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

Perfilado Demográfico de Celebridades de Redes Sociales

Juan Carlos Alonso Sánchez [1], Aldo Isaac Hernández Antonio [1], José Alfredo Romero González [1], Hugo Iván Lozoyo Belman [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.
{jc.alonsosanchez, ai.hernandezantonio, ja.romerogonzalez, hi.lozoyobelman, lm.lopezsantamaria, jc.gomez}@ugto.mx

Resumen

El perfilado de autor en redes sociales es una tarea que trata de predecir de forma automática los atributos demográficos de una población objetivo de usuarios a partir de la información que estos comparten y generan en las redes sociales. El perfilado de autor permite segmentar a los usuarios dependiendo de sus atributos demográficos. Con esta segmentación, distintas empresas y organizaciones pueden 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 el proyecto de investigación que analiza los mensajes de texto publicados por los seguidores de celebridades (usuarios populares) en Twitter, con el fin de predecir el perfil demográfico de tales celebridades, conformado por su género, ocupación y año de nacimiento. Para esta tarea se utilizan dos conjuntos de datos: el de entrenamiento y el de prueba. El conjunto de datos de entrenamiento contiene 5,066,608 tweets pertenecientes a 1,920 celebridades de Twitter. El conjunto de datos de prueba está conformado por 34,893,195 tweets generados por los seguidores de 400 celebridades (con al menos 10 seguidores). A partir de estos datos se realizaron experimentos extrayendo una serie de características textuales de los tweets y con ellas se construyeron diversos modelos de aprendizaje de máquina. Para evaluar los modelos se midió el área bajo la curva ROC. Los resultados indican que algunos atributos como el año de nacimiento son complicados de predecir. Se observa de igual forma, que características como los vectores de palabras presentan buen desempeño sobre todo en combinación con modelos de aprendizaje discriminativos.

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

1. Introducción

El perfilado de autor se entiende como el análisis del contenido generado o compartido por un usuario con la finalidad de predecir de forma automática atributos demográficos que caractericen a ese usuario, tales como su edad, género, ocupación [1], rasgos de personalidad [2], nivel educativo, orientación política [3], entre otros. Esta tarea ha tomado mucha relevancia en los últimos años ya que en la actualidad millones de las personas tienen acceso a los medios electrónicos, generan perfiles en redes sociales y generan contenido diariamente. En este contenido, los usuarios suelen expresar sus gustos, opiniones e ideas a partir de imágenes, videos y texto.

El perfilado de autor en redes sociales tiene distintas aplicaciones, ya que permite sectorizar a los usuarios por grupos dependiendo de sus atributos demográficos. Con esta sectorización, distintas empresas y organizaciones pueden ajustar el contenido y las herramientas que proveen a los usuarios con fines de mercadotecnia, promoción política, programas sociales, información educativa, entretenimiento, entre otros. Por ejemplo, en la mercadotecnia puede apoyar a las empresas para realizar campañas de productos para usuarios con características específicas. Adicionalmente, con propósitos de seguridad, usando el perfilado de autor se puede lograr una identificación primaria de usuarios que tienen un comportamiento anómalo (acoso, hostigamiento, intento de robo de información, terrorismo) dentro de las redes sociales y cuya información demográfica está ofuscada.

En el presente artículo se realiza un estudio sobre el perfilado demográfico de celebridades de redes sociales. Una celebridad se considera un usuario de la red que tiene un número considerable de seguidores dentro de la misma. La tarea consiste en analizar los mensajes de texto publicados o compartidos por los seguidores de las celebridades, y con base en ello predecir los atributos demográficos de género, ocupación y año de nacimiento.

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 tweets de 1,920 celebridades y por tweets de seguidores de 400 celebridades respectivamente. Para este trabajo, de los conjuntos de datos excluimos los tweets que utilizaran un alfabeto no occidental y que no estuvieran en inglés. 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).

