Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 71-74 recibido 03-11-2017 revisado 07-01-2018 aceptado 04-02-2018

Nuevos Paradigmas de Análisis Basados en Contenidos para la Detección del Spam en RRSS

New approaches for content-based analysis towards Online Social Network spam detection

Enaitz Ezpeleta
Mondragon Unibertsitatea
Goiru kalea 2, 20500 Arrasate, Spain
eezpeleta@mondragon.edu

Resumen: Tesis doctoral realizada por Enaitz Ezpeleta Gallastegi en Mondragon Unibertsitatea, dentro del grupo de Sistemas Inteligentes para Sistemas Industriales, dirigida por los Doctores Urko Zurutuza Ortega (Mondragon Unibertsitatea) y José María Gómez Hidalgo (Pragsis Technologies). La defensa se efectuó el 30 de septiembre de 2016 en Arrasate. El tribunal estuvo conformado por el Dr. Manel Medina Llinas (Universitat Politecnica de Catalunya), el Dr. Magnus Almgren (Chalmers University of Technology), el Dr. Igor Santos Grueiro (Universidad de Deusto), el Dr. José Ramón Méndez Reboredo (Universidad de Vigo) y el Dr. D. Iñaki Garitano Garitano (Mondragon Unibertsitatea). La tesis obtuvo una calificación de Sobresaliente Cum Laude y la mención "Doctor Europeus".
Palabras clave: Spam, redes sociales, PLN, análisis de sentimiento, polaridad, reconocimiento de personalidad, seguridad

Abstract: PhD Thesis written by Enaitz Ezpeleta Gallastegi at Mondragon University supervised by Dr. Urko Zurutuza Ortega (Mondragon Unibertsitatea) and Dr. José María Gómez Hidalgo (Pragsis Technologies). The viva voce was held on the 30th September 2016 and the members of the commission were Dr. Manel Medina Llinas (Universitat Politecnica de Catalunya), el Dr. Magnus Almgren (Chalmers University of Technology), el Dr. Igor Santos Grueiro (Universidad de Deusto), el Dr. José Ramón Méndez Reboredo (Universidad de Vigo) y el Dr. D. Iñaki Garitano Garitano (Mondragon Unibertsitatea). The thesis obtained the grade of Excellent Cum Laude and the mention "Doctor Europeus".
Keywords: Spam, online social networks, NLP, sentiment analysis, polarity, personality recognition, security

1 Introducción

Las campañas de correo electrónico no deseado siguen siendo una de las mayores amenazas que afectan a millones de usuarios al día. Aunque las técnicas de detección de spam son capaces de detectar un porcentaje muy alto de spam, el problema está lejos de ser solventado, sobre todo por la cantidad tan alta de tráfico spam existente entre el tráfico global de correo electrónico, y las nuevas estrategias utilizadas por los atacantes.

Además, el auge del número de usuarios de las redes sociales (RRSS) en Internet (como Facebook, Twitter, Instagram...), muchos de los cuales publican mucha información de forma abierta en sus perfiles, han proporcionado que estos sitios se conviertan en objetivos atractivos para los atacantes, principalmente por dos razones: la posibilidad de explotar la información pública almacenada en los perfiles de los usuarios, y por la facilidad para entrar en contacto directo con los usuarios mediante los perfiles, los grupos, las páginas... Como consecuencia, cada vez se detectan más actividades ilegales en estas redes. Entre ellas, el spam es una de las que mayor impacto causa. Actualmente, la venta comercial, la creación de alarma social, campañas de sensibilización, distribución de *malware*, etc. son los principales objetivos de los men-

ISSN 1135-5948. DOI 10.26342/2018-60-8

sajes de spam. Tomando en cuenta esto, partimos de la hipótesis de que su forma de ser escrito conlleva una intencionalidad implícita, que el autor desea explotar para su detección.

