VOLUMEN 16
XXVII Verano De la Ciencia
ISSN 2395-9797
www. jóvenesenlaciencia.ugto.mx

Identificación del Perfil Demográfico de Influencers en Twitter utilizando *Deep Learning*

Juan Carlos Alonso Sánchez [1], Aldo Isaac Hernández Antonio [1], Diana Martínez Frías [1], Raúl Uriel Silva Ramírez[1], Bruno Adrián Franco Ortiz [1], Juan Aguilera Huerta [1], Juan Carlos Gómez Carranza [1*]
[1]Departamento de Ingeniería Electrónica, División de Ingenierías Campus Irapuato-Salamanca, Universidad de Guanajuato.
jc.alonsosanchez
ai.hernandezantonio
d.martinezfrias
j.aguilerahuerta
jc.gomez@ugto.mx
* Autor de correspondencia

Resumen

El perfilado de autor en redes sociales consiste en predecir de forma automática los atributos demográficos de una población de usuarios a partir de la información que estos comparten y generan en las redes sociales. El perfilado de autor segmenta a los usuarios de acuerdo con sus atributos demográficos, lo que permite a empresas y organizaciones ajustar el contenido que proveen a los usuarios con fines de mercadotecnia, promoción política, programas sociales, información educativa, entretenimiento, entre otros. En este artículo se presenta un análisis del desempeño de distintos modelos de aprendizaje profundo (Deep learning) para realizar la predicción del género, ocupación y año de nacimiento de influencers en Twitter. La predicción se hace de manera indirecta, analizando no el contenido de los influencers, si no los mensajes de texto publicados por los seguidores de estos. Para evaluar el desempeño de los modelos se utilizó el conjunto de datos publicado en la conferencia PAN@CLEF 2020 que contiene en su parte de entrenamiento más de 166 millones de tweets de los seguidores de 1920 influencers; y en su parte de prueba cerca de 35 millones de tweets de los seguidores de 400 influencers. A estos datos se les extrajeron como características textuales las palabras y con éstas se realizaron los experimentos, evaluando el desempeño de los modelos utilizando la métrica macro F1. Los resultados indican que los modelos de aprendizaje profundo tienen dificultades para alcanzar una predicción precisa del perfil demográfico, especialmente para el rasgo del año del nacimiento.

Palabras clave: Perfilado de autor, minería de datos, aprendizaje de máquina, deep learning, redes sociales.

Introducción

El perfilado de autor se define como el análisis del contenido generado o compartido por un usuario con el objetivo de determinar de forma automática atributos demográficos que caractericen a ese usuario, tales como su edad, género, ocupación [0], rasgos de personalidad [0], nivel educativo, orientación política [0], entre otros. Esta tarea ha sido relevante en los últimos años ya que actualmente millones de personas generan diariamente un gran volumen de contenido en los medios electrónicos, como el internet y particularmente en las redes sociales. En estas redes, los usuarios suelen expresar sus ideas, gustos u opiniones utilizando mensajes de texto, imágenes, audios o videos.

El perfilado de autor en redes sociales tiene diversas aplicaciones, ya que permite segmentar a los usuarios por grupos dependiendo de sus atributos demográficos. A partir de esta segmentación, empresas y organizaciones pueden ajustar el contenido y las herramientas que proporcionan a los usuarios con fines de mercadotecnia, promoción política, programas sociales, información educativa, entretenimiento, entre otros. Por ejemplo, en la mercadotecnia, el perfilado de autor puede apoyar a las empresas a realizar campañas de productos o servicios focalizadas a usuarios con características específicas. De igual forma, con propósitos de seguridad, usando el perfilado de autor se puede lograr una identificación primaria de usuarios que

presenten un comportamiento anómalo (acoso, hostigamiento, intento de robo de información, terrorismo, etc.) dentro de las redes sociales y cuya información demográfica está ofuscada.

