Clasificación automática de correos electrónicos

Juan M. Fernandez[1], Nicolás Cavasín[1], Agustín Rodriguez[1], Marcelo Errecalde[2]
{jmfernandez, ncavasin, arodriguez}@unlu.edu.ar, merreca@unsl.edu.ar
[1]Departamento de Ciencias Básicas, Universidad Nacional de Luján
[2]LIDIC, Universidad Nacional de San Luis

Resumen

El correo electrónico es una de las herramientas de comunicación asincrónica más extendidas en la actualidad, habiendo desplazado a los canales más clásicos de comunicación debido a su alta eficiencia, costo extremadamente bajo y compatibilidad con muchos tipos diferentes de información [30].

En el último tiempo, con el objetivo de mejorar su uso y aprovechar a los correos electrónicos como fuente de conocimiento, se han aplicado diversas técnicas de aprendizaje automático a este tipo de información. En este sentido, existe un área particular del aprendizaje automático, denominada minería de textos, donde el conocimiento es generado a partir de la adopción de bases de datos exclusivamente textuales como fuente de datos [32]. A su vez, el correo electrónico posee características particulares respecto de otros elementos de texto que hace que existan diferencias y problemáticas particulares entre la minería de textos tradicional y la minería de correos electrónicos, conocida como *email mining* [5].

En este trabajo, se describen las acciones abordadas en el proyecto de investigación "Clasificación automática de correos electrónicos", así como las líneas de I+D comprendidas en el mismo.

Contexto

Este trabajo se encuentra en el marco del Proyecto de Investigación "Clasificación automática de correos electrónicos", aprobado por Disposición CDD-CB N°086/2020, del Departamento de Ciencias Básicas de la Universidad Nacional de Luján.

El objetivo general consiste en estudiar y analizar el conocimiento existente sobre técnicas de aprendizaje automático aplicadas a la clasificación automática de textos, particularmente de correos electrónicos, y generar modelos que aborden problemas concretos.

Asimismo, se está trabajando en conjunto con el Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC) de la Universidad Nacional de San Luis a efectos de integrar los resultados encontrados en este proyecto con desarrollos de ese Laboratorio en torno a la clasificación de preguntas para sistemas conversacionales (*chatbots*).

Introducción

Existen trabajos que han recogido estimaciones respecto de la utilización mundial del correo electrónico donde se afirma que actualmente existen más de 3930 millones de usuarios y se proyectan 4371 millones para el año 2023 [11], alcanzando el tráfico actual de 293.6 billones de correos enviados diariamente [28].

Muchos de estos correos electrónicos son enviados a centros de contacto de organizaciones públicas y privadas, debido a que este medio se ha constituido en un canal de comunicación estándar [27].

Sin embargo, este canal de comunicación requiere una importante afectación de recursos humanos. A efectos de graficar este costo, algunos autores han relevado este aspecto a través de estudios de casos; por ejemplo, se demostró que responder un correo electrónico de un ciudadano enviado a la Agencia de Pensiones de Suecia lleva unos 10 minutos y, por lo tanto, los 99000 mensajes que reciben por año pueden necesitar hasta 10 empleados de tiempo completo para responderlos [13].

Con el fin de mejorar su uso y aprovechar a los correos electrónicos como fuente de conocimiento se han aplicado diversas técnicas de minería de textos, donde el conocimiento es generado mediante la adopción de bases de datos exclusivamente textuales como fuentes de datos [32]. Estas técnicas se enfrentan a problemáticas muy complejas dentro del área de la ciencia de la computación, debido principalmente a la dificultad del análisis del lenguaje (derivada de su ambigüedad), fundamentalmente en la etapa de análisis semántico, como así también, a los relativamente escasos materiales de entrenamiento y a la capacidad de cómputo necesaria para correr determinados algoritmos muy demandantes en recursos de hardware [7].