Los principales objetivos de esta tesis son: (1) demostrar que es posible desarrollar spam personalizado usando información publicada en redes sociales que eluda los sistemas actuales de detección; y (2) diseñar y validar nuevos métodos para la detección y filtrado de spam usando técnicas de Procesamiento de Lenguaje Natural (PLN). Además, estos sistemas deberán ser efectivos con el spam que se propaga dentro de las redes sociales.

2 Organización de la Tesis

Este trabajo de tesis está organizado en los siguientes capítulos:

1. Indroducción: Se explica la motivación para la realización del trabajo, así como los objetivos y las hipótesis a los que se intenta dar respuesta.
2. Estado de la cuestión: En este capítulo se resume como se ha abordado la detección identificando los diferentes sistemas actuales. También se presentan diferentes propuestas basadas en técnicas de PLN, y se realiza una introducción a los problemas de seguridad de las redes sociales.
3. Efectividad del spam personalizado: Capítulo en el que se presenta el trabajo realizado de cara a demostrar que es posible crear spam personalizado capaz de saltarse los sistemas anti-spam actuales.
4. Análisis de sentimiento: Se resume como se puede conseguir mejorar el filtrado spam utilizando la polaridad de los mensajes.
5. Reconocimiento de personalidad: En este capítulo se describe el modelo creado utilizando las dimensiones de la personalidad del texto.
6. Combinación de ambas técnicas: Presentación del tercer modelo donde se combina la utilización de técnicas de análisis de sentimiento y reconocimiento de personalidad.
7. Conclusiones: Capítulo en el que se resumen las aportaciones más significativas del trabajo, así como las lineas futuras identificadas.

3 Contribuciones y Resultados Experimentales

Para validar el primer objetivo de este trabajo se ha diseñado y desarrollado un sistema que permite enviar campañas de spam personalizado (Ezpeleta, Zurutuza, y Hidalgo, 2015; Ezpeleta, Zurutuza, y Gómez Hidalgo, 2016c). Mediante este sistema, se ha podido demostrar que utilizando información pública personal de los usuarios de las redes sociales (Facebook) es posible crear spam personalizado que alcance ratios de click-through muy superiores a los del spam. Para ello el sistema recolecta direcciones de correo electrónico en Internet, para después extraer la información personal guardada de forma pública por el propietario de la cuenta vinculada a esa dirección en Facebook. Con esa información se crean diferentes perfiles que son usados para enviar correos electrónicos personalizados. Finalmente se han desarrollado experimentos donde se demuestra la eficacia de este tipo de spam frente al spam típico/común. Esta información sirve para subrayar el problema que supone publicar información personal en las redes sociales, así como para entender posibles riesgos futuros a los que la comunidad científica se deberá enfrentar, como es el caso del spam personalizado. Y por último ofrece las bases para el desarrollo de sistemas capaces de detectar este tipo de mensajes, tal y como se ha hecho en la segunda fase de esta tesis.

En la segunda parte de la tesis se presentan tres nuevos modelos para el filtrado de nuevos tipos de spam. Estos métodos tienen como objetivo detectar la intencionalidad comercial no evidente en los textos que luego ayuden a clasificarlos. Siendo este el objetivo, se identificó la necesidad de utilizar técnicas de PLN para analizar el contenido de los mensajes y poder extraer información que pudiera ser interesante a la hora de detectar mensajes no deseados. Debido al auge experimentado por estas técnicas en los últimos años, se ha podido realizar un estudio exhaustivo de gran variedad de técnicas para identificar las que mejor resultado ofrecían para este objetivo. De esta forma se han diseñado dos modelos independientes, donde uno de ellos utiliza Análisis de Sentimiento (AS) y el otro el Reconocimiento de Personalidad (RP) de los mensajes para mejorar la detección del spam.

