Volumen 10
XXVI Verano de la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Identificación de Usuarios entre Redes Sociales

José Alfredo Romero González[1], Hugo Iván Lozoyo Belman[1], Juan Carlos Alonso Sánchez[1], Aldo Isaac Hernández Antonio[1], Luis Miguel López Santamaría[1], Juan Carlos Gómez Carranza[1]

[1]Departamento de Ingeniería Electrónica, División de Ingenierías Campus Irapuato-Salamanca,
Universidad de Guanajuato
{ja.romerogonzalez, hi.lozoyobelman, jc.alonsosanchez, ai.hernandezantonio, lm.lopezsantamaria, jc.gomez}@ugto.mx

Resumen

La identificación de usuarios entre redes sociales es una tarea que ha tomado relevancia en los últimos años. Esto se debe a que las redes sociales han tenido un aumento significativo y constante en el número de usuarios. El poder identificar las cuentas de estos usuarios a través de distintas redes sociales podría permitir diferentes aplicaciones como recomendación de productos, marketing, identificación tanto de cuentas falsas como de usurpación de identidad, ciencias forenses, entre otros. En este artículo, se presenta un estudio sobre el análisis de las publicaciones y tweets generados por usuarios verificados en Facebook y Twitter, respectivamente, con la finalidad de identificar a un usuario de una red social en otra. Para el trabajo, se utilizaron dos conjuntos de datos correspondientes a las redes sociales antes mencionadas. Estos conjuntos de datos cuentan con un total de 453 usuarios verificados. El conjunto de datos de Facebook cuenta con un total de 787,649 publicaciones, mientras que el conjunto de datos de Twitter cuenta con 1,303,885 tweets. Lo conjuntos de datos están en dos idiomas, inglés y español. Utilizando el contenido textual de las publicaciones y tweets generados por los usuarios, se extrajeron las palabras y con ellas se construyeron y probaron una serie de modelos de aprendizaje de máquina para resolver la tarea. Para evaluar el desempeño de los modelos se utilizaron las métricas de *accuracy* y t*op-n accuracy*.

Palabras clave: Redes sociales; minería de datos; identificación de usuarios; aprendizaje de máquina; contenido generado por usuarios.

Introducción

En la actualidad, las redes sociales se han convertido en uno de los medios de comunicación más populares entre personas y organizaciones. A través de estas redes, los usuarios pueden expresar sus preferencias o ideas por medio de fotos, videos, publicaciones, etc. Twitter y Facebook son dos de las redes sociales más populares entre los usuarios, según cifras del sitio web *Our World in Data*, en 2019 Facebook ocupó el primer lugar entre las redes sociales más populares a nivel mundial con 2.38 billones de usuarios[1], procesando 2,500 millones de publicaciones al día[2], mientras que Twitter ocupó la segunda posición con 330 millones de usuarios[3] procesando 500 millones de tweets al día[4].

Por lo general, las personas suelen estar registradas en más de una red social simultáneamente [1], pero la información personal que registran en las redes sociales suele estar incompleta o no se puede tener acceso a esta debido a las políticas de privacidad de cada red social. El vínculo de identidad de usuario (*User Identity Linkage* o UIL) es una de las problemáticas que se plantea a la hora de identificar las diferentes cuentas que pertenezcan a una misma persona física en distintas redes sociales [2]. Una de las tareas que ha planteado una solución a estas problemáticas es la predicción de identidad en redes sociales, la cual busca realizar un mapeo entre las diferentes cuentas de usuario en distintas redes sociales para encontrar las identidades reales de las personas propietarias de las cuentas [3].