En el presente artículo se realiza un estudio sobre el perfilado demográfico de *influencers* en la red social Twitter. Un *influencer* se considera un usuario de la red que tiene un número considerable de seguidores dentro de la misma. La tarea se realiza de manera indirecta, analizando no el contenido de los influencers, sino los mensajes de texto publicados por los seguidores de tales *influencers*, y con base en ellos predecir los atributos demográficos de género, ocupación y año de nacimiento de los *influencers*.

Para conducir el estudio, se utilizaron los conjuntos de datos de entrenamiento y de prueba publicados en el evento PAN@CLEF 2020[1] , los cuales están formados por más de 166 millones de tweets de los seguidores de 1,920 celebridades para el entrenamiento; y por cerca de 35 millones de tweets de los seguidores de 400 celebridades para la prueba. Para este trabajo, de los conjuntos de datos excluimos los tweets que utilizaran un alfabeto no occidental y que no estuvieran en inglés, así como los retweets y las respuestas de estos [1]. En los conjuntos de datos, las celebridades están clasificadas en dos géneros (masculino, femenino), cuatro ocupaciones (político, creador, artista, deportista) y en 60 años de nacimiento (entre 1940 y 1999).

A los tweets de ambos conjuntos se les extrajeron las palabras, consideradas como una secuencia de caracteres alfabéticos más los caracteres '-' y '_', ignorando otros elementos como los links, menciones (ats), etiquetas (hashtags), emoticones/emojis y etiquetas HTML, ya que se ha estudiado que las palabras representan la mayor parte del contenido de las publicaciones [0].

Con esas características, se entrenaron y probaron dos clasificadores basados en modelos de aprendizaje profundo (*Deep Learning*), particularmente modelos de redes neuronales recurrentes (RNN), llamados Long Short-Term Memory (LSTM) y Gated Recurrent Unit (GRU) para realizar la predicción de los atributos demográficos. Se realizó la experimentación variando los valores de diferentes parámetros de los modelos que afectan su desempeño.

El desempeño de los modelos se midió utilizando la métrica macro F1, que es una métrica popular en clasificación de textos, principalmente cuando se tienen clases desbalanceadas (donde algunas clases tienen mayor cantidad de ejemplos que otras), lo cual es el caso para las clases encontradas en el presente problema.

La contribución de nuestro trabajo radica en el estudio del desempeño de dos modelos de aprendizaje de profundo probando con diferentes parámetros de los modelos para la tarea de perfilado demográfico indirecto de influencers en la red social Twitter, intentando responder las siguientes preguntas de investigación: 1) ¿Hay una arquitectura de aprendizaje profundo con mejor desempeño? 2) ¿Es recomendable usar aprendizaje profundo para esta tarea? 3) ¿Qué parámetro de los modelos afecta más su desempeño?

El resto del artículo está organizado como sigue. En la Sección 2 se presenta una breve descripción de algunos trabajos relacionados con la tarea de perfilado de autor en redes sociales. En la Sección 3 se presenta la metodología seguida en el presente artículo para la solución de la tarea. La Sección 4 muestra los resultados obtenidos de la experimentación con los modelos de aprendizaje profundo. Finalmente, en la Sección 5 se presentan las conclusiones del trabajo y algunos posibles caminos para futuras investigaciones.

Trabajos Relacionados

El estudio de perfilado de autor en redes sociales a partir del análisis del contenido textual que generan los usuarios se ha abordado a lo largo de los años siguiendo diferentes enfoques. Dentro de los atributos demográficos que se han estudiado para esta tarea se incluyen la edad, el género, la ocupación [0], el nivel

[1] Disponible en: https://pan.webis.de/clef20/pan20-web/celebrity-profiling.html

socioeconómico, los rasgos de personalidad [0], entre otros; siendo la predicción de edad y género los atributos más populares para determinar [0].