A su vez, el correo electrónico como fuente de datos posee un conjunto de características particulares respecto de otros elementos de texto que hace que existan diferencias y problemáticas particulares entre la minería de textos tradicional y la minería de correos electrónicos, conocida como *email mining* [5]. Por un lado, los correos electrónicos poseen información adicional en el encabezado que pueden ser explotadas para la obtención de conocimiento. Asimismo,

poseen una extensión reducida que hace que muchas técnicas de minería de textos sean ineficientes para estas fuentes de datos. A su vez, este tipo de comunicaciones muchas veces se da en un contexto informal o inmersos en una cultura organizacional particular, por lo tanto, los errores ortográficos y gramaticales, así como las abreviaturas o acrónimos aparecen con frecuencia. Por otro lado, además de los datos textuales, los correos electrónicos pueden contener tipos más ricos de datos, como enlaces URL, marcas HTML e imágenes. Aprovechar al máximo esos datos no textuales existentes en los correos electrónicos representan un problema interesante y desafiante para abordar [30].

Antecedentes

En lo relativo a clasificación de correos electrónicos, existen abordajes desde el procesamiento y generación de resúmenes [31], utilización de redes neuronales [4], clasificación para respuesta automática de correos [26], aplicación de técnicas basadas en máquinas vector-soporte, Naive Bayes, clasificadores TF-IDF [30] y utilización de multi-view y semi-supervised learning [17, 35], entre otras.

Algunos autores que abordaron la clasificación de correos electrónicos para la respuesta automática [26] categorizan las técnicas de acuerdo a, básicamente, tres enfoques de recuperación de texto: categorización de texto por aprendizaje automático, cálculo de similitud estadística de texto y coincidencia de patrones de texto y plantillas.

En relación a la categorización de texto mediante técnicas de aprendizaje automático, existen trabajos [6, 24] que desarrollaron modelos utilizando las técnicas de KNN, Naïve Bayes, RIPPER y SVM, encontrando que SVM fue la técnica que demostró mejor performance. En el mismo sentido, existen trabajos [9] en los cuales se compara la precisión de técnicas como K-means++, KNN y Naïve Bayes, alcanzando niveles de precisión muy altos, por encima del 96 %, para K-means++. En otras experiencias se realizan comparaciones entre los métodos de clasificación de Naive Bayes, SMO, J48 y Random Forest [23], observando que el algoritmo Random Forest fue el que obtuvo la mejor precisión, siendo esta de un 95.5 % mientras que el algoritmo Naive Bayes fue el más veloz en la construcción del clasificador. Siguiendo la misma línea, en otra experiencia [30] se comparó la precisión de diferentes algoritmos de clasificación, árboles de decisión, redes neuronales, Naive Bayes, K-Nearest Neighbor y SVM. Se utilizaron datos académicos para predecir la performance de los alumnos encontrando que los árboles de decisión y redes neuronales fueron los que mejor performance obtuvieron.

Otros abordajes a partir de la clasificación de correos electrónicos mediante el cálculo de similitud estadística también obtuvieron resultados alentadores [3]. En estos casos, el modelo mantiene respuestas estándar asociadas a una variedad de preguntas etiquetadas como preguntas frecuentes. Cuando llega un correo electrónico de consulta, el sistema hace coincidir las oraciones en la consulta con las preguntas de la etiqueta considerando la distancia entre conceptos en las oraciones utilizando WordNet.

Un enfoque alternativo es el basado en coincidencia de patrones de texto y plantillas [2], donde el sistema mantiene un diccionario que contiene palabras y la probabilidad de que una palabra aparezca en un mensaje de una determinada categoría de texto, categorizando los mensajes en base a esa probabilidad junto con información adicional que toma de los mensajes de consulta.

También existen técnicas de clasificación de correos electrónicos utilizando un enfoque denominado de múltiples vistas o *multi-view* [35]; lo cual implica generar múltiples grupos de características de los correos y aprovechar los algoritmos de *Disagreement-based Semi-Supervised Learning* que proporcionan herramientas para ser entrenados en diferentes vistas. En algunas experiencias [17, 35], se generaron dos grupos de características de los correos, internas y externas, donde las primeras explotan el cuerpo del correo y las últimas aprovechan otras como el asunto y los destinatarios y luego se utilizó *Disagreement-based Semi-Supervised Learning* para generar varios modelos a múltiples vistas y permitirles colaborar para explotar ejemplos no etiquetados.

Por último y de forma más reciente, surgen los abordajes basados en *Deep Learning* [9] que implementan una red neuronal basada en un modelo *Long-Short-Term-Memory* para clasificar correos no deseados. Para resolver el problema de la gran cantidad de datos etiquetados necesarios para los métodos de *Deep Learning*, utilizaron un método de aprendizaje activo. Este método selecciona diferentes muestras y sólo entrena esas, buscando disminuir el costo del etiquetado manual de los datos. Este modelo demostró una mejor performance con respecto a los tradicionales CNN y RNN.

