ISSN 1870-4069

Perfilado demográfico de celebridades de redes sociales

Juan-Carlos Alonso-Sánchez, Luis-Miguel López-Santamaría, Juan Carlos Gomez

Universidad de Guanajuato,
Departamento de Ingeniería Electrónica,
México

{jc.alonsosanchez, lm.lopezsantamaria, jc.gomez}@ugto.mx

Resumen. El perfilado de autor en redes sociales es una tarea que ha tomado auge en los últimos años para tratar de predecir de forma automática los atributos demográficos de una población objetivo de usuarios, a partir de la información que éstos generan o comparten. Estos atributos pueden ser aprovechados por distintas organizaciones y compañías para propósitos de seguridad, mercadotecnia, educación, estadísticas poblacionales, entre otros. En este artículo se presenta un estudio sobre el análisis de los mensajes de texto publicados por celebridades de redes sociales (usuarios populares), para con base en ello predecir el perfil demográfico de tales usuarios, formado por su género, ocupación y año de nacimiento. Para la tarea se utiliza un conjunto de datos de 1,920 celebridades de Twitter, formado por 5,066,608 tweets principalmente en inglés. 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. Se realizó una evaluación del uso de características y modelos siguiendo una validación cruzada estratificada de 10 partes y 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 fastText 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.

Demographic Profiling of Celebrities in Social Networks

Abstract. Author profiling in social media is a task that has become popular in recent years to automatically predict the demographic attributes from a population of users, based on the information they generate and share. These attributes can be exploited by different organizations and companies for purposes of security, marketing, education, population statistics, among others. This article

presents a study on the analysis of text messages posted by social network celebrities (popular users), to predict the demographic profile of these users, conformed by their gender, occupation and year of birth. A dataset of 5,066,608 tweets, mainly in English, by 1,920 Twitter celebrities is used for the task. From this data, experiments were conducted by extracting a series of textual features from the tweets and with these features various machine learning models were built. An evaluation of the different features and models was performed following a stratified 10-fold cross-validation, measuring their performance with the area under the ROC. The results indicate that some attributes such as year of birth are difficult to predict. It is also observed, that features such as fastText word vectors perform well, especially in combination with discriminative learning models.

Keywords: Author profiling, data mining, machine learning, social networks.

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 [20], rasgos de personalidad [9], nivel educativo, orientación política [2], entre otros.

Esta tarea ha tomado relevancia en los últimos años gracias a la abundante información que las personas generan y comparten en distintos medios a través de internet.

Uno de los medios más populares donde las personas crean y comparten información es en las redes sociales, que cuentan con millones de usuarios que todos los días expresan sus gustos, opiniones e ideas a través de publicaciones que contienen información en varias modalidades, como texto, imágenes y video.

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 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 oculta, esto con propósitos de seguridad.

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 considerables de seguidores dentro de la misma. La tarea consiste en analizar los mensajes de texto publicados o compartidos por la celebridad y con base en ello predecir los atributos demográficos de género, ocupación y año de nacimiento.

Para conducir el estudio, se utilizó el conjunto de datos de entrenamiento publicado en PAN@CLEF 2020[1] que está formado por los tweets de 1,920 celebridades.

En este conjunto, un usuario se considera celebridad si tiene al menos 10 seguidores. Para este trabajo, del conjunto de datos original se filtraron los tweets que utilizaran un alfabeto no occidental, quedando un total de 5,066,608 tweets, con un promedio de 2,639 tweets por celebridad.

Los tweets en su mayoría se encuentran en inglés, con algunos en otros idiomas como el español. En el conjunto de datos, las celebridades están clasificadas en dos géneros (hombre, mujer), cuatro ocupaciones (político, creador, artista, deportista) y en 60 años de nacimiento (entre 1940 y 1999).

Partiendo de estos datos, a los tweets de las celebridades se les extrajeron las siguientes características textuales: palabras, emoticones/emojis, etiquetas (# o hashtags), menciones (@ o ats), abreviaturas y los vectores de palabras fastText. 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 y probaron 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 modelos de aprendizaje y características utilizando una validación cruzada estratificada de 10 partes, con el fin de obtener resultados consistentes estadísticamente. El desempeño de cada combinación se midió utilizando la métrica del área bajo la curva ROC (AUC), 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 para la tarea de perfilado demográfico de celebridades de redes sociales, intentando responde 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?