Uno de los principales eventos donde se han presentado investigaciones sobre el estudio de perfilado de autor en redes sociales es en las conferencias de PAN[2]. PAN forma parte de CLEF (Conference and Labs of Evaluation Forum), en donde desde el 2013 se realiza anualmente la tarea de perfilado de autor para la predicción de edad, género, idioma nativo, ocupación rasgos de personalidad, entre otros [0, 0, 0, 0, 0, 0]. En estas conferencias se han utilizado diversos conjuntos de datos extraídos de Twitter, los cuales contienen el texto de las publicaciones generadas por los usuarios. Los conjuntos de datos se han conformado principalmente por publicaciones en inglés, aunque también se han agregado otros idiomas como español, portugués, italiano, neerlandés y árabe.

A través de las ediciones de PAN@CLEF se han presentado una diversidad de trabajos que han hecho uso de diferentes enfoques para la tarea de perfilado de autor. Se han utilizado diferentes características textuales como palabras, emoticonos/emojis [0], n-gramas [0], diccionario de palabras, entre otras. De igual manera, se han utilizado diferentes modelos de aprendizaje de máquina como máquinas de vectores de soporte [0], regresión logística [0], clasificadores bayesianos y modelos de aprendizaje profundo (*deep learning*) [0].

Recientemente, en las conferencias de PAN@CLEF se ha presentado el estudio de perfilado de celebridades. Considerando a una celebridad como un usuario de una red social que tiene un número considerable de seguidores. El objetivo es la predicción de variables demográficas como el género, edad, ocupación y grado de fama utilizando el contenido generado en Twitter [0] ya sea por la celebridad o por sus seguidores [0].

Para el perfilado de celebridades utilizando el contenido generado por las mismas, en [0] utilizaron máquinas de vectores de soporte y regresión logística para la predicción de ocupación, edad y género. Los autores en [0] utilizaron un modelo de regresión logística para predecir la edad, género y grado de fama, mientras que para predecir la ocupación utilizaron un modelo multimodal simple de Bayes. De igual manera, utilizaron un número promedio de palabras por tweet, emojis, longitud de palabras, hashtags, hipervínculos, menciones, entre otra. En [0], los autores emplearon vectores tf-idf (*term-document frequency inverse document frequency*) formados a partir de unigramas de palabras, así como también trigramas de caracteres delimitados por palabras. Los autores usaron clasificadores como máquinas de vectores de soporte con kernels lineales y RBF, regresión logística, bosques aleatorios, y clasificadores de potenciación de gradiente.

En cuanto al perfilado de celebridades utilizando el contenido generado por sus seguidores, los autores en [0] usaron una matriz de tf-idf que se introdujo en una red neuronal LSTM para la predicción. Los autores en [0] utilizaron características como el promedio de todos los vectores de palabras de los tweets de los seguidores, palabras vacías (stopwords), hashtags, emojis, menciones y links; los cuales fueron usados con modelos de regresión logística, máquinas de vectores de soporte y bosques aleatorios para la predicción. Por otro lado, en [0], los autores utilizaron representaciones léxicas en conjunto con clasificadores de regresión logística para la predicción de la edad y ocupación, mientras que para la predicción del género usan un modelo de máquinas de vectores de soporte.

Metodología

La metodología de este trabajo se encuentra conformada por tres fases, la extracción de datos, el procesamiento de los datos y la experimentación. Las tres fases se encuentran descritas a continuación.

Descripción del conjunto de datos

En este artículo se utilizaron los conjuntos de datos de entrenamiento y de prueba de la conferencia PAN@CLEF 2020 para la tarea de *celebrity profiling*[3]. Las celebridades para este conjunto de datos fueron muestreadas del Webis Celebrity Profiling Corpus 2019 [0]. Ese corpus contiene los IDs de 71,706 celebridades e información demográfica sobre ellos obtenida de WikiData [0]. Se seleccionaron las

[2] https://pan.webis.de/

[3] Disponible en: https://pan.webis.de/clef20/pan20-web/celebrity-profiling.html

celebridades de las cuales se contaba simultáneamente con los atributos de género, ocupación y año de nacimiento. Posteriormente, en Twitter se hizo una selección de aquellas celebridades que tuvieran al menos 10 seguidores, con al menos 100 tweets en inglés por seguidor, obteniendo un total de 10,585 celebridades. Después, se hizo una segunda selección, buscando la muestra más grande posible de perfiles balanceados por ocupación y género, dando como resultado 2,320 celebridades. Se seleccionaron de forma aleatoria 10 seguidores por celebridad. Este conjunto se separó aproximadamente 80:20 dando como resultado un conjunto de entrenamiento con las publicaciones de los seguidores de 1920 celebridades y uno de prueba con las publicaciones de los seguidores de 400 celebridades. De ambos conjuntos de datos se eliminaron aquellas publicaciones que fueran retweets o respuestas a otros tweets.