[1] https://bit.ly/2K1M5dt
[2] https://bit.ly/3elUbtQ
[3] https://bit.ly/2K1M5dt
[4] https://bit.ly/3km1cPe

La identificación de usuarios a través de redes sociales ofrece algunos beneficios como por ejemplo dar recomendaciones más acertadas al usuario como parte de una estrategia de marketing basado en el contenido que publica. De igual manera, el seguimiento de usuarios entre redes permite construir mecanismos de confianza, ya que es común que las redes sociales no cuenten con algún sistema de autenticación de identidad, lo cual permite la creación de perfiles falsos, la introducción de información engañosa o la falta de información [4]. Adicionalmente, puede ayudar a confirmar las identidades de los usuarios que se podrían aplicar en diferentes áreas como ciencias forenses, seguridad y marketing.

En el presente artículo se realiza un estudio sobre la identificación de usuarios entre las redes sociales más populares, Facebook y Twitter. Para dicho estudio, se utilizó la información pública de cuentas que hayan sido verificadas tanto por Facebook con la *insignia de verificación*[5], como por Twitter con *la insignia azul de verificación*[6]. La tarea consiste en identificar a un usuario de una red social (Facebook o Twitter) en otra red social (Twitter o Facebook); utilizando para ello las palabras extraídas del contenido textual generado por los usuarios.

Para llevar a cabo el estudio, se tomaron de manera aleatoria usuarios que tuvieran sus cuentas verificadas en Facebook y en Twitter, obteniendo un total de 453 usuarios verificados. Lo usuarios fueron separados por dos idiomas: inglés y español, obteniendo 258 usuarios correspondientes a inglés y 195 a español. El conjunto de datos correspondiente a las publicaciones de los usuarios en Facebook consiste en 417,532 publicaciones en inglés y 370,117 publicaciones en español. El conjunto de datos correspondiente a Twitter consiste en 715,143 tweets en inglés y 588,712 tweets en español. De los cuatro conjuntos de datos se extrajeron las palabras como característica textual para su análisis.

Empleando las palabras se realizó la construcción de modelos de aprendizaje de máquina para realizar la identificación de los usuarios entre redes sociales. Los modelos empleados para el entrenamiento y prueba fueron k vecinos más cercanos (*K-Nearest Neighbors* o KNN), regresión logística (*Logistic Regression* o LR), multinomial simple de Bayes (*Multinomial Naïve Bayes* o MNB), bosques aleatorios (*Random Forest* o RF), descenso de gradiente estocástico (*Stochastic Gradient Descent* o SGD) y máquinas de vectores de soporte lineales (*Linear Support Vector Machines* o LSVM).

Para el estudio, se construyeron diferentes modelos variando los hiperparámetros de cada modelo de aprendizaje. Para medir el desempeño de cada modelo con un hiperparámetro en particular, se consideraron las métricas de *accuracy* y *top-n accuracy*, que son populares en la literatura para la evaluación de modelos de aprendizaje de máquina.

La aportación de nuestro trabajo se enfoca en el estudio del desempeño de las palabras y modelos de aprendizaje de máquina para la tarea de identificación de usuarios en redes sociales, intentando responder la siguiente pregunta de investigación, ¿Qué modelo de aprendizaje tiene el mejor desempeño para la tarea?

El resto del presente artículo se organiza de la siguiente manera, la sección 2 muestra los trabajos relacionados al presente trabajo, donde se explican diferentes análisis para la identificación de usuarios. En la sección 3, se explica la metodología que se utilizó para el estudio, entre ellos la descripción del conjunto de datos y la experimentación a detalle. En la sección 4, se muestran los resultados que se obtuvieron de los diferentes experimentos realizados. Por último, en la sección 5 se muestran las conclusiones del estudio y se proponen algunas ideas para trabajos futuros.

[5] https://bit.ly/3BgwSeS
[6] https://bit.ly/2VLVETR

Trabajos Relacionados

La identificación o el seguimiento de usuarios a través de redes sociales es una tarea que se ha vuelto popular en los últimos años. Como consecuencia, se han realizado algunos trabajos de investigación utilizando diferentes enfoques.