El AS realizado, extrayendo la polaridad (mensaje positivo, negativo o neutro) de ca-

da mensaje, ofrece a la comunidad científica bases para demostrar que, teniendo los mensajes spam en su mayoría intención de vender productos, el contenido de los mensajes se escribe con una connotación más positiva que en los mensajes legítimos. Gracias a ello, al añadir esta información a los clasificadores de spam, se ha demostrado, tal y como se recoge en (Ezpeleta, Zurutuza, y Gómez Hidalgo, 2016a; Ezpeleta, Zurutuza, y Hidalgo, 2016), que los resultados obtenidos mejoran sustancialmente. Es decir, se ha demostrado que el AS ayuda a mejorar los resultados del filtrado de mensajes no deseados.

En el caso del segundo modelo presentado en (Ezpeleta, Zurutuza, y Gómez Hidalgo, 2016b; Ezpeleta, Zurutuza, y Gómez Hidalgo, 2016), se han mejorado los resultados del filtrado spam añadiendo información sobre la personalidad de cada mensaje, demostrando que las técnicas de RP también resultan de interés a la hora de mejorar los sistemas de detección de spam actuales.

Con la presentación de estos dos nuevos métodos, se ofrece tanto a la comunidad científica, así como a las empresas y organismos del sector, la posibilidad de ofrecer sistemas anti spam más eficaces a los usuarios, aportando seguridad y privacidad a las millones de personas que todos los días sufren las campañas de correo electrónico no deseados.

Finalmente, se ha presentado un nuevo modelo para la detección de spam donde se combinan los dos modelos anteriormente descritos, consiguiendo un sistema más eficaz tal y como se presenta en (Ezpeleta, Zurutuza, y Gómez Hidalgo, 2016d; Ezpeleta et al., 2017; Ezpeleta, Zurutuza, y Gómez Hidalgo, 2017). De esta forma, se demuestra que la combinación de técnicas de AS y RP mejora los resultados de las técnicas actuales de filtrado de spam.

Cabe destacar que las tres técnicas presentadas han sido validadas utilizando diferentes tipos de spam como son el spam en emails, spam en mensajes SMS y spam social o spam recogido en las redes sociales, y además han sido utilizados más de un conjunto de datos por cada tipo, con el objetivo de contrastar y refrendar la validez de los resultados obtenidos.

3.1 Resultados: eficacia del spam personalizado

Para validar el primero de los objetivos, se extrajeron direcciones de correo electrónico a través de un famoso buscador, y se contrastó la existencia de una vinculación a la red social Facebook de cada una de ellas, obteniendo una base de 22.654 usuarios con los cuales se pudieron crear perfiles para llevar a cabo el envío de diferentes campañas.

Los resultados demuestran que el spam personalizado es más eficaz que el spam habitual. Esto se refleja sobre todo en el porcentaje de usuarios que hacen click en la URL personalizada que se incluye en el contenido del correo enviado, siendo 18 veces más alto en el caso del spam personalizado, con un *click-through* del spam típico de un 0,41 % y un 7,62 % en el caso del personalizado.

3.2 Resultados: nuevos modelos para la detección de Spam

Una vez demostrado el riesgo que suponen las nuevas técnicas de creación de spam, se han diseñado y desarrollado tres nuevos modelos para la detección de nuevos tipos de spam. A la hora de realizar los experimentos para evaluar la eficacia de estos modelos, se han aplicado diferentes clasificadores tanto sobre los conjuntos de datos originales de cada tipo de spam, así como sobre los conjuntos de datos creados después de añadir los atributos creados con resultados de las diferentes técnicas utilizadas (AS, RP y combinación de ambas). Finalmente se ha llevado a cabo una comparativa en términos de precisión y el número de falsos positivos.

La Figura 1 muestra la precisión máxima obtenida en los distintos tipos de spam y utilizando los tres modelos presentados en este trabajo. El mejor resultado ha sido obtenido con el modelo que combina ambas técnicas (AS y RP). Cabe destacar que en el caso del número de falsos positivos, este se reduce significativamente en la mayoría de los casos.