Partiendo de estos datos, a los tweets de los conjuntos de datos se les extrajeron las siguientes características textuales: palabras, emoticones/emojis, etiquetas (# o hashtags), menciones (@ o ats), abreviaturas y los vectores de palabras fastText, word2vec y GloVe. Cada una de estas características revela diferentes aspectos del contenido que generan o comparten los usuarios.

Empleando las características extraídas se construyeron modelos de aprendizaje de máquina para realizar la predicción de los atributos demográficos. Se entrenaron con el conjunto de entrenamiento y probaron con el conjunto de prueba los modelos de clasificadores multinomiales simples de Bayes (MNB o Multinomial Naïve Bayes), k vecinos más cercanos (KNN o K-Nearest Neighbors), bosques aleatorios (RF o Random Forest), regresión logística (LR o Logistic Regression) y máquinas de vectores de soporte lineales (LSVM o Linear Support Vector Machines).

Para el estudio, se experimentó con las combinaciones de los modelos de aprendizaje y características textuales usando los conjuntos de entrenamiento para entrenar un modelo y el de prueba para evaluar el desempeño de éste. El desempeño de cada combinación se midió utilizando la métrica del área bajo la curva ROC (AUC o Area Under the Curve), que es una métrica popular en clasificación de textos, principalmente cuando se tienen clases desbalanceadas (donde algunas clases tienen mayor cantidad de ejemplos de entrenamiento que otras).

La contribución de nuestro trabajo radica en el estudio del desempeño de diferentes características textuales y modelos de aprendizaje de máquina para la tarea de perfilado demográfico de redes sociales, intentando responder las siguientes preguntas de investigación: 1) ¿Hay un modelo de aprendizaje de máquina con mejor desempeño? 2) ¿Hay una característica textual con un mejor desempeño? 3) ¿Hay una combinación de modelo de aprendizaje y característica textual con un mejor desempeño?

2. 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 la tarea de perfilado se incluyen la edad, el género, la ocupación, el nivel socioeconómico, entre otros; siendo la predicción de edad y género los atributos más populares para determinar [4]. Sin embargo, otras subtareas como la identificación de rasgos de personalidad [2] u ocupación [1], también han cobrado relevancia en años recientes.

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 y rasgos de personalidad [5, 6, 7, 8, 9, 10]. 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 [11], bolsa de palabras (bag-of-words), n-gramas, diccionario de palabras, vectores de palabras, entre otras. De igual manera, se han utilizado diferentes modelos de aprendizaje de máquina como máquinas de vectores de soporte, regresión logística, clasificadores bayesianos y modelos de aprendizaje profundo (*deep learning*).

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

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 [12] ya sea por la celebridad o por sus seguidores [1].

Para el perfilado de celebridades utilizando el contenido generado por las mismas, en [13] utilizaron máquinas de vectores de soporte y regresión logística para la predicción de ocupación, edad y género. Los autores en [14] 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 [15], 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 [16] usaron una matriz de tf-idf que se introdujo en una red neuronal LSTM para la predicción. Los autores en [17] 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 [18], 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.

3. Metodología

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

3.1. Adquisición 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], los cuales fueron extraídos directamente de Twitter por los organizadores de la conferencia. El conjunto de datos de prueba está formado por el contenido textual de las publicaciones realizadas por los seguidores de 400 celebridades; mientras que el conjunto de datos de entrenamiento está conformado por el contenido textual de las publicaciones realizadas por 1,920 celebridades. De ambos conjuntos de datos se eliminaron aquellas publicaciones con un alfabeto diferente al occidental. Los tweets en su mayoría se encuentran en inglés, con algunos en otros idiomas como el español. 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 ambos conjuntos de datos. Como se puede ver en la tabla 1, la distribución de usuarios del conjunto de prueba es homogénea tanto para género como para las ocupaciones. En contraste, en la tabla 2 se observa que el número de usuarios hombres (56%) es ligeramente mayor al número de usuarios mujeres (44%), mientras que se observa una distribución más homogénea para cada una de las clases del atributo ocupación.