A pesar de que el conjunto de datos ya viene seleccionado con tweets en inglés, en el presente trabajo se encontró que aún existían varios tweets que no estaban en algún idioma occidental, por lo que se decidió eliminarlos.

Las celebridades de ambos conjuntos de datos se encuentran etiquetadas con tres atributos demográficos: género (hombre y mujer), año de nacimiento (entre 1940 y 1999) y ocupación (político, creador, artista y deportista).

En las Tablas 1 y 2 se observa la distribución de usuarios por género y ocupación de los conjuntos de datos de entrenamiento y prueba respectivamente. Como se puede ver en la Tabla 2, la distribución de usuarios del conjunto de prueba es homogénea para el género sobre las ocupaciones. Por otro lado, en el conjunto de entramiento se observa que para la clase 'político' hay un número ligeramente mayor de hombres que de mujeres, lo que crea un pequeño desbalance en el conjunto total (56% de hombre y 44% de mujeres).

Tabla 1. *Distribución de usuarios por género y ocupación (conjunto de datos de entrenamiento).*

Género	Político	Creador	Artista	Deportista	Total
Mujer	128	240	240	240	848
Hombre	352	240	240	240	1072
Total	480	480	480	480	1920

Tabla 2. *Distribución de usuarios por género y ocupación (conjunto de datos de prueba).*

Género	Político	Creador	Artista	Deportista	Total
Mujer	50	50	50	50	200
Hombre	50	50	50	50	200
Total	100	100	100	100	400

Por motivos de ilustración, se agruparon los años de nacimiento en décadas, y su distribución con respecto al género se muestran en las Tablas 3 y 4 para los conjuntos de entrenamiento y prueba respectivamente. Tanto para el conjunto de entrenamiento como para el conjunto de prueba se observa un predominio de usuarios nacidos en los años 1980s, seguidos de los nacidos en los años 1970s. En la tarea de predicción, se considera el año exacto de nacimiento.

Tabla 3. *Distribución de usuarios por género y década de nacimiento (conjunto de datos de entrenamiento).*

Género	1940s	1950s	1960s	1970s	1980s	1990s	Total
Mujer	20	64	119	217	285	143	848
Hombre	68	150	237	264	257	96	1072
Total	88	214	356	481	542	239	1920

Tabla 4. *Distribución de usuarios por género y década de nacimiento (conjunto de datos de prueba).*

Género	1940s	1950s	1960s	1970s	1980s	1990s	Total
Mujer	12	36	35	41	51	25	200
Hombre	10	22	26	39	72	31	200
Total	22	58	61	80	123	56	400

Procesamiento de Datos

De cada tweet en ambos conjuntos (entrenamiento y prueba) se extrajeron las palabras como características para representarlo. Una palabra se definió como una secuencia de caracteres alfabéticos, más los caracteres '-' y '_' con el fin de capturar palabras compuestas. Para realizar esta extracción se utilizó una expresión regular. Posteriormente, se eliminaron las palabras muy cortas (longitud < 3), muy largas (longitud > 35) y las palabras vacías (*stopwords*). Para este último filtro, se utilizó una lista de palabras vacías en inglés proporcionada por la librería NLTK en Python. Posteriormente, se concatenaron todas las listas de palabras de los tweets correspondientes a cada celebridad, quedando un único documento largo para representar a cada celebridad.

