoW+SVM Uno de los métodos más simples para la representación de documentos, y también uno de los más antiguos, es el denominado *Bag of Words* (bolsa de palabras) o modelo de espacio vectorial [11]. Esta técnica genera un vector que representa un documento usando el recuento de frecuencia de cada término en el documento [6] y se denomina así puesto que las palabras son tomadas como características y los documentos se tratan simplemente como colecciones de palabras desordenadas [8]. Esta estrategia de representación tiene como ventaja la simplicidad y a su vez la posibilidad de aplicar a la representación resultante cualquiera de las técnicas de clasificación existentes. Una de las usualmente utilizadas es máquina de vectores de soporte (SVM), presentada a mediados de 1990, fue ganando popularidad debido a algunas características atractivas y su rendimiento empírico. SVM se basa en el principio de Minimización del Riesgo Estructural (SRM) de la teoría del aprendizaje estadístico, el cual consiste en encontrar un hiperplano óptimo para el que se pueda garantizar el error verdadero más bajo [7]. Para el cálculo de distancias y la búsqueda de los hiperplanos, las SVM utilizan funciones denominadas kernels [12].

#2: Word2Vec+LSTM Una línea de investigación bastante actual comprende la utilización de información contextual junto con modelos simples de redes neuronales para obtener representaciones de palabras y frases en el espacio vectorial[14]. Uno de los modelos más populares es Word2Vec, el cual dispone de dos arquitecturas diferentes, a saber, CBoW y Skip-gram [10]. Estos modelos de incrustaciones de palabras usualmente se complementan con redes neuronales recurrentes como LSTM (*Long short-term memory*). Estas redes neuronales proveen dos características que mejoran sustancialmente el rendimiento de las

redes neuronales convencionales para el tratamiento de texto: permiten identificar el orden de las secuencias de texto en los documentos y pueden procesar documentos de diferentes longitudes [1].

#3: BERT Como una evolución a la estrategia anterior, en 2017, se propone una nueva arquitectura de red neuronal, más simple y paralelizable, denominada *Transformer* [13], basada únicamente en mecanismos de atención, prescindiendo por completo de recurrencia y convoluciones. Estos mecanismos de atención se pueden describir como la asignación de una consulta y un conjunto de pares clave-valor a una salida, donde la consulta, las claves, los valores y la salida son todos vectores. A partir de esta lógica nace, lo que en la literatura se conoce como el estado del arte actual de los modelos de representación del lenguaje, denominado BERT *Bidirectional Encoder Representations from Transformers* [4]. Sintéticamente, este framework consta de dos pasos: pre-entrenamiento inicial y ajuste fino posterior. Durante el entrenamiento previo, el modelo se entrena con datos sin etiquetar en diferentes tareas. Luego, para el ajuste fino, el modelo BERT se inicializa primero con los parámetros del modelo pre-entrenado, los cuales se ajustan en esta etapa utilizando datos etiquetados de las tareas posteriores.

3 Metodología de la investigación

Para la ejecución de los experimentos se utilizó un conjunto de datos generado a partir consultas académicas realizadas mediante correo electrónico por parte estudiantes de la Universidad Nacional de Luján al staff administrativo, sobre trámites derivados de la actividad académica. Sobre una base de 24700 correos, se seleccionaron 1000 interacciones que fueron etiquetadas en torno al tema de la consulta por un experto del dominio. Para los experimentos se utilizaron los correos de consulta originales sin supervisión humana sobre errores semánticos ni de sintaxis. Para el enfoque de aprendizaje basado en **BoW+SVM** se normalizó el texto, se eliminaron palabras vacías, se incorporaron atributos estáticos (largo de la consulta, utilización de signos de puntuación, etc) y se experimentó con diferentes variaciones de *n-gramas* de palabras y caracteres. En cambio, para los enfoques basados en **Word2Vec+LSTM** y **BERT** se utilizaron las secuencias de texto relativas a la consulta únicamente con el texto normalizado y se eliminaron palabras vacías solo para **Word2Vec+LSTM** puesto que para **BERT** en las pruebas experimentaba una baja de la performance. Para el enfoque **Word2Vec+LSTM** se utilizaron incrustaciones de palabras pre-entrenados disponibles para el idioma español[2]. En cuanto a **BERT**, se experimentó con dos modelos pre-entrenados, uno para nativo para el idioma español [3] y otro, denominado *Multilingual* [4], desarrollado para múltiples idiomas. Para la evaluación de los modelos se utilizó una validación cruzada con *5-fold* sobre el 80% de las instancias en la etapa de entrenamiento mientras que luego se *testeó* el modelo mediante las 20% restante de las instancias a partir de las métricas *accuracy*, *precision*, *recall* y *f1-score*.