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

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

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

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 entrenamiento).

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. Tanto para el conjunto de prueba como para el conjunto de entrenamiento 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.

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

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

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 entrenamiento).

3.2. Procesamiento de Datos

De cada tweet se extrajeron cinco características textuales: palabras, emoticones/emojis, etiquetas (# o hashtags), menciones (@ o ats) y abreviaturas comunes. Primero se concatenaron todos los tweets correspondientes a un usuario en una sola cadena de texto. El proceso se aplicó a todos los usuarios, de tal forma que un usuario queda expresado como una cadena de larga de texto. Posteriormente, se emplearon una serie de expresiones regulares para la extracción de las cinco características textuales. Para las palabras se removieron aquellas muy cortas (longitud < 3), muy largas (longitud > 35) y las palabras vacías (*stopwords*). Para ello, se utilizó una lista de palabras vacías en inglés proporcionada por la librería NLTK en Python. En el caso de las abreviaturas, se recopiló a través de internet una lista de las 1,374 abreviaturas más comunes en Twitter y solo se reconocieron las abreviaturas que aparecían en la lista.

Al final del proceso de limpieza, agrupamiento de información y extracción de características, se obtuvieron cinco archivos por cada conjunto de datos. Cada archivo del conjunto de prueba contiene 400 líneas, cada archivo del

conjunto de entrenamiento contiene 1,920 líneas. Cada línea de los archivos del conjunto de prueba representa las características de los tweets de los seguidores de una celebridad, cada línea de los archivos de entrenamiento representa las características de los tweets de las celebridades.

Para cada una de las características textuales se extrajo un vocabulario (conjunto de características únicas). En las tablas 5 y 6 se muestran los tamaños de cada vocabulario para el conjunto de datos de prueba y para el conjunto de datos de entrenamiento respectivamente. En ambas tablas se observa que las características con vocabularios más extensos son las menciones y las palabras; mientras que las abreviaturas tienen el vocabulario más pequeño.

Palabras	Emoticones	Etiquetas	Menciones	Abreviaturas
1,195,082	1,519	930,010	2,413,799	1000

Tabla 5. Tamaño del vocabulario por característica (conjunto de datos de prueba).

Palabras	Emoticones	Etiquetas	Menciones	Abreviaturas
662,308	1,334	407,959	1,068,617	864

Tabla 6. Tamaño del vocabulario por característica (conjunto de datos de entrenamiento).

Utilizando el vocabulario correspondiente de cada una de las características, se realizó un proceso de vectorización con el método tf-idf (*term frequency inverse document frequency*), el cual se encuentra definido por la ecuación 1.

$$tfidf(t,d) = tf(t,d)\ x\ idf(t) \quad (1)$$

En donde *tf(t,d)* es la frecuencia en la que ocurre la término *t* en el documento *d*, y el término *idf* se encuentra definido por la ecuación 2.

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

En donde *df(t)* es el número de documentos que contienen al término *t*, y el término n_d es el número total de documentos. Para cada característica del conjunto de entrenamiento se calculó el término *idf* el cual sería utilizado para transformar el conjunto de prueba con la característica correspondiente.

Adicionalmente a las cinco características textuales mencionadas, se construyeron matrices usando como características los vectores de palabras fastText, word2vec y GloVe. Estos modelos miden estadísticas de coocurrencia entre palabras a partir de un conjunto de datos de entrenamiento. Para este trabajo se utilizaron modelos preentrenados sobre conjuntos de datos en inglés. En el caso de fastText se utilizó un modelo creado a partir de datos de Wikipedia y Common Crawl[4] el cual contiene un diccionario de más de 2 millones de palabras, cada una representada con un vector de 300 características. En el caso de word2vec se utilizó un modelo creado a partir de noticias de Google[5] en inglés, el cual contiene un diccionario de más de 3 millones de palabras, cada una representada con un vector de 300 características. En el caso de GloVe se utilizó un modelo creado con 2 billones de publicaciones de Twitter[6], representando cada palabra como un vector de 200 características. Para estas características de vectores se calculó el vector promedio de todos los vectores de palabras encontradas en los tweets de un usuario. De esta manera, cada celebridad se presenta como un vector promedio de 300 características densas. El proceso de vectorización se aplicó a todos los usuarios en cada conjunto de entrenamiento y prueba.