Líneas de I+D

Las principales líneas de investigación en las cuales se trabaja actualmente en el marco de este proyecto consisten en:

- Análisis y aplicación de diferentes estrategias de etiquetado de correos electrónicos.
- Implementación y comparación de estrategias de representación de documentos convencionales a correos electrónicos.
- Evaluación de técnicas de aprendizaje automático para la clasificación de correos electrónicos.

- Construcción de un clasificador automático como solución a la respuesta de consultas académicas.

A continuación se presenta una descripción breve de las lineas de I+D previamente enunciadas.

a. Estrategias de etiquetado

Una de las primeras tareas que se deben llevar a cabo para la construcción de un clasificador automático es la clasificación inicial de un conjunto de documentos que luego serán utilizados como conjuntos de entrenamiento y prueba para el entrenamiento y validación del clasificador. La estrategia tradicional para el etiquetado de documentos consiste en que esta tarea sea realizada por un humano, de forma manual. En muchas ocasiones, este etiquetado manual debe ser realizado por expertos en el tema que forma parte del problema que se desea abordar.

Si bien estas etiquetas de expertos proporcionan la piedra angular tradicional para evaluar los modelos de aprendizaje automático, el acceso limitado o costoso a los expertos representa un cuello de botella. A su vez, para caracterizar con precisión la efectividad de un sistema, la experiencia ha demostrado que los sistemas de IR (recuperación de información) deben evaluarse a la escala operativa en la que se utilizarán en la práctica, lo cual resulta una limitación para esta metodología puesto que debido a que los tamaños de las colecciones de datos han crecido rápidamente en los últimos años, se ha vuelto cada vez menos factible etiquetar manualmente tantos ejemplos usando el etiquetado experto tradicional [16].

En este sentido, han surgido metodologías alternativas que aportan mayor escalabilidad como la inferencia automática de etiquetas en función del comportamiento de los usuarios [14], la "supervisión distante", en la que los datos de entrenamiento son etiquetados a partir de algunas características del texto, como tags, emoticones y otros metadatos [10] o el etiquetado de palabras representativas [19].

En el marco de este proyecto, se indagan las estrategias existentes, así como se exploran nuevas soluciones, que resulten escalables y efectivas para la clasificación de correos electrónicos.

b. Representación de correos

Aunque un documento de texto expresa una gran variedad de información, lamentablemente carece de la estructura impuesta en una base de datos tradicional; por lo tanto, los datos no estructurados deben transformarse en datos estructurados previo a la aplicación de técnicas de aprendizaje automático. Después de convertir datos no estructurados en datos estructurados, necesitamos tener un modelo de representación de documentos efectivo para construir un sistema de clasificación eficiente [12]. En el marco de este proyecto, se evalúan y aplican diferentes estrategias de representación de documentos usualmente utilizadas como *bag of words* [18], *topic modeling* [34], *embeddings* [33] y BERT [29]. El objetivo de esta línea de investigación es evaluar las diferentes técnicas aplicadas a la clasificación de correos electrónicos.

c. Evaluación de algoritmos de clasificación

Luego de obtener los datos, realizar el preprocesamiento de los mismos para la extracción de características, realizar el etiquetado y avanzar en un esquema de representación, se entrena el clasificador utilizando distintos enfoques o algoritmos [22], como el aprendizaje bayesiano [21], la regresión logística, redes neuronales, árboles de decisión y máquinas de vectores soporte [15].

El modelo generado a partir del entrenamiento debe ser capaz de capturar las características distintivas de los documentos del conjunto de entrenamiento para luego poder analizar otros textos no observados previamente, lográndose así la capacidad de generalización del clasificador que se suele evaluar sobre otro conjunto de prueba separado [20].

A la fecha, y debido a la cantidad de algoritmos de aprendizaje existentes, resulta muy complejo sistematizar todos los abordajes posibles. De acuerdo a los antecedentes estudiados, los algoritmos escogidos que se están utilizando en el marco de esta investigación son el clasificador de Naive Bayes [21], la regresión logística [25], las máquinas de vector soporte [15], las redes neuronales recurrentes LSTM (*Long short-term memory*) [1] y XGBoost [8].