4 Experimentos

En todos los casos se realizó una búsqueda de los mejores hiperparámetros para cada estrategia, obteniendo los siguientes resultados[5]:

Table 1. Resultados de los experimentos con las distintas estrategias de aprendizaje.

Estrategia	**Accuracy**	**Precision**	**Recall**	**F1-score**
BoW+SVM	0.870	0.862	0.830	0.840
Word2Vec+LSTM	0.835	0.814	0.841	0.820
BERT (Multilingual)	0.860	0.838	0.842	0.840
BERT (BETO)	**0.890**	**0.870**	**0.885**	**0.876**

Los resultados obtenidos muestran que el abordaje más efectivo para la clasificación de este conjunto de datos es **BERT**, con el modelo pre-entrenado para el idioma español. No obstante, las diferencias observadas entre las métricas obtenidas respecto a **BOW+SVM** no supera para ninguna métrica el 0.03.

5 Conclusiones

A partir de los resultados obtenidos, y teniendo en cuenta que entre entre el abordaje **BOW+SVM** y **BERT** han pasado cerca de treinta años de evolución en esta disciplina, se verifica que los métodos tradicionales de representación y clasificación siguen siendo una opción competitiva. Sin embargo, es importante tener en cuenta que los correos electrónicos en general, y este conjunto de datos en particular, tienen características que no benefician a los modelos más actuales dada la informalidad y los errores de sintaxis propios de la dinámica de este medio de comunicación, por lo cual se espera que en conjuntos de datos con textos más depurados las diferencias se amplien entre los modelos más tradicionales y los más actuales. A su vez, podría verificarse esta situación a partir del preprocesamiento de la colección de documentos con correctores ortográficos para la depuración del lenguaje.

References

1. Aggarwal, C.C., et al.: Neural networks and deep learning. Springer **10**, 978–3 (2018)
2. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (August 2019), https://crscardellino.github.io/SBWCE/
3. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)

[5] Experimentos disponibles en github.com/jumafernandez/clasificacion_correos/jcc

4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Fanny, F., Muliono, Y., Tanzil, F.: A comparison of text classification methods k-nn, naïve bayes, and support vector machine for news classification. Jurnal Informatika: Jurnal Pengembangan IT **3**(2), 157–160 (2018)
6. Harish, B.S., Guru, D.S., Manjunath, S.: Representation and classification of text documents: A brief review. IJCA, Special Issue on RTIPPR (2) pp. 110–119 (2010)
7. Islam, M.R., Chowdhury, M.U., Zhou, W.: An innovative spam filtering model based on support vector machine. In: CIMCA-IAWTIC'06. vol. 2, pp. 348–353. IEEE (2005)
8. Li, Z., Xiong, Z., Zhang, Y., Liu, C., Li, K.: Fast text categorization using concise semantic analysis. Pattern Recognition Letters **32**(3), 441–448 (2011)
9. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development **1**(4), 309–317 (1957)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11), 613–620 (1975)
12. Skiena, S.S.: The data science design manual. Springer (2017)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
14. Wu, L., Yen, I.E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.Y., Ravikumar, P., Witbrock, M.J.: Word mover's embedding: From word2vec to document embedding. arXiv preprint arXiv:1811.01713 (2018)

Agradecimientos

Los autores agradecen al Centro de Investigación, Docencia y Extensión en TIC de la Universidad Nacional de Luján (CIDETIC) la provisión de recursos computacionales para la ejecución de los experimentos de este proyecto.