3.3. Experimentación

Al terminar el proceso de vectorización, se realizó una experimentación con diferentes modelos de aprendizaje de máquina. Los modelos que se utilizaron siguen varios enfoques: probabilístico (clasificador simple de Bayes o MNB), basado en instancias (k vecinos más cercanos o KNN), reglas de decisión (bosques aleatorio o RF), y discriminativos (máquinas de vectores de soporte lineales o LSVM, y regresión logística o LR).

[4] Disponible en: https://fasttext.cc/docs/en/supervised-models.html
[5] Disponible en: https://radimrehurek.com/gensim/models/word2vec.html
[6] Disponible en: https://nlp.stanford.edu/projects/glove/

Con los modelos LSVM, LR, RF y KNN, se realizó una validación cruzada estratificada de 3 partes para cada conjunto de entrenamiento, con el fin de encontrar los valores óptimos para sus hiperparámetros. En la tabla 7 se pueden observar los diferentes valores que se consideraron en la optimización del hiperparámetro de cada modelo. Una vez encontrado el valor óptimo, se construye el modelo final con ese valor y con todo el conjunto de entrenamiento.

Modelo	Parámetro	Descripción	Valores
KNN	k	Número de vecinos	[1, 2, 3, 5, 10]
RF	r	Número de árboles	[50, 100, 150, 200]
LR	c	Parámetro de Regularización	[0.1, 1, 10, 100]
LSVM	c	Parámetro de Regularización	[0.1, 1, 10, 100]

Tabla 7. Valores considerados para los hiperparámetros.

Para medir el desempeño de los modelos, se utilizó una métrica basada en la matriz de confusión, formada por las celdas de: verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Esta matriz muestra la relación entre las clases reales de los usuarios contra las clases predichas por los modelos. La métrica utilizada para evaluar a cada uno de los modelos fue el área bajo la curva ROC (*Receiver Operating Characteristic*). La curva ROC grafica la razón de verdaderos positivos contra la razón de falsos positivos en varios umbrales. El área bajo la curva ROC (AUC o *Area Under the Curve*) evalúa el grado de separabilidad, midiendo la probabilidad de que un modelo clasifique a un usuario elegido aleatoriamente en una clase, más que a un usuario de otra clase elegido aleatoriamente. Esta métrica es particularmente útil cuando la distribución entre clases no es uniforme, como es el caso del año de nacimiento en nuestro conjunto de prueba.

Todos los códigos para el procesamiento y experimentación se realizaron en Python utilizando las librerías NLTK, emoji, scikit-learn y fasttext.

4. Resultados

En las tablas 8, 9 y 10 se muestran los resultados obtenidos por las distintas características textuales y los diferentes modelos de aprendizaje de máquina utilizados para la predicción de los atributos de género, ocupación y año de nacimiento respectivamente.

En las tablas, los renglones 3 a 7 indican los modelos de aprendizaje probados: MNB, KNN, RF, LR, LSVM. Las columnas 2 a 9 indican las características textuales extraídas de las publicaciones para construir y probar los modelos: palabras, emojis/emoticones, etiquetas, menciones, abreviaturas y los vectores de palabras GloVe, word2vec y fastText. Las celdas muestran la métrica AUC para el uso de un modelo con una característica. El renglón 8 muestra el promedio de la métrica de forma transversal para todos los modelos por característica. De forma similar, la columna 10 muestra el promedio de la métrica de forma transversal para todas las características por modelo.