Utilizando el conjunto de entrenamiento se extrajo el vocabulario base, que consiste en el conjunto de palabras única dentro de este conjunto. Con el vocabulario, tanto el conjunto de entrenamiento como el de prueba se convirtieron a secuencias de enteros y posteriormente se rellenaron (padding) para formar matrices densas de enteros. Este proceso se realizó con el tokenizador y la función de relleno de Keras en Python. El resultado de este proceso se almacenó en archivos para tener un procesamiento más eficiente en las fases posteriores.

Experimentación

Para el proceso de experimentación usamos dos modelos de aprendizaje profundo basados en las redes neuronales recurrentes o RNN (Recurrent Neural Networks). Este tipo de redes son útiles para analizar datos de series temporales permitiendo también tratar la dimensión de “tiempo”.

Para este trabajo se utilizaron los modelos GRU (Gated Recurrent Unit) y LSTM (Long Short Term Memory). Las neuronas de ambos modelos de red guardan un estado y además pueden recordar el pasado. El modelo LSTM aprende por medio de secuencias de término que son largas , mientras que el modelo GRU hace la incorporación de dos mecanismos de compuertas (la compuerta de actualización y la compuerta de reseteo).

Para evaluar el desempeño de los modelos, se utilizó la métrica F1 que está basada a su vez en las métricas *precision* y *recall.* La *precision* mide la proporción de usuarios clasificados correctamente, es decir, se centra en lo que el modelo dice y luego lo compara con la realidad; mientras que *recall* mide cuántos usuarios positivos son correctamente clasificados. La métrica F1 está definido por las siguientes ecuaciones:

VOLUMEN 16
XXVII Verano De la Ciencia
ISSN 2395-9797
www. jóvenesenlaciencia.ugto.mx

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = 2\,\frac{Precision * Recall}{Precision + Recall} \quad (3)$$

En las ecuaciones 1 y 2, TP es el número de usuarios positivos verdaderos (usuarios positivos clasificados como positivos), TN el número de usuarios negativos verdaderos (usuarios negativos clasificados como negativos), FP el número de usuarios falsos positivos (usuarios negativos clasificados como positivos) y FN el número de usuarios falsos negativos (usuarios positivos clasificados como negativos).

La métrica F1 corresponde a la media armónica de las métricas precisión y recall, lo cual lo hace particularmente útil en problemas en los que el conjunto de datos a analizar está desbalanceado, alcanzando su máximo en 1 que simboliza un modelo perfecto y su mínimo en 0 que significa que el modelo es totalmente incorrecto.

Al entrenar la red neuronal con parámetros definidos, no es posible recrear los mismos resultados al volver a entrenar el modelo nuevamente con los mismos parámetros, debido a se entrenaron y evaluaron los modelos 10 veces. Comparamos el desempeño de los modelos probando diferentes valores en dos parámetros de la red neuronal, como se muestra en la Tabla 5:

***Tabla 5**. Valores considerados para los parámetros de la red neuronal.*

Parámetro	Descripción	Valores
emb_dim	Tamaño de dimensión de la capa de embedding	[50, 100, 200, 300]
max_len	Longitud máxima de palabras que entra al modelo	[7500, 8750, 10000, 11250, 12500]

Para cada modelo LSTM y GRU se entrenaron modelos con la combinación de los parámetros indicados en la Tabla 5, para un total de 40 modelos (20 para LSTM y 20 para GRU). Cada modelo a su vez se ejecutó 10 veces, obteniendo así 400 resultados. Adicionalmente, se entrenaron modelos de forma independiente para cada rasgo demográfico de año de nacimiento, género y ocupación, lo que da un total de 1200 experimentos. Para cada rasgo y modelo LSTM y GRU con la combinación de parámetros se obtuvo la mediana de la métrica F1.

Todos los códigos para el procesamiento y experimentación se realizaron en Python utilizando las bibliotecas scikit-learn, pandas, numpy, Keras y TensorFlow

Resultados

En la Tabla 6 se muestran los resultados para la métrica F1 de la predicción de los tres atributos demográficos, año de nacimiento, género y ocupación, usando las arquitecturas GRU y LSTM evaluando las arquitecturas con sus combinaciones de los parámetros para el amaño de dimensión de embedding y el número de palabras que toma como entrada la red neuronal. Los resultados en cada celda indican la mediana los 10 experimentos realizados con cada combinación. En negritas se muestran los valores más altos obtenidos para cada rasgo.