El resto del presente artículo se organiza de la siguiente manera. La sección 2 presenta una revisión de los trabajos relacionados encontrados en la literatura. La sección 3 explica la metodología utilizada para el estudio, incluyendo la descripción del conjunto de datos y los detalles de la experimentación. La sección 4 muestra los resultados obtenidos con el estudio de modelos de aprendizaje y características textuales. Finalmente, la sección 5 presenta las conclusiones y algunas ideas para trabajos futuros.

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

2. Trabajos relacionados

El estudio de perfilado de autor en redes sociales utilizando el 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 extraído 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 [3]. Sin embargo, otras subtareas como la identificación de rasgos de personalidad [9] u ocupación [20], 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 [14, 12, 16, 17, 15, 13].

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 el español, el portugués, el italiano, el neerlandés y el 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 [7], 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).

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 determinado 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 [19] por la celebridad o por sus seguidores [20].

Para el perfilado de celebridades utilizando el contenido generado por las mismas, en [11] utilizaron máquinas de vectores de soporte y regresión logística para la predicción de ocupación, edad y género. Los autores en [10] 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 [8], los autores emplearon vectores tf-idf (term-document frequencye inverse document frequency) formados a partir de unigramas de palabras, así como también trigramas de caracteres delimitados por palabras.

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

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 [1] usaron una matriz de tf-idf generada a partir del contenido textual generado por los seguidores.

Esta matriz se introdujo en una red neuronal LSTM para la predicción. Los autores en [5] 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.

Utilizaron modelos de regresión logística, máquinas de vectores de soporte y bosques aleatorios. Por otro lado, en [6], 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 se encuentra conformada por tres fases que son la adquisición de los datos, el procesamiento de los datos y la experimentación. En la última fase se describen los procesos de la construcción de modelos y la evaluación de estos mismos. Las tres fases se encuentran descritas a continuación.

3.1. Adquisición de los datos

En este artículo se utilizó el conjunto de datos de la conferencia PAN@CLEF 2020 para la tarea de celebrity profiling[3], el cual se extrajo directamente de Twitter por los organizadores de la conferencia. Este conjunto de datos está conformado por el contenido textual de las publicaciones realizadas por 1,920 celebridades.

Del conjunto original se eliminaron aquellas publicaciones con un alfabeto diferente al occidental, quedando un total de 5,066,608 tweets, para un promedio de 2,639 tweets por celebridad. Los tweets en su mayoría se encuentran en inglés, con algunos en otros idiomas como el español. Las celebridades 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 la Tabla 1 se observa la distribución de usuarios por género y ocupación. Como se puede ver en la tabla, el número de usuarios hombres (56 %) es ligeramente mayor al número de usuarios mujeres (44 %).

Esta distribución de género refleja en cierta medida la que se encuentra en Twitter, donde el 68.5 % de los usuarios son hombres[4]. Por otro lado, se observa una distribución más homogénea para cada una de las clases del atributo ocupación.

Por motivos de ilustración, se agruparon los años de nacimiento en décadas, y su distribución con respecto al género se muestra en la Tabla 2.

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

[4] https://bit.ly/2QqCiRs

Tabla 1. Distribución de usuarios por género y ocupación.

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 década de nacimiento.

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

En la tabla 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.

3.2. Procesamiento de datos

En esta paso se procesaron los tweets para extraer diferentes características textuales. 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 extracción de cinco características textuales: palabras, emoticones/emojis, etiquetas (# o hashtags), menciones (@ o ats) y abreviaturas comunes. Para la palabras se realizó un proceso en donde se removieron aquellas palabras que eran muy cortas (longitud < 3), muy largas (longitud > 35) y palabras vacías (stopwords).

Para ello, se utilizó una lista de palabras vacías en inglés proporcionada por la librería NLTK. En el caso de las abreviaturas, se recopiló a través de internet una lista de las 1,374 abreviaturas más comunes en Twitter.

Al final del proceso limpieza, agrupamiento de información y extracción de características, se obtuvieron cinco archivos. Cada archivo contenía 1,920 líneas; siendo cada una de estás líneas las características de una celebridad. Utilizando una validación cruzada estratificada a 10 partes, se dividió cada archivo en 10 conjuntos de entrenamiento y 10 conjuntos de prueba.