En lo que respecta al género, la combinación de palabras, emoticones, etiquetas o vectores de palabras con los modelos LR o LSVM produjeron resultados similares, alcanzando entre 0.63 y 0.65 en el AUC, siendo estos los valores más altos para la predicción de este atributo. Considerando el promedio general, las características de vectores de palabras se desempeñaron mejor que las demás características en la predicción de este atributo, siendo las menciones las que peor se desempeñaron. Se puede establecer que las características de vectores de palabras son más adecuadas para predecir el género de las celebridades, debido a sus buenos resultados y su corto tiempo de entrenamiento y prueba en comparación a otras características. Los promedios generales de los modelos muestran que LR y LVSM fueron los que mejor se desempeñaron, pero considerando el tiempo de ejecución, el más adecuado para predecir el género de las celebridades es LR.
Se puede especular que los errores para predecir este atributo pueden deberse a la superposición de palabras entre géneros, es decir, que las celebridades de ambos géneros utilizan palabras similares con la misma frecuencia.

Modelo	Características								
	Palabras	Emoticones	Etiquetas	Menciones	Abreviaturas	GloVe	Word2vec	fastText	Promedio
MNB	0.52	0.58	0.58	0.52	0.50	--	--	--	0.54
KNN	0.57	0.56	0.60	0.53	0.50	0.57	0.59	0.60	0.57
RF	0.62	0.56	0.53	0.59	0.59	0.65	0.6	0.63	0.60
LR	0.64	0.64	0.65	0.56	0.61	0.64	0.63	0.64	0.63
LSVM	0.64	0.63	0.64	0.57	0.59	0.64	0.64	0.64	0.62
Promedio	0.60	0.59	0.60	0.55	0.56	0.63	0.62	0.63	

Tabla 8. Resultados (AUC) para género.

Para el atributo de ocupación, se obtuvieron resultados más favorables que con el género, siendo las mejores combinaciones de modelo y atributo nuevamente LR o LVSM con palabras, etiquetas o los vectores de palabras, que obtienen un AUC entre 0.83 y 0.85. La característica recomendable para predecir el atributo de ocupación en las celebridades son las etiquetas (hashtags), gracias a su desempeño y a su menor tiempo de ejecución. Se puede especular que es debido a que las personas de diferentes ocupaciones tienden a usar etiquetas similares con mayor frecuencia. En cuanto a los modelos de clasificación, la mejor opción es evaluar las etiquetas con los modelos MNB, LR o LSVM.

Modelo	Características								
	Palabras	Emoticones	Etiquetas	Menciones	Abreviaturas	Glove	Word2Vec	FastText	Promedio
MNB	0.79	0.81	0.84	0.81	0.72	--	--	--	0.79
KNN	0.74	0.70	0.80	0.76	0.69	0.77	0.75	0.76	0.75
RF	0.80	0.74	0.80	0.81	0.77	0.82	0.82	0.82	0.80
LR	0.83	0.75	0.84	0.82	0.74	0.84	0.84	0.84	0.81
LSVM	0.83	0.76	0.84	0.81	0.73	0.84	0.82	0.85	0.81
Promedio	0.80	0.75	0.82	0.80	0.73	0.82	0.81	0.82	

Tabla 9. Resultados (AUC) para ocupación

Para el atributo de año de nacimiento, las mejores combinaciones son nuevamente usando los modelos LR o LSVM en conjunto con las palabras o las características de vectores de palabras, con las cuales se obtiene un AUC entre 0.64 y 0.67. Los promedios transversales más altos fueron obtenidos por GloVe y fastText con un 0.61 en ambas características. El modelo con un promedio transversal más alto en este caso fue el LSVM con un 0.63. Comparando los resultados de este atributo con los anteriores parecen resultados poco favorables, pero en comparación con los atributos anteriores, la predicción del año de nacimiento es una tarea más compleja. Una razón importante es por el gran número de clases posibles (60 años/clases en comparación de 2 para género y 4 de ocupación); considerando que, para la clasificación de textos, en general, entre mayor número de clases existe, más complejo es el problema, como se ha observado en otros ámbitos [19]. Una segunda razón es que tiene un desbalanceo entre clases más pronunciado que en los otros atributos, lo cual afecta en mayor medida el desempeño.