Para cada rasgo de forma individual, la mejor arquitectura para el año de nacimiento es GRU utilizando una dimensión de embedding de 200 y una longitud de 7500 palabras. En el caso del género la mejor arquitectura es GRU con una dimensión de embedding de 100 y una longitud de 12500 palabras. Finalmente, en el caso de la ocupación la mejor arquitectura es LSTM con una dimensión de embedding de 300 y una longitud de 10000 palabras.

Para ambas arquitecturas. LSTM y GRU, los resultados más bajos se dieron al predecir la edad. En este caso, al ser un problema con 60 clases, la predicción es más compleja en comparación con el género (dos clases) y la ocupación (cuatro clases). En general es conocido que entre menos clases haya, la predicción se facilita; lo cual se puede observar en los resultados del género, que con dos clases tiene los resultados más altos.

***Tabla 6**. Resultados (macro F1) para los tres atributos demográficos.*

Parámetros		LSTM			GRU		
emb_dim	max_len	Año/nac	Género	Ocupación	Año/nac	Género	Ocupación
50	7,500	0.0019	0.3574	0.3093	0.0024	0.3892	0.3049
50	8,750	0.0011	0.4402	0.2975	0.0018	0.3709	0.2907
50	10,000	0.0016	0.4133	0.3365	0.0019	0.4627	0.3443
50	11,250	0.0013	0.4249	0.3144	0.0025	0.3755	0.2825
50	12,500	0.0013	0.4129	0.3102	0.0023	0.4293	0.3098
100	7,500	0.0018	0.3398	0.3122	0.0026	0.3607	0.3239
100	8,750	0.0017	0.4290	0.3069	0.0016	0.3639	0.3246
100	10,000	0.0018	0.4698	0.3553	0.0019	0.3993	0.3469
100	11,250	0.0017	0.4464	0.3400	0.0021	0.4281	0.3338
100	12,500	0.0015	0.3827	0.3334	0.0014	**0.4759**	0.3083
200	7,500	0.0028	0.3474	0.3536	**0.0032**	0.3827	0.3332
200	8,750	0.0017	0.4147	0.3492	0.0024	0.3707	0.3317
200	10,000	0.0025	0.4093	0.3505	0.0015	0.4521	0.3455
200	11,250	0.0017	0.4324	0.3367	0.0024	0.3974	0.3335
200	12,500	0.0019	0.4182	0.3262	0.002	0.3838	0.3223
300	7,500	0.0018	0.3333	0.3715	0.0027	0.4346	0.3391
300	8,750	0.003	0.3755	0.3545	0.0017	0.3606	0.3462
300	10,000	0.0008	0.4306	**0.3822**	0.0021	0.4256	0.3555
300	11,250	0.0025	0.3683	0.3615	0.0019	0.4243	0.3336
300	12,500	0.0023	0.381	0.3309	0.0015	0.4165	0.324

En la Tabla 7 se muestran los resultados agregados (mediana) para el tamaño de la dimensión de embedding. Para la arquitectura LSTM se observa una tendencia a tener un mejor desempeño conforme se aumenta el valor en este parámetro para todos los rasgos, alcanzando un mejor desempeño con tamaño de 300. Para el caso de la arquitectura GRU solamente con el rasgo de ocupación se observa un efecto de mejora, para los otros rasgos parece que este parámetro no es relevante (año de nacimiento) o disminuye el desempeño (género).