4 Conclusiones

Al ser el spam un problema que afecta diariamente a millones de usuarios, la presentación de este tipo de modelos ayuda a que la experiencia de los usuarios vaya mejorando, y que dichos usuarios no sufran de posibles peligros derivados de este tipo de ataques contra su seguridad y privacidad.

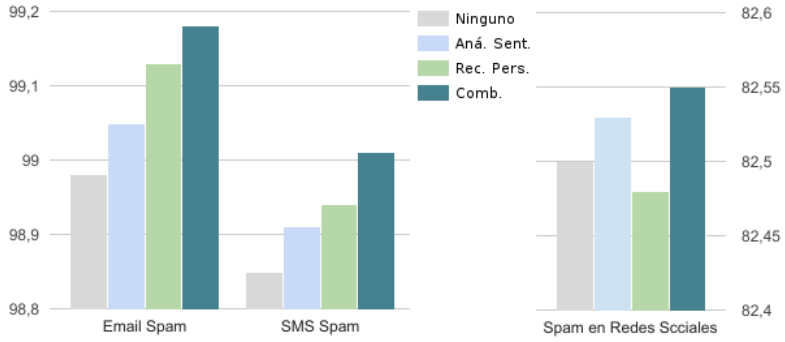

Figura 1: Comparativa de las precisiones obtenidas

En este trabajo se demuestra el potencial de las redes sociales a la hora de crear spam personalizado, el cual no es detectado por los sistemas de detección actuales. Tras presentar tres modelos novedosos en el ámbito de análisis de contenido para la detección del spam, se demuestra que se pueden mejorar los resultados de los sistemas actuales tanto en spam en emails, así como en mensajes SMS, y también en el spam que se propaga dentro de las redes sociales.

Muestra de la aplicabilidad de estos métodos en entornos reales es que actualmente, dentro del proyecto SocialSPAM (PI_2014_1_102), financiado por el Gobierno Vasco, se está desarrollando una aplicación nativa para Facebook. Esta herramienta analiza los mensajes de los usuarios, utilizando los métodos presentados en este trabajo, con el objetivo de detectar posibles mensajes spam, y filtrarlos.

Agradecimientos

Este trabajo ha sido realizado en el grupo de Sistemas Inteligentes para Sistemas Indutriales (Mondragon Unibertsitatea) en el proyecto SocialSPAM(PI_2014_1_102) ambos parcialmente financiados por el Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco.

Bibliografía

Ezpeleta, E., I. Garitano, I. Arenaza-Nuño, U. Zurutuza, y J. M. Gómez Hidalgo. 2017. Novel comment spam filtering method on youtube: Sentiment analysis and personality recognition. En *Proceedings of Current Trends In Web Engineering - ICWE 2017 International Workshops*.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2016a. Does sentiment analysis help in bayesian spam filtering? En *Springer Int. Publishing*, páginas 79–90.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2016b. Short messages spam filtering using personality recognition. En *Proceedings of the 4th Spanish Conference on Information Retrieval*, CERI '16, páginas 1–7, New York, NY, USA. ACM.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2016c. A study of the personalization of spam content using facebook public information. *Logic Journal of IGPL*.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2016d. Using personality recognition techniques to improve bayesian spam filtering. *Procesamiento del Lenguaje Natural*, 57:125–132.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2017. Short messages spam filtering combining personality recognition and sentiment analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. In press.

Ezpeleta, E., U. Zurutuza, y J. M. Gómez Hidalgo. 2016. Los spammers no piensan: usando reconocimiento de personalidad para el filtrado de spam en mensajes cortos. En *Actas de la XIV Reunión Española sobre Criptología y Seguridad de la Información*.

Ezpeleta, E., U. Zurutuza, y J. M. G. Hidalgo. 2015. An analysis of the effectiveness of personalized spam using online social network public information. En *Springer Int. Publishing*, páginas 497–506.

Ezpeleta, E., U. Zurutuza, y J. M. G. Hidalgo. 2016. Short messages spam filtering using sentiment analysis. En *Springer Int. Publishing*, páginas 142–153.