Modelo	Características								
	Palabras	Emoticonos	Etiquetas	Menciones	Abreviaturas	GloVe	Word2vec	fastText	Promedio
MNB	0.56	0.60	0.58	0.56	0.55	--	--	--	0.57
KNN	0.51	0.50	0.51	0.50	0.50	0.51	0.51	0.51	0.51
RF	0.56	0.52	0.56	0.60	0.53	0.59	0.56	0.60	0.57
LR	0.64	0.56	0.58	0.57	0.63	0.66	0.65	0.65	0.62
LSVM	0.67	0.60	0.59	0.59	0.58	0.66	0.66	0.66	0.63
Promedio	0.59	0.56	0.56	0.56	0.56	0.61	0.60	0.61	

Tabla 10. Resultados (AUC) para año de nacimiento

En la tabla 11 se muestra el promedio de los resultados para los tres atributos en conjunto (género, ocupación y año de nacimiento). Los valores más altos de desempeño en conjunto se obtienen con las combinaciones de palabras y vectores de palabras en conjunto con los modelos LR y LSVM. Los vectores de palabras obtienen los mejores promedios transversales entre características, mientras que los modelos LR y LSVM obtienen los mejores promedios entre modelos. Considerando los tiempos de ejecución, las mejores características serían los vectores de palabras GloVe o fastText; mientras que el mejor modelo sería LSVM.

Modelo	Características								
	Palabras	Emoticonos	Etiquetas	Menciones	Abreviaturas	GloVe	Word2vec	fastText	Promedio
MNB	0.62	0.66	0.66	0.63	0.59	--	--	--	0.63
KNN	0.60	0.58	0.63	0.59	0.56	0.61	0.61	0.62	0.60
RF	0.66	0.60	0.63	0.66	0.63	0.68	0.66	0.68	0.65
LR	0.70	0.65	0.69	0.65	0.66	0.71	0.70	0.71	0.68
LSVM	0.71	0.67	0.69	0.65	0.64	0.71	0.70	0.71	0.69
Promedio	0.66	0.63	0.66	0.64	0.62	0.68	0.67	0.68	

Tabla 11. Promedio de resultados para los tres atributos (AUC)

En la tabla 12 se muestra el promedio de tiempo de ejecución total de cada uno de los modelos para los tres atributos demográficos. El tiempo de ejecución total es la suma del tiempo de lectura, tiempo de agrupación de característica y de etiquetas para ambos conjuntos de prueba y entrenamiento, además de los tiempos de validación, entrenamiento y predicción del modelo. Los modelos que usan las palabras como característica son los que más tardados, esto es debido a que esta característica contiene las palabras de 40,049,803 tweets, para estos modelos el proceso de agrupación por usuario del conjunto de prueba tomaba un promedio de 28 horas. Los modelos que usan vectores de palabras como característica (glove, word2vec y fasttext) son los modelos que presentaron el menor tiempo de ejecución total, esto es debido a que el tiempo de agrupamiento es nulo ya que construimos los vectores promedio de palabras previamente, al momento del proceso de lectura los datos ya vienen agrupados. En la décima columna está el total de tiempo de ejecución de cada modelo, el modelo que presentó menor tiempo de ejecución fue LR.