Las personas usan distintas redes sociales para diferentes propósitos, lo cual nos hace pensar que en cualquier red social las podemos encontrar a partir de su nombre de usuario. Los autores en [5] utilizaron la similitud que existe entre los nombres de usuario para vincular diferentes perfiles en servicios web y poder identificar al usuario. En [6] los autores se centran en comprobar si en los diferentes sitios web donde existen usuarios con nombres de usuario idénticos, estos pertenecen a una misma persona. En [3] los investigadores propusieron una modelo que utiliza el nombre de usuario como la mínima información que se puede conocer sobre el/ella en una red social, con la finalidad de identificar sus características, por ejemplo, sus intereses individuales o sus opiniones políticas.

En [7] se trabajó bajo una metodología que se enfoca en poder identificar a un usuario o tener candidatos predichos con base en una serie de atributos biográficos. En [8] los autores propusieron un enfoque que explotaba un conjunto de propiedades para la identificación de un usuario, como el nombre de usuario, el nombre completo, su ubicación, su correo electrónico, etc. En [9] se realizó un trabajo más enfocado en la detección de perfiles falsos o incompletos, en donde los autores presentaron un enfoque basado en campos aleatorios condicionales.

En [10] se menciona un método basado en estructuras sociales para mejorar el desempeño del mapeo de usuarios. Los investigadores propusieron un algoritmo de aprendizaje subespacial nombrado MAH (*Manifold Alignment on Hypergraph*), el cual utiliza probabilidades para encontrar al usuario en la otra red social.

A pesar de la diversidad de investigaciones que se han hecho en el campo de la identificación de usuarios a través de las redes sociales, el problema sigue abierto a la investigación para crear sistemas más robustos que ayuden a obtener mejores resultados en el proceso del mapeo. Lo anterior con la finalidad de tener una mayor certeza en las predicciones y poder construir sistemas con mejor funcionalidad y utilidad.

Metodología

La metodología de este trabajo está formada por tres fases que son la adquisición de datos, el procesamiento de datos y la experimentación. Las tres fases se describen a continuación.

Adquisición de datos

Como primera fase de la metodología, se realizó la búsqueda de manera aleatoria de 500 usuarios que fueran figuras públicas y cuyas cuentas tuvieran la insignia de verificación en Facebook y Twitter. Esta insignia indica que la red social confirma que el perfil del usuario es la presencia auténtica del usuario. Se seleccionaron figuras públicas con la finalidad de asegurar que el contenido que generan sea de acceso totalmente abierto. Analizando a los usuarios y a algunas de sus publicaciones y tweets en las redes sociales, se detectó que había usuarios que manejaban distintos idiomas, por lo tanto, se decidió trabajar únicamente con aquellos usuarios que sus textos estuvieran en español o en inglés, quedando un total de 453 usuarios verificados, de los cuales 258 escriben en inglés, y 195 en español

Para la obtención del contenido generado por los usuarios en sus cuentas de Twitter, se realizó una aplicación para realizar un *scrapping* obteniendo distintas características como la fecha de publicación, el idioma, el tweet, la ubicación, etc. Dejando únicamente el contenido textual del tweet para conformar nuestro conjunto de datos de Twitter. De igual manera, para la obtención de las publicaciones de los usuarios en sus cuentas de Facebook, se utilizó la librería de Python llamada *facebook scraper*[7], la cual ayudo a obtener distintas características como la publicación, la fecha, el número de comentarios, las reacciones, etc. Dejando únicamente el contenido textual de la publicación para conformar nuestro conjunto de datos de Facebook.

[7] Disponible en: https://pypi.org/project/facebook-scraper/

El conjunto de datos recolectado de Facebook contiene 787,649 publicaciones, con 417,532 publicaciones en inglés y 370,117 publicaciones en español. El conjunto de datos recolectado de Twitter contiene 1,303,885 tweets, con 715,143 tweets en inglés y 588,712 tweets en español. Las estadísticas de los subconjuntos se pueden ver en la tabla 1.

Tabla 1. Distribución de las publicaciones por idioma y red social.