Para cada una de las características textuales se extrajo un vocabulario (características únicas).

En la Tabla 3 se muestran los tamaños de cada vocabulario para el conjunto de datos completo. En esta tabla se observa que las características con vocabularios más extensos son las menciones y las palabras; mientras que las abreviaturas tienen el vocabulario más pequeño.

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:

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

Tabla 3. Tamaño del vocabulario por característica.

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

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

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

En la ecuación 2, $df(d, t)$ es el número de documentos d que contienen al término t, el término n_d es el número total de documentos. Para cada conjunto de entrenamiento de cada característica se calculó el término idf el cual sería utilizado para la vectorización del conjunto de prueba.

Adicionalmente, se construyeron matrices usando como características los vectores de palabras fastText. El modelo fastText mide estadísticas de coocurrencia entre palabras a partir de un conjunto de datos de entrenamiento.

Para este trabajo se utilizó un modelo preentrenado sobre un conjunto de datos en inglés de Wikipedia y Common Crawl[5], el cual contiene un diccionario de más de 2 millones de palabras, cada una representada con un vector de 300 características. Para las características de vectores fastText 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 aplico 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).

Se decidió utilizar estos modelos ya que, como ha sido mencionado por otros autores en la tarea de perfilado de autor utilizando información textual, los modelos basado en aprendizaje profundo no han logrado mejorar el desempeño de modelos más tradicionales [18, 5].

Con cada uno de los modelos se aplicó la validación cruzada estratificada de 10 partes, para que los resultados obtenidos fueran sólidos estadísticamente.

Con los modelos LSVM, LR, RF y KNN, se realizó una subvalidación cruzada de 3 partes para cada conjunto de entrenamiento, con el fin de encontrar los valores óptimos para sus hiperparámetros. En la Tabla 4 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.

[5] Disponible en: https://fasttext.cc/docs/en/supervised-models.html

Tabla 4. Valores considerados para los hiperparámetros.

Modelo	Parámetro	Descripción	Valores
KNN	k	Número de vecinos	[1, 2, 3, 5, 10]
RF	r	Número de árboles	[5, 10, 15, 20]
LR	c	Parámetro de regularización	[0.1, 1, 10, 100]
LSVM	c	Parámetro de regularización	[0.1, 1, 10, 100]

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 de los atributos género y año de nacimiento.

Como base de comparación se consideran dos modelos. El primero es uno totalitario, el cual asignaría todos los usuarios del conjunto de prueba a la clase mayoritaria. El segundo es uno aleatorio uniforme, el cual asignaría un usuario a una clase aleatoria con la misma probabilidad para todas. Para ambos modelos la métrica AUC sería de 0.5.

Todos los códigos para el procesamiento y experimentación se realizaron en Python utilizando las librerías NLTK, emoji, scikit-learn y fasttext. El código está disponible en el siguiente repositorio https://github.com/jcgcarranza/rcs_celebrity_profiling. Los datos procesados como fueron usados en este artículo están disponibles en https://zenodo.org/record/4767751.

4. Resultados

En las tablas 5, 6 y 7 se muestran los resultados obtenidos por las distintas características textuales extraídas y los diferentes modelos de aprendizaje 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 7 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 fastText.

Las celdas muestran el promedio de la métrica AUC para el uso de un modelo con una característica siguiendo la validación cruzada, con la desviación estándar entre paréntesis.

Tabla 5. Resultados (AUC) para género.

	Característica						
Modelo	**Palabras**	**Emoticones**	**Etiquetas**	**Menciones**	**Abreviaturas**	**fastText**	**Promedio**
MNB	0.71(0.37)	0.76(0.38)	0.71(0.36)	0.64(0.34)	0.57(0.31)	–	0.68
KNN	0.75(0.36)	0.72(0.37)	0.69(0.35)	0.70(0.36)	0.68(0.35)	0.77(0.39)	0.72
RF	0.77(0.39)	0.77(0.39)	0.72(0.37)	0.72(0.37)	0.74(0.38)	0.77(0.39)	0.75
LR	**0.88(0.44)**	0.79(0.40)	0.71(0.37)	0.68(0.36)	0.78(0.40)	**0.88(0.44)**	**0.79**
LSVM	**0.88(0.44)**	0.78(0.40)	0.71(0.37)	0.69(0.36)	0.76(0.39)	**0.88(0.44)**	0.78
Promedio	0.80	0.76	0.71	0.69	0.71	**0.83**	–

El renglón 8 muestra el promedio de la métrica de forma transversal para todos los modelos por característica. De forma similar, la columna 8 muestra el promedio de la métrica de forma transversal para todas las características por modelo.