***Tabla 7**. Resultados (macro F1) agregados por tamaño de dimensión de embedding.*

	LSTM			GRU		
emb_dim	**Año/nac**	**Género**	**Ocupación**	**Año/nac**	**Género**	**Ocupación**
50	0.0013	0.4133	0.3102	0.0023	0.3892	0.3049
100	0.0013	0.4133	0.3122	0.0023	0.3755	0.3098
200	0.0016	0.4133	0.3122	0.0023	0.3755	0.3239
300	0.0017	0.4249	0.3122	0.0023	0.3755	0.3239

En la Tabla 8 se muestran los resultados agregados (mediana) por la longitud de palabras que toma como entrada la red neuronal. El efecto de este parámetro en los resultados depende del rasgo que se quiera predecir, para el año de nacimiento un valor bajo de 7500 palabras produce buenos resultados en ambas arquitecturas. Mientras que, para el género y la ocupación, un valor intermedio de 10000 palabras es mejor. Sin embargo, no hay una tendencia que indique que a un mayor número de palabras el desempeño mejorará.

***Tabla 8**. Resultados (macro F1) agregados por longitud de palabras.*

	LSTM			GRU		
max_len	**Año/nac**	**Género**	**Ocupación**	**Año/nac**	**Género**	**Ocupación**
7,500	0.0019	0.3436	0.3329	0.0027	0.3860	0.3286
8,750	0.0017	0.4219	0.3281	0.0018	0.3673	0.3282
10,000	0.0017	0.4220	0.3529	0.0019	0.4389	0.3462
11,250	0.0017	0.4287	0.3384	0.0023	0.4109	0.3333
12,500	0.0017	0.3978	0.3286	0.0018	0.4229	0.3161

Finalmente, en la Tabla 9 se muestran los resultados agregados (mediana) por arquitectura. Aquí se puede observar que de manera generar, independientemente de los valores de los parámetros, LSTM se desempeña mejor para los rasgos de género y ocupación, mientras que GRU lo hace para el año de nacimiento. No obstante, las diferencias son pequeñas.

***Tabla9**. Resultados (macro F1) agregados por arquitectura.*

LSTM			GRU		
Año/nac	**Género**	**Ocupación**	**Año/nac**	**Género**	**Ocupación**
0.0018	0.4131	0.3366	0.0021	0.3984	0.3325

Conclusiones

De acuerdo con los experimentos realizados en este trabajo para la tarea del perfilado demográfico de influencers en Twitter, ambas arquitecturas de redes neuronales probadas, LSTM y GRU, produjeron resultados semejantes, con diferencia de centésimas. LSTM probó ser ligeramente mejor al predecir el género y la ocupación de los influencers mientras que GRU obtuvo mejores resultados prediciendo su año de nacimiento. Respondiendo a la primera pregunta de la investigación, no hay una arquitectura que empíricamente sea mejor que la otra dada la diferencia minúscula de sus resultados.

Para el parámetro de dimensión de embedding, los resultados indican que para esta tarea entre mayor sea el tamaño de dimensión, los resultados llegan a ser más altos. Sin embargo, la diferencia parece no ser significativa. Por lo cual concluimos que este parámetro no afecta fuertemente el desempeño de ninguna de las dos arquitecturas.

Respondiendo a la tercera pregunta de investigación, el parámetro que sí afecta de manera significativa a los resultados es el número máximo de palabras por celebridad que toma como entrada la red neuronal recurrente. Por cada influencer se tienen entre 10 mil y 4.4 millones de palabras correspondientes a los tweets de sus seguidores, al truncar el número de palabras que se alimenta a la red neuronal, 10,000 palabras resultó ser la cantidad indicada para obtener los mejores resultados. Si se incrementa o decrementa este número de palabras, los resultados comienzan a disminuir.

Respondiendo a la segunda pregunta de investigación, no es recomendable el uso de aprendizaje profundo para esta tarea. En comparación con los resultados obtenidos al implementar modelos de aprendizaje de máquina [0], el aprendizaje profundo apenas alcanza la mitad de eficiencia de los mejores resultados obtenidos con los modelos de aprendizaje discriminativos (como Máquinas de Vectores de Soporte Lineales y Regresión Logística).