Idioma	Número de usuarios	Facebook	Twitter	Total
Ingles	258	417,532	715,143	1,132,675
Español	195	370,117	588,712	958,829
Total	453	787,649	1,303,885	

Procesamiento de datos

En esta fase se realizó el procesamiento de las publicaciones y los tweets, de los cuales se extrajeron como característica textual a las palabras. El conjunto de datos consta de 4 archivos (uno por cada red social y por cada idioma), en cada uno, una línea representa una publicación o un tweet ya sea en inglés o en español. Posteriormente, se realizó un proceso de limpieza donde se removieron palabras que eran muy cortas (longitud < 2), muy largas (longitud > 35) y palabras vacías (*stopwords*). Se hizo uso de listas de palabras vacías en inglés y español utilizando la librería NLTK en Python. Al final del proceso de limpieza, se obtuvieron otros 4 archivos, un archivo por red social y por idioma, donde cada línea de cada archivo contiene solo las palabras filtradas de cada publicación o tweet de cada usuario.

Haciendo uso de los conjuntos obtenidos después del proceso de limpieza, se aplicó la técnica conocida como td-idf (*term frequency inverse document frequency*), que está definida por la ecuación 1.

$$tfidf(t,d) = tf(t,d)\ x\ idf(t) \tag{1}$$

donde $td(t,d)$ determina la frecuencia relativa de un término t específico en un documento d y el término $idf(t)$ que está definido por la Ecuación 2.

$$idf(t) = log\frac{1+n_d}{1+df(t)} + 1 \tag{2}$$

donde df es el número de documentos en los que el término t, y el termino n_d corresponde al número total de documentos. Se computo el término idf para cada conjunto de entrenamiento que sería utilizado posteriormente para la vectorización del conjunto de prueba.

Experimentación

Una vez finalizado el proceso de vectorización, se utilizaron seis modelos de aprendizaje de máquina para la experimentación, los cuales fueron k vecinos más cercanos (*K-Nearest Neighbors* o KNN), regresión logística (*Logistic Regression* o LR), multinomial simple de Bayes (*Multinomial Naïve Bayes* o MNB), bosques aleatorios (*Random Forest* o RF), descenso de gradiente estocástico (*Stochastic Gradient Descent* o SGD) y máquinas de vectores de soporte lineales (*Linear Support Vector Machines* o LSVM). Exceptuando el modelo MNB, para el resto de los modelos se utilizaron diferentes valores para sus hiperparámetros, como se muestra en la tabla 2, construyendo un modelo con cada valor.

Tabla 2. Valores de hiperparámetros considerados para cada modelo.

Modelo	Hiperparámetro	Descripción	Valores
KNN	K	Número de vecinos	[1, 2, 3, 5, 10]

Volumen 10
XXVI Verano de la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

LR	C	Parámetro de regularización	[0.01, 0.1, 1, 10, 100]
SGD	C	Parámetro de regularización	[0.01, 0.1, 1, 10, 100]
SVM	C	Parámetro de regularización	[0.01, 0.1, 1, 10, 100]
RF	R	Número de árboles	[10, 50, 100, 200, 500]

En la experimentación primero se crearon modelos de aprendizaje entrenados con las publicaciones o tweets de una red social y posteriormente los modelos fueron probados con las publicaciones o tweets de la otra red. Se espera que un modelo sea capaz de clasificar de manera correcta las publicaciones como correspondientes al usuario que las generó (identificación del usuario). De esta forma, se crearon modelos tanto en inglés como en español entrenando con datos de Facebook y probando con datos de Twitter, y entrenando con datos de Twitter y probando con datos de Facebook.

Para medir el desempeño de los modelos, se utilizó la matriz de confusión que está formada por las celdas de: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). La relación que existe entre las clases reales de los usuarios contra las clases predichas por los modelos se puede visualizar en esta matriz. Como métrica de desempeño, primero se calculó el *accuracy*, la cual se define como $accuracy = \frac{TP+TN}{TP+FP+FN+TN}$, que representa la fracción de predicciones que el modelo realizó de manera correcta. Adicionalmente, cada modelo produce una probabilidad de que las publicaciones de prueba sean generadas por cada usuario, de esta forma se midió el *top-n accuracy*, que representa si el usuario correcto está dentro de los n usuarios más probables.