En lo que respecta al género, se observa que tanto la combinación de las palabras con los modelos LR y LSVM, como la combinación de los vectores fastText con los mismos modelos, producen ambos resultados similares, alcanzando un 0.88 en AUC, siendo el resultado más alto para la predicción de este atributo y 38 % más alto que la base de comparación.

Si se revisan los promedios generales por modelo, se observa que LR es el que presenta el mejor desempeño con un 0.79, considerando el uso transversal de las características; aunque LSVM tiene un comportamiento similar. En cuanto a los promedios generales de características, se observa que fastText presenta el mejor resultado a lo largo del uso de distintos modelos con un 0.83.

Considerando que los vectores fastText producen una representación más pequeña que el uso de palabras, y por lo tanto un tiempo de entrenamiento y prueba más rápido, se puede considerar a esta característica como más adecuada para predecir el género de las celebridades. En cuanto al modelo de aprendizaje, LR usa el mismo principio que LSVM pero su tiempo de entrenamiento es menor, por lo que se puede considerar más adecuado para predecir el género de las celebridades. Por otro lado, dada la escala de los valores obtenidos en la predicción de este atributo, el atributo no es tan difícil de predecir, pero hay margen para mejorar.

Se puede especular que los errores en 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. Adicionalmente, también es posible que el desbalanceo entre las clases del atributo tenga una afectación negativa en el desempeño.

Analizando los resultados para la ocupación, se observa que los valores más altos de desempeño se obtienen con la combinación de palabras o vectores fastText con los modelos LR o LSVM. Todas estas combinaciones producen un valor de 0.94 para AUC, siendo 44 % más alto que la base de comparación.

Revisando los promedios generales por modelo, se determina que el valor más alto se obtiene con los modelos LR y LSVM, ambos obteniendo un desempeño de 0.89

Tabla 6. Resultados (AUC) para ocupación.

	Característica						
Modelo	**Palabras**	**Emoticones**	**Etiquetas**	**Menciones**	**Abreviaturas**	**fastText**	**Promedio**
MNB	0.91(0.38)	0.83(0.30)	0.89(0.35)	0.90(0.38)	0.80(0.31)	–	0.87
KNN	0.90(0.36)	0.76(0.26)	0.86(0.32)	0.90(0.37)	0.77(0.26)	0.91(0.37)	0.85
RF	0.89(0.35)	0.79(0.28)	0.84(0.32)	0.87(0.35)	0.83(0.30)	0.91(0.38)	0.86
LR	**0.94(0.40)**	0.84(0.31)	0.88(0.35)	0.90(0.38)	0.85(0.33)	**0.94(0.40)**	**0.89**
LSVM	**0.94(0.40)**	0.84(0.32)	0.89(0.35)	0.90(0.36)	0.85(0.33)	**0.94(0.34)**	**0.89**
Promedio	**0.92**	0.81	0.87	0.90	0.82	**0.92**	–

promediado de forma transversal a todas las características. En cuanto a los promedios generales por característica, tanto las palabras como los vectores fastText tienen el mejor desempeño con el uso de todos los modelos, con un desempeño de 0.92.

Por las mismas razones que con el género, se puede considerar a los vectores fastText y al modelo LR como las mejores opciones en cuanto a característica y modelo para predecir la ocupación de las celebridades.

En este caso, la escala de valores para la métrica es mayor que la del género, por lo que es un atributo más sencillo de predecir. Se puede especular que hay una mejor separación entre las distribuciones de palabras.

Es decir, que las celebridades de las diferentes ocupaciones utilizan palabras diferentes con diferentes frecuencias. Adicionalmente, el balanceo de los usuarios entre las clases de este atributo, afecta positivamente el desempeño.