Las aspiraciones en torno a esta línea de estudio son encontrar las características de los problemas que hacen más acorde la utilización de un algoritmo de aprendizaje automático por encima del resto, sin descuidar las estrategias de representación de documentos ni las características del proceso de preprocesamiento realizado previamente.

d. Clasificador automático para la respuesta de consultas

La Universidad Nacional de Luján cuenta con un sistema informático propio para llevar adelante la gestión académica así como los trámites que de éstas se desprenden. Este sistema posee una funcionalidad que permite a los estudiantes realizar consultas vía correo electrónico al staff administrativo. Al cuerpo de ese correo, además del texto escrito por el estudiante, se agregan datos académicos y de la persona. Ante la llegada de un correo electrónico, el personal administrativo debe dar respuesta dentro de las 48 horas de realizada la solicitud. Como política de res-

guardo de la información, la Universidad Nacional de Luján realiza periódicamente una copia de seguridad con estas consultas y las respuestas brindadas a los estudiantes en cada caso, llegando a un total actual de 24700 correos electrónicos.

Utilizando esa base de conocimiento, se aborda el desafío de generar un modelo mediante técnicas de aprendizaje automático para clasificar cual es el tema de cada consulta realizada en función del contenido de los mensajes enviados y así poder responder las consultas de forma focalizada.

Objetivos

Los objetivos específicos que se persiguen en el marco de este proyecto se plantean a continuación:

1. Evaluar alternativas de pre-procesamiento de correos electrónicos y enriquecimiento de la representación de los mismos.

2. Evaluar nuevas técnicas de aprendizaje automático que permitan mejorar la precisión en problemas de clasificación de correos electrónicos.

3. Generar modelos que permitan clasificar correos electrónicos relacionados con casos concretos de acuerdo a los tópicos de los mismos.

4. Integrar y adaptar las técnicas y estrategias abordadas en este proyecto a sistemas de clasificación de preguntas para sistemas conversacionales (*chatbots*).

Complementariamente, su transferencia a la sociedad para resolver problemas del mundo real resulta de alto interés. Las herramientas a desarrollar pueden ser aplicadas a espacios tanto académicos como comerciales con lo que hay una oportunidad concreta de transferencia al sector público y privado.

Recursos Humanos

Se espera que este proyecto contribuya a consolidar un grupo de investigación en la temática y brindar un marco adecuado para la formación de recursos humanos en la Universidad Nacional de Luján a partir de la incorporación de saberes y competencias provenientes de la participación en actividades de investigación.

Concretamente, se ha incluido como integrantes del proyecto a dos docentes auxiliares del Departamento de Ciencias Básicas y un estudiante de la Carrera de Licenciatura en Sistemas de Información. A su vez, el director trabaja con estos temas en su tesis de Maestría, esperando culminarla en el marco del proyecto.

Por otro lado, se espera que este proyecto brinde la posibilidad a estudiantes de la Licenciatura en Sistemas de Información de la Universidad Nacional de Luján de realizar sus Tesinas de Grado en temas relacionados con la temática del proyecto.

Referencias

[1] Aggarwal, C. C., et al. Neural networks and deep learning. *Springer 10* (2018), 978–3.

[2] Al-Alwani, A. Improving email response in an email management system using natural language processing based probabilistic methods. *Journal of Computer Science 11*, 1 (2015), 109.

[3] Alfalahi, A., Eriksson, G., and Sneiders, E. Shadow answers as an intermediary in email answer retrieval. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (2015), Springer, pp. 209–214.

[4] Alghoul, A., Al Ajrami, S., Al Jarousha, G., Harb, G., and Abu-Naser, S. S. Email classification using artificial neural network. *ACM* (2018).

[5] Bogawar, P. S., and Bhoyar, K. K. Email mining: a review. *International Journal of Computer Science Issues(IJCSI) 9*, 1 (2012).

[6] Busemann, S., Schmeier, S., and Arens, R. G. Message classification in the call center. *arXiv preprint cs/0003060* (2000).

[7] Cardenas, M. E., Castillo, J. J., Navarro, M., Hernández, N., and Velazco, M. Herramientas para el desarrollo de sistemas de análisis de textos no estructurados. In *XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan)*. (2019).