Los códigos para el procesamiento y experimentación fueron codificados en Python utilizando librerías como NLTK, scikit-learn y NumPy.

Resultados

Las tablas 3 a 6 muestran los resultados obtenidos durante la experimentación. Las tablas 3 y 4 corresponden al idioma inglés, las tablas 5 y 6 al idioma español. Las tablas 3 y 5 corresponden al entrenamiento con Facebook y la prueba con Twitter. Las tablas 4 y 6 corresponden al entrenamiento con Twitter y la prueba con Facebook.

En las tablas, los renglones 2 a 7 indican los modelos de aprendizaje probados: LSVM, LR, MNB, KNN, SGCD y RDF con el hiperparámetro que muestra mejores resultados. Las columnas 2 a 5 indican las métricas utilizadas para determinar el mejor desempeño de cada modelo: accuracy, top-2 accuracy, top-5 accuracy, top-10 accuracy. La columna 6 indica el tiempo de ejecución de cada uno de los modelos. El modelo LSVM no presenta valores para las métricas top-2 accuracy, top-5 accuracy y top-10 accuracy ya que no calcula las probabilidades por usuario.

En la tabla 3, se observa que el mayor valor para accuracy se obtiene con el modelo RDF, con un valor de 0.81. Para el top-2 accuracy, tanto el modelo RDF y como el LR obtuvieron el mismo valor. Sin embargo, para top-5 y top-10 accuracy, el modelo LR obtiene los mejores resultados entre todos los modelos. A pesar de los buenos resultados obtenidos por LR, también es el modelo que más tiempo toma en realizar la ejecución.

Tabla 3. Resultados entrenando con Facebook y probando con Twitter para el idioma inglés.

Modelo	Accuracy	Top-2 accuracy	Top-5 accuracy	Top-10 accuracy	Tiempo (s)
SVM, C=0.01	0.69	--	--	--	43.38
LR, C=100	0.79	**0.86**	**0.92**	**0.94**	83.73
MNB	0.67	0.75	0.82	0.86	28.19

Volumen 10
XXVI Verano de la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

KNN, N=10	0.09	0.16	0.51	0.87	26.13
SGCD, C=0.01	0.01	0.01	0.03	0.68	35.45
RDF, C=500	**0.81**	**0.86**	0.89	0.92	70.54

En la tabla 4, se observan los resultados de entrenar con datos de Twitter y probar con datos de Facebook en inglés. Los valores de desempeño más alto se obtienen con los modelos SVM y MNB, arrojando un valor de 0.93 para el accuracy. Para la métrica top-2 accuracy, los modelos LR, MNB y KNN obtuvieron valores similares de 0.95. Para el resto de las métricas, el modelo LR obtuvo los mejores resultados con 0.98 y 0.99 para top-5 accuracy y top-10 accuracy respectivamente. En este caso, el modelo con mayor tiempo de ejecución es RDF.

Tabla 4. Resultados entrenando con Twitter y probando con Facebook para el idioma inglés.

Modelo	Accuracy	Top-2 accuracy	Top-5 accuracy	Top-10 accuracy	Tiempo (s)
LSVM, C=0.01	0.93	--	--	--	43.97
LR, C=0.1	**0.91**	**0.95**	**0.98**	**0.99**	63.64
MNB	**0.93**	**0.95**	0.96	0.97	28.06
KNN, N=2	0.45	**0.95**	0.95	0.96	26.97
SGCD, C=10	0.69	0.76	0.84	0.89	44.00
RDF, C=500	0.64	0.70	0.78	0.84	74.28

Analizando los resultados de la tabla 5, donde se realizó el entrenamiento con Facebook y la prueba con Twitter para español, se puede ver que los valores con el mejor desempeño se obtienen el con modelo RDF, dando un valor de 0.75 para el accuracy, 0.82 para el top-2 accuracy, 0.90 para el top-5 accuracy y 0.92 para el top-10 accuracy. Este modelo es también el que tiene un mayor tiempo de ejecución.