En el caso de los resultados para el año de nacimiento, los valores de desempeño más altos se obtienen con la combinación de palabras o vectores fastText con el modelo LSVM, dando un valor de 0.69 para AUC, que es 19 % más alto que la base de comparación.

El promedio más alto para un modelo a lo largo de las características lo obtiene LSVM con 0.64; mientras que el promedio más alto para una característica a lo largo de los modelos lo obtiene tanto las palabras como los vectores fastText con 0.60.

De nueva cuenta, se puede considerar a los vectores fastText como la mejor característica para predecir el año de nacimiento de las celebridades, aunque en este caso el modelo recomendado es LSVM. Por las escalas de los valores de la métrica AUC, se observa que predecir el año de nacimiento es más complejo que los otros dos atributos.

Una razón importante es por el gran número de clases posibles (60 años/clases); 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 [4].

Una segunda razón es que se tiene un desbalanceo entre clases más pronunciado que en los otros atributos, lo cual afecta en mayor medida el desempeño.

En la Tabla 8 se presenta el promedio del desempeño de las combinaciones de características y modelos para los tres atributos: género, ocupación y año de nacimiento. Las combinaciones que obtienen mejores resultados son el uso de las palabras o los

Tabla 7. Resultados (AUC) para año de nacimiento.

	Característica						
Modelo	**Palabras**	**Emoticones**	**Etiquetas**	**Menciones**	**Abreviaturas**	**fastText**	**Promedio**
MNB	0.59(0.02)	0.65(0.02)	0.60(0.02)	0.59(0.02)	0.59(0.02)	–	0.60
KNN	0.51(0.01)	0.52(0.01)	0.51(0.02)	0.51(0.03)	0.51(0.01)	0.51(0.02)	0.51
RF	0.54(0.01)	0.54(0.01)	0.52(0.01)	0.53(0.02)	0.56(0.02)	0.55(0.02)	0.54
LR	0.66(0.02)	0.61(0.01)	0.57(0.02)	0.60(0.02)	0.63(0.03)	0.66(0.02)	0.62
LSVM	**0.69(0.01)**	0.68(0.02)	0.59(0.02)	0.57(0.02)	0.66(0.02)	**0.69(0.02)**	**0.64**
Promedio	**0.60**	0.60	0.56	0.56	0.59	**0.60**	–

Tabla 8. Promedio de resultados para los tres atributos (AUC).

	Característica						
Modelo	Palabras	Emoticones	Etiquetas	Menciones	Abreviaturas	fastText	**Promedio**
MNB	0.74	0.75	0.74	0.71	0.66	–	0.72
KNN	0.72	0.67	0.69	0.74	0.66	0.73	0.70
RF	0.74	0.70	0.70	0.71	0.71	0.74	0.71
LR	0.82	0.75	0.72	0.73	0.76	0.83	**0.77**
LSVM	**0.84**	0.77	0.73	0.72	0.76	**0.84**	**0.78**
Promedio	0.77	0.72	0.71	0.71	0.70	**0.78**	–

vectores fastText con el modelo LSVM alcanzando un valor de 0.84, un 34 % más alto que la base de comparación.

El mejor desempeño general de LSVM con respecto a LR se debe al desempeño de LSVM con el atributo de año de nacimiento, ya que en los otros dos atributos ambos modelos se comportan igual.

5. Conclusiones

En este artículo se desarrolló un estudio del comportamiento de distintas características textuales en combinación con diversos modelos de aprendizaje de máquina para la tarea de perfilado demográfico de celebridades en redes sociales.

Para esta tarea se analizó los mensajes de texto publicados o compartidos por una celebridad y con base en ellos se predijeron los atributos demográficos de género, ocupación y año de nacimiento.

Para ello se experimentó con un conjunto de 5,066,608 tweets, mayormente en inglés, correspondientes a 1,920 celebridades de Twitter. De acuerdo con los experimentos, para el predecir el perfil demográfico de celebridades de Twitter, se concluye lo siguiente:

– Los vectores de palabras fastText, como característica para representar el contenido textual de las celebridades, tienen el mejor desempeño para predecir los atributos demográficos de éstas, tanto de forma individual como en su desempeño agregado.
– 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 características textuales, emoticones, etiquetas, 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, tienen 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. 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.