Algunas ideas para trabajos futuros incluyen el uso de otras arquitecturas de redes neuronales de aprendizaje profundo, como los transformers, en particular la arquitectura BERT, además de explorar otras características más complejas como los n-gramas o las características estilísticas.

Referencias

Wiegmann, M., Stein, B., Potthast, M.: Overview of the celebrity profiling task at pan 2020. In: CLEF (2020)

Moreno, D. R. J., Gomez, J. C., Almanza-Ojeda, D. L., & Ibarra-Manzano, M. A. (2019). Prediction of personality traits in twitter users with latent features. In *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)* (pp. 176-181). IEEE.

Cohen, R., & Ruths, D. (2013, June). Classifying political orientation on Twitter: It's not easy! In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 7, No. 1).

Garcia-Guzman, R., Andrade-Ambriz, Y. A., Ibarra-Manzano, M. A., Ledesma, S., Gomez, J. C., & Almanza-Ojeda, D. L. (2020). Trend-based categories recommendations and age-gender prediction for pinterest and twitter users. *Applied Sciences*, *10*(17), 5957.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., & Inches, G. (2013). Overview of the author profiling task at PAN 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation* (pp. 352-365). CELCT.

Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., & Daelemans, W. (2014). Overview of the 2nd author profiling task at pan 2014. In *CEUR Workshop Proceedings* (Vol. 1180, pp. 898-927). CEUR Workshop Proceedings.

Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015, September). Overview of the 3rd Author Profiling Task at PAN 2015. In *CLEF* (p. 2015). sn.

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B. (2016). Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. *Working Notes Papers of the CLEF*, *2016*, 750-784.

Rangel, F., Rosso, P., Potthast, M., & Stein, B. (2017). Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, 1613-0073.

Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., & Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working Notes Papers of the CLEF*, 1-38.

López-Santamaría, L. M., Gomez, J. C., Almanza-Ojeda, D. L., & Ibarra-Manzano, M. A. (2019). Age and gender identification in unbalanced social media. In *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)* (pp. 74-80). IEEE.

Wiegmann, M., Stein, B., & Potthast, M. (2019, September). Overview of the Celebrity Profiling Task at PAN 2019. In *CLEF (Working Notes)*.

Radivchev, V., Nikolov, A., & Lambova, A. (2019, September). Celebrity Profiling using TF-IDF, Logistic Regression, and SVM. In *CLEF (Working Notes)*.

Moreno-Sandoval, L. G., Puertas, E., Plaza-del-Arco, F. M., Pomares-Quimbaya, A., Alvarado-Valencia, J. A., & Alfonso, L. (2019). Celebrity Profiling on Twitter using Sociolinguistic.

Martinc, M., Skrlj, B., & Pollak, S. (2019, September). Who is Hot and Who is Not? Profiling Celebs on Twitter. In *CLEF (Working Notes)*.

Alroobaea, R., Almulihi, A. H., Alharithi, F. S., Mechti, S., Krichen, M., & Belguith, L. H. (2020). A Deep Learning Model to Predict Gender, Age and Occupation of the Celebrities based on Tweets Followers. In *CLEF (Working Notes)*.

Hodge, A., & Price, S. (2020). Celebrity Profiling using Twitter Follower Feeds.

Koloski, B., Pollak, S., & Skrlj, B. (2020). Know your Neighbors: Efficient Author Profiling via Follower Tweets. In *CLEF (Working Notes)*.

Gomez, J. C. (2019). Analysis of the effect of data properties in automated patent classification. *Scientometrics*, *121*(3), 1239-1268.

Bevendorff, J., Potthast, M., Hagen, M., Stein, B.: Heuristic Authorship Obfuscation. In: Korhonen, A., Màrquez, L., Traum, D. (eds.) 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), pp. 1098–1108, Association for Computational Linguistics (Jul 2019).

Gomez, J. C., Alonso-Sánchez, J. C., López, L. M., Hernández, A. I., Lozoyo, H. I., Romero, J. A. (2021). Perfilado demográfico de celebridades y seguimiento de usuarios en redes sociales. Jóvenes en la Ciencia (10)