Tabla 5. Resultados entrenando con Facebook y probando con Twitter para el idioma español.

Modelo	Accuracy	Top-2 accuracy	Top-5 accuracy	Top-10 accuracy	Tiempo (s)
LSVM, C=0.01	0.68	--	--	--	28.59
LR, C=0.01	0.72	0.76	0.83	0.89	31.06
MNB	0.69	0.74	0.79	0.86	16.79
KNN, N=10	0.11	0.21	0.43	0.85	15.11
SGCD, C=0.01	0.01	0.16	0.67	0.78	21.40
RDF, C=500	**0.75**	**0.82**	**0.90**	**0.92**	45.88

Volumen 10
XXVI Verano de la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

En la tabla 6, que muestra los resultados de entrenar con Twitter y probar con Facebook en español, el modelo que tuvo un mejor desempeño fue el LR obteniendo un valor de 0.87 en accuracy, 0.92 para top-2 accuracy, 0.96 para top-5 accuracy y 0.97 para top-10 accuracy. Solo en top-2 accuracy, el modelo KNN obtiene valores similares a LR. El modelo LR es también el que presenta un mayor tiempo de ejecución.

Tabla 6. Resultados entrenando con Twitter y probando con Facebook para el idioma español.

Modelo	Accuracy	Top-2 accuracy	Top-5 accuracy	Top-10 accuracy	Tiempo (s)
LSVM, C=0.01	0.85	--	--	--	31.29
LR, C=10	**0.87**	**0.92**	**0.96**	**0.97**	56.25
MNB	0.86	0.91	0.95	**0.97**	17.55
KNN, N=2	0.43	**0.92**	0.92	0.92	16.49
SGCD, C=0.01	0.01	0.04	0.30	0.86	23.41
RDF, C=500	0.41	0.59	0.78	0.88	50.10

En la tabla 7 se observan los valores promedio de todos los experimentos previos para cada modelo y cada métrica, con lo que se puede tener una visión del comportamiento general de los modelos. No se muestran los valores de los hiperparámetros porque éstos varían con cada experimento y el promedio no tiene un sentido válido. Es posible observar que el modelo con los mejores valores para todas las métricas al promediar sobre todos los experimentos es LR. Este modelo tiene el segundo mayor tiempo promedio de ejecución. El modelo LR es un modelo discriminativo que separa los datos a partir de una función, es claro que los usuarios de redes sociales pueden ser separados utilizando este enfoque. No obstante, aún existe un margen de mejora.

Tabla 7. Promedio de los resultados en todos los experimentos.

Modelo	Accuracy	Top-2 accuracy	Top-5 accuracy	Top-10 accuracy	Tiempo (s)
LSVM	0.79	--	--	--	36.81
LR	**0.82**	**0.87**	**0.92**	**0.95**	58.67
MNB	0.79	0.84	0.88	0.92	22.65
KNN	0.27	0.56	0.70	0.90	21.18
SGCD	0.18	0.24	0.46	0.80	31.07
RDF	0.65	0.74	0.84	0.89	60.20

Conclusiones

En este artículo se presentó un estudio de la identificación de usuarios entre redes sociales, utilizando el contenido textual de las publicaciones y tweets en conjunto con distintos modelos de aprendizaje de máquina. Los modelos fueron entrenados con datos de una red social y probados con datos de la otra red social para identificar qué usuario era el que había generado un conjunto de publicaciones. Se trabajo con un conjunto de datos dividido en dos idiomas, español e inglés, para las redes sociales de Facebook y Twitter. En total se tenían 453 usuarios, de los cuales 258 eran

en inglés, y 195 en español. De estos usuarios se colectaron 417,532 publicaciones en inglés y 370,117 publicaciones en español para Facebook; y 715,143 tweets en inglés y 588,712 tweets en español para Twitter.