[8] Chen, T., and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794.

[9] Chen, Z., Tao, R., Wu, X., Wei, Z., and Luo, X. Active learning for spam email classification. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence* (2019), pp. 457–461.

[10] Go, A., Bhayani, R., and Huang, L. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford 1*, 12 (2009), 2009.

[11] Group, T. R. Email statistics report, 2019-2023. url: http://www.radicati.com, 2019.

[12] Harish, B. S., Guru, D. S., and Manjunath, S. Representation and classification of text documents: A brief review. *IJCA, Special Issue on RTIPPR (2)* (2010), 110–119.

[13] Henkel, M., Perjons, E., and Sneiders, E. Examining the potential of language technologies in public organizations by means of a business and it architecture model. *International Journal of Information Management 37*, 1 (2017), 1507–1516.

[14] Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (2002), pp. 133–142.

[15] Joachims, T., et al. Transductive inference for text classification using support vector machines. In *Icml* (1999), vol. 99, pp. 200–209.

[16] Jung, H. J., and Lease, M. Evaluating classifiers without expert labels. *arXiv preprint arXiv:1212.0960* (2012).

[17] Li, W., Meng, W., Tan, Z., and Xiang, Y. Design of multi-view based email classification for iot systems via semi-supervised learning. *Journal of Network and Computer Applications 128* (2019), 56–63.

[18] Li, Z., Xiong, Z., Zhang, Y., Liu, C., and Li, K. Fast text categorization using concise semantic analysis. *Pattern Recognition Letters 32*, 3 (2011), 441–448.

[19] Liu, B., Li, X., Lee, W. S., and Yu, P. S. Text classification by labeling words. In *AAAI* (2004), vol. 4, pp. 425–430.

[20] Mariñelarena-Dondena, L., Errecalde, M. L., and Solano, A. C. Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología. *Revista Argentina de Ciencias del Comportamiento 9*, 2 (2017), 65–76.

[21] McCallum, A., Nigam, K., et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (1998), Citeseer, pp. 41–48.

[22] Russell Stuart, J., and Norvig, P. *Artificial intelligence: a modern approach*. Prentice Hall, 2009.

[23] Saha, S., DasGupta, S., and Das, S. K. Spam mail detection using data mining: A comparative analysis. In *Smart Intelligent Computing and Applications*. Springer, 2019, pp. 571–580.

[24] Scheffer, T. Email answering assistance by semi-supervised text classification. *Intelligent Data Analysis 8*, 5 (2004), 481–493.

[25] Skiena, S. S. *The data science design manual*. Springer, 2017.

[26] Sneiders, E. Review of the main approaches to automated email answering. In *New advances in information systems and technologies*. Springer, 2016, pp. 135–144.

[27] Sneiders, E., Sjöbergh, J., and Alfalahi, A. Automated email answering by text-pattern matching: Performance and error analysis. *Expert Systems 35*, 1 (2018), e12251.

[28] Statista. Most popular global mobile messenger apps as of july 2019, based on number of monthly active users (in millions). url: http://www.statista.com/, 2019.

[29] Sun, C., Qiu, X., Xu, Y., and Huang, X. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics* (2019), Springer, pp. 194–206.

[30] Tang, G., Pei, J., and Luk, W.-S. Email mining: tasks, common techniques, and tools. *Knowledge and Information Systems 41*, 1 (2014), 1–31.

[31] Upadhyay, M., Radhakrishnan, D., and Natarajan, M. Summarization and processing of email on a client computing device based on content contribution to an email thread using weighting techniques, Oct. 16 2018. US Patent 10,102,192.

[32] Usai, A., Pironti, M., Mital, M., and Mejri, C. A. Knowledge discovery out of text data: a systematic review via text mining. *Journal of knowledge management* (2018).

[33] Wu, L., Yen, I. E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.-Y., Ravikumar, P., and Witbrock, M. J. Word mover's embedding: From word2vec to document embedding. *arXiv preprint arXiv:1811.01713* (2018).

[34] YILDIRIM, S., and YILDIZ, T. A comparison of different approaches to document representation in turkish language. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi 22*, 2 (2018), 569–576.

[35] Zhao, J., Xie, X., Xu, X., and Sun, S. Multi-view learning overview: Recent progress and new challenges. *Information Fusion 38* (2017), 43–54.