Modelo	Características								
	Palabras	Emoticones	Etiquetas	Menciones	Abreviaturas	GloVe	Word2vec	fastText	Total
MNB	26.58	1.85	3.87	6.04	0.36	--	--	--	38.70
KNN	26.58	1.84	3.88	6.04	0.37	0.002	0.002	0.002	38.71
RF	31.45	1.36	3.21	6.36	0.35	0.005	0.004	0.005	42.74
LR	26.75	2.04	3.21	5.33	0.33	0.007	0.005	0.007	37.68
LSVM	35.27	1.85	2.85	5.22	0.34	0.012	0.012	0.012	45.56
Promedio	29.32	1.78	3.40	5.80	0.35	0.006	0.005	0.006	

Tabla 12. Promedio de tiempos de ejecución de los modelos para los tres atributos (horas)

5. Conclusiones

En esta investigación se estudió el comportamiento de distintas características textuales en combinación con diversos modelos de aprendizaje de máquina para la tarea de perfilado demográfico de celebridades de redes sociales.

Para esta tarea se analizó los mensajes de texto publicados o compartidos por los seguidores de una celebridad y con base en ellos se predijeron los atributos demográficos de la celebridad, que consistían en el género, la ocupación y el año de nacimiento. Se utilizaron un conjunto de datos de entrenamiento de 5,066,608 tweets, correspondientes a 1,920 celebridades de Twitter, y un conjunto de datos de prueba de 34,893,195 tweets, correspondientes a los seguidores de 400 celebridades de Twitter.

De acuerdo con los experimentos, para predecir el perfil demográfico de celebridades de Twitter, se concluye lo siguiente:

- Los vectores de palabras GloVe, fastText y word2vec, como características para representar el contenido textual de los usuarios, tienen el menor tiempo de ejecución total y el mejor desempeño para predecir los atributos demográficos, tanto de forma individual como su desempeño agregado.
- Las etiquetas como característica tienen resultados muy buenos, tan solo 3% debajo del mejor resultado obtenido por los vectores de palabras. Presentaron un tiempo de ejecución de casi diez veces menor que los modelos que usaron palabras como característica.
- Otras características textuales como las palabras también muestran un buen desempeño en la predicción; sin embargo, su uso implica una representación más extensa que consume más memoria, y requiere de un mayor tiempo de entrenamiento y prueba de los modelos de aprendizaje.
- El resto de las características textuales, emoticones, menciones y abreviaturas, presentan un desempeño moderado. Es de resaltar el uso de las abreviaturas, que con un vocabulario tan pequeño mantienen un comportamiento aceptable, con valores entre 3% a 10% abajo de los mejores resultados.
- Los modelos de aprendizaje que siguen un enfoque discriminativo, LR y LSVM, tiene el mejor desempeño para predecir los atributos de género y ocupación para las celebridades; mientras que el modelo LSVM tiene el mejor desempeño para predecir el año de nacimiento. De forma agregada, el modelo LSVM es el que tiene el mejor desempeño promedio por cuestión de un par de milésimas en comparación de LR. No obstante, el modelo LR presenta mejores tiempos de entrenamiento y prueba, por lo que para los atributos de género y ocupación sería recomendable su uso.

Referencias

1. Wiegmann, M., Stein, B., Potthast, M.: Overview of the celebrity profiling task atpan 2020. In: CLEF (2020)
2. 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.
3. 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).
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. Wiegmann, M., Stein, B., & Potthast, M. (2019, September). Overview of the Celebrity Profiling Task at PAN 2019. In *CLEF (Working Notes)*.
13. Radivchev, V., Nikolov, A., & Lambova, A. (2019, September). Celebrity Profiling using TF-IDF, Logistic Regression, and SVM. In *CLEF (Working Notes)*.
14. 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.
15. Martinc, M., Skrlj, B., & Pollak, S. (2019, September). Who is Hot and Who is Not? Profiling Celebs on Twitter. In *CLEF (Working Notes)*.
16. 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)*.
17. Hodge, A., & Price, S. (2020). Celebrity Profiling using Twitter Follower Feeds.
18. Koloski, B., Pollak, S., & Skrlj, B. (2020). Know your Neighbors: Efficient Author Profiling via Follower Tweets. In *CLEF (Working Notes)*.
19. Gomez, J. C. (2019). Analysis of the effect of data properties in automated patent classification. *Scientometrics*, *121*(3), 1239-1268.