Algunas ideas por explorar para trabajos futuros incluyen el uso de otros modelos de clasificación como las redes neuronales profundas, las cuales pueden funcionar adecuadamente con características densas como los vectores fastText. También se puede considerar el uso de características estilísticas, como las partes del discurso, o las frecuencias de palabras funcionales, puntuaciones o errores gramaticales.

Por último, sería interesante explorar el uso de métodos de extracción de características latentes tales como Latent Dirichlet Allocation, Latent Semantic Indexing, Principal Component Analasis, Biased Discriminant Analysis y Non Negative Matrix Factorization, los cuales se encargan de calcular asociaciones entre palabras para agruparlas en tópicos o temas.

Referencias

1. Alroobaea, R., Almulihi, A. H., Alharithi, F. S., Mechti, S., Krichen, M., Belguith, L. H.: A deep learning model to predict gender, age and occupation of the celebrities based on tweets followers. In: CLEF (Working Notes) (2020)
2. Cohen, R., Ruths, D.: Classifying political orientation on twitter: It's not easy! In: Proceedings of the International AAAI Conference on Web and Social Media (2013)
3. Garcia-Guzman, R., Andrade-Ambriz, Y. A., Ibarra-Manzano, M. A., Ledesma, S., Gomez, J. C., Almanza-Ojeda, D. L.: Trend-based categories recommendations and age-gender prediction for pinterest and twitter users. Applied Sciences, vol. 10, no. 17, pp. 5957 (2020)
4. Gomez, J. C.: Analysis of the effect of data properties in automated patent classification. Scientometrics, vol. 121, no. 3, pp. 1239–1268 (2019)
5. Hodge, A., Price, S.: Celebrity profiling using twitter follower feeds. In: Proceedings of the Working Notes of the Conference and Labs of the Evaluation Forum (2020)
6. Koloski, B., Pollak, S., Škrlj, B.: Know your neighbors: Efficient author profiling via follower tweets. In: CLEF (Working Notes) (2020)

7. López-Santamaría, L. M., Gomez, J. C., Almanza-Ojeda, D. L., Ibarra-Manzano, M. A.: Age and gender identification in unbalanced social media. In: 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP), pp. 74–80 (2019)
8. Martinc, M., Skrlj, B., Pollak, S.: Who is hot and who is not? profiling celebs on twitter. In: Proceedings of the Working Notes of the Conference and Labs of the Evaluation Forum (2019)
9. Moreno, D. R. J., Gomez, J. C., Almanza Ojeda, D. L., Ibarra Manzano, M. A.: Prediction of personality traits in twitter users with latent features. In: Proceedings of the International Conference on Electronics, Communications and Computers (CONIELECOMP), pp. 176–181 (2019)
10. Moreno-Sandoval, L. G., Puertas, E., Plaza-del Arco, F. M., Pomares-Quimbaya, A., Alvarado-Valencia, J. A., Alfonso, L.: Celebrity profiling on twitter using sociolinguistic. In: CLEF (Working Notes) (2019)
11. Radivchev, V., Nikolov, A., Lambova, A., Cappellato, L., Ferro, N., Losada, D., Müller, H.: Celebrity profiling using TF-IDF, logistic regression, and SVM. In: CLEF (Working Notes) (2019)
12. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. In: CEUR Workshop Proceedings, vol. 1180, pp. 898–927 (2014)
13. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. In: Proceedings of the Working Notes Papers of the CLEF, pp. 1–38 (2018)
14. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation, pp. 352–365 (2013)
15. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working notes papers of the CLEF, pp. 1613–0073 (2017)
16. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: Proceedings of the Conference and Lab of the Evaluation ForumEF, pp. 2015 (2015)
17. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: Cross-genre evaluations. Working Notes Papers of the CLEF, vol. 2016, pp. 750–784 (2016)
18. Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., Ohkuma, T.: Text and image synergy with feature cross technique for gender identification. In: Proceedings of the Working Notes Papers of the CLEF (2018)
19. Wiegmann, M., Stein, B., Potthast, M.: Overview of the celebrity profiling task at PAN 2019. In: Proceedings of the Working Notes of the Conference and Lab of the Evaluation Forum (Working Notes) (2019)
20. Wiegmann, M., Stein, B., Potthast, M.: Overview of the celebrity profiling task at pan 2020. In: Proceedings of the Conference and Lab of the Evaluation Forum (2020)