Con base en la experimentación, podemos concluir que para identificar a un usuario de una red social en otra que:

- El modelo de aprendizaje que sigue un enfoque discriminativo, LR, fue el que tuvo un mejor desempeño en general en los diferentes experimentos para las métricas de accuracy, top-2 accuracy, top-5 accuracy y top-10 accuracy; aunque en algunos experimentos no obtuvo el valor más alto. También es el segundo modelo que en general toma más tiempo en realizar el proceso.
- El modelo MNB se podría considerar como un buen modelo para este trabajo, ya que el valor de la exactitud obtenido para cada experimento se aproximó en varios experimentos al del modelo LR, pero el tiempo de ejecución del modelo MNB es hasta una tercera parte del tiempo de ejecución de LR.
- Los experimentos donde se utilizó el conjunto de datos de Twitter como entrenamiento y el conjunto de datos de Facebook como prueba obtuvieron resultados más altos, en ambos idiomas, que en los experimentos contrarios (entrenar con Facebook y probar con Twitter). Una causa posible es que el conjunto de datos de Twitter es más grande, por lo que se tiene más información para construir un modelo más robusto. También es posible que las haya palabras que los usuarios usan en Facebook que no aparecen en Twitter, lo que dificultaría la identificación de los usuarios.

Los resultados de este artículo proporcionan ideas para futuras investigaciones como la extracción de otras características del contenido generado por los usuarios como emoticones/emojis, etiquetas (hashtags), enlaces (links), menciones (@ o ats) y abreviaturas para analizar si se obtiene mejores resultados en la identificación de los usuarios. Adicionalmente, sería interesante investigar el uso de vectores de palabras como fastText o GloVe, que miden la estadística de coocurrencia entre palabras a partir de un conjunto de datos de entrenamiento. Finalmente, se podría explorar el uso de redes neuronales profundas durante la clasificación.

Referencias

Bartunov, S., Korshunov, A., Park, S. T., Ryu, W., & Lee, H. (2012, August). Joint link-attribute user identity resolution in online social networks. In Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. ACM.

Carmagnola, F., & Cena, F. (2009). User identification for cross-system personalisation. Information Sciences, 179(1-2), 16-32.

Fu, S., Wang, G., Xia, S., & Liu, L. (2020). Deep multi-granularity graph embedding for user identity linkage across social networks. Knowledge-Based Systems, 193, 105301.

Li, Y., Zhang, Z., Peng, Y., Yin, H., & Xu, Q. (2018). Matching user accounts based on user generated content across social networks. Future Generation Computer Systems, 83, 104-115.

Liu, J., Zhang, F., Song, X., Song, Y. I., Lin, C. Y., & Hon, H. W. (2013, February). What's in a name? An unsupervised approach to link users across communities. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 495-504).

Motoyama, M., & Varghese, G. (2009, November). I seek you: searching and matching individuals in social networks. In Proceedings of the eleventh international workshop on Web information and data management (pp. 67-75).

Nie, Y., Jia, Y., Li, S., Zhu, X., Li, A., & Zhou, B. (2016). Identifying users across social networks based on dynamic core interests. Neurocomputing, 210, 107-115.

Perito, D., Castelluccia, C., Kaafar, M. A., & Manils, P. (2011, July). How unique and traceable are usernames?. In International Symposium on Privacy Enhancing Technologies Symposium (pp. 1-17). Springer, Berlin, Heidelberg.

Tan, S., Guan, Z., Cai, D., Qin, X., Bu, J., & Chen, C. (2014, June). Mapping users across networks by manifold alignment on hypergraph. In twenty-eighth AAAI conference on artificial intelligence.

Zafarani, R., & Liu, H. (2013, August). Connecting users across social media sites: a behavioral-modeling approach. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 41-49).