Predicción automática del nivel educativo en usuarios de *Twitter* en México

Automatic Prediction of the Educational Level of Twitter Users in Mexico

Juan Carlos Gomez, Luis Miguel López Santamaría, Mario Alberto Ibarra Manzano y Dora Luz Almanza Ojeda*

Para resolver la situación planteada en este trabajo, se extrajo una serie de características del contenido textual de los tuits publicados por los usuarios, que se utilizaron para construir modelos basados en aprendizaje automático, los cuales predicen si un usuario tiene estudios universitarios o no. Ambos se probaron con un conjunto de datos extraído de forma directa del sitio, compuesto por más de un millón de tuits en español, correspondientes a 195 usuarios ubicados en México. Con él, se hicieron experimentos siguiendo una validación cruzada de 10 partes. La evaluación se realizó utilizando las métricas macro *F1* y el área bajo la curva ROC (AUC). Los resultados indican que la tarea es compleja, siendo las mejores características las abreviaturas, que alcanzaron valores arriba de 60 % para ambas métricas, mientras que los modelos de máquinas de vectores de soporte y árboles de decisión presentaron un desempeño similar.

Palabras clave: analítica de datos; *Twitter*; perfilado de autor; aprendizaje automático.

To solve the task raised in this work, a series of characteristics were extracted from the textual content of the tweets published by users, which were used to build models based on machine learning, which predict whether a user has a university degree or not. Both were tested with a data set extracted directly from the site, composed of more than one million tweets in Spanish, corresponding to 195 users located in Mexico. With it, experiments were made following a 10-fold cross-validation. The evaluation was performed using the F1 macro metrics and the area under the ROC (AUC) curve. The results indicate that the task is complex, the best characteristics being the abbreviations, which reached values above 60% for both metrics, while the support vector and decision tree machine models showed similar performance.

Key words: data analytics; Twitter; author profiling; machine learning.

Recibido: 14 de noviembre de 2019.
Aceptado: 9 de julio de 2020.

* Universidad de Guanajuato, Campus Irapuato-Salamanca, jc.gomez@ugto.mx (autor de correspondencia), lm.lopezsantamaria@ugto.mx, ibarram@ugto.mx, dora.almanza@ugto.mx, respectivamente.

Nota: esta investigación fue apoyada por el Fondo Sectorial CONACYT-INEGI, proyecto número 290910.

Futuristic cyborgs learning about humans/gremlin/iStock

Introducción

El análisis de las redes sociales visualiza al entorno social como un conjunto de patrones y conexiones entre unidades de interacción (usuarios) a partir de los cuales es posible estudiar relaciones económicas, políticas y afectivas existentes entre los usuarios (Wasserman & Faust, 1994). En este contexto, surge la tarea denominada *perfilado de autor* (*Author Profiling*), la cual se refiere al análisis del contenido generado o compartido en estos medios (p. ej. *Facebook* y *Twitter*, entre otros) para predecir diferentes atributos de quien los realiza, como género, edad, personalidad u orientación política.

El perfilado de autor tiene una amplia gama de aplicaciones en áreas como mercadotecnia, política, seguridad y educación; por ejemplo, en esta puede apoyar a las plataformas digitales para adecuar las estrategias de enseñanza para perfiles de estudiantes específicos. Por otro lado, muchas compañías y organizaciones lo utilizan para segmentar a sus clientes/usuarios para determinar qué contenido proveerles con fines de publicidad, campañas políticas, programas sociales o entretenimiento, entre otros.

Uno de los enfoques más populares para perfilar los atributos de interés (edad, género, etc.) de una persona consiste en entrenar modelos basados en aprendizaje automático (*Machine Learning*) utilizando el texto publicado por un grupo de personas de las cuales ya se tiene conocimiento de los atributos de interés. Una vez entrenados los modelos, se pueden usar para predecir esas particularidades en individuos no conocidos de forma previa.

Dentro de las redes sociales, *Twitter* es una de las más populares en diversos países; cuenta con,

al menos, 330 millones de usuarios activos,[1] de los cuales se conectan diario cerca de 126 millones.[2] En México, es la tercera más popular con cerca de 5 millones de usuarios (solo detrás de *Facebook* y *YouTube*), quienes publican mensajes (llamados tuits), con una longitud máxima de 280 caracteres para compartir lo que piensan, hacen o lo que está sucediendo a su alrededor y, aunque pueden incluir contenido multimedia (imágenes, audio o video), en su mayoría es texto debido a la facilidad de su escritura y envío. De manera adicional, todos los usuarios pueden interactuar entre sí, respondiendo o volviendo a publicar los tuits de otros.

En este trabajo se presenta un estudio sobre la predicción automática del nivel educativo en usuarios de *Twitter* en México a través del análisis del texto de las publicaciones que han hecho en la red social. Su hipótesis central es que algunas de las características textuales extraídas de las publicaciones son adecuadas para identificar si un usuario de *Twitter* en México tiene estudios superiores o no, partiendo del supuesto de que un entorno universitario favorece el dominio de conocimientos y usos lingüísticos específicos (Ramos, 2012).

Para conducir la investigación presentada aquí, se hizo una recopilación de 1 101 241 textos en español correspondientes a 195 personas, sin distinción de género, ubicados en México, quienes fueron seleccionados haciendo un muestreo aleatorio de usuarios activos utilizando la *Application Program Interface (API)* nativa de *Twitter* y fueron cuestionados de forma directa sobre su género y nivel de estudios; los que respondieron, se colocaron en dos clases: nivel superior (positivo) y no superior (negativo).

Después, de los tuits se extrajo una serie de características textuales que representan distintos aspectos del contenido publicado: palabras, menciones (@ o *ats*), etiquetas (# o *hashtags*), ligas (*http links)*, abreviaturas, los vectores de palabras *GloVe* y los promedios de uso de cada característica.

Utilizando los distintivos anteriores, se construyeron modelos basados en aprendizaje automático para la predicción. En específico, se entrenaron y probaron modelos de máquinas de vectores de soporte lineales (SVM, por sus siglas en inglés), clasificadores Naïve Bayes (MNB, por sus siglas en inglés) y árboles de decisión (DT, por sus siglas en inglés). Se experimentó con ellos siguiendo una validación cruzada de 10 partes para tener resultados más consistentes. Su desempeño se midió usando dos métricas populares en la clasificación de textos: macro F1 y el área bajo la curva ROC (AUC).

Las contribuciones de este trabajo son dos: la primera es la presentación de la tarea de la predicción automática del nivel educativo en usuarios de *Twitter* a partir del análisis del contenido textual que publican; la segunda, el estudio del desempeño de diferentes características textuales y modelos de predicción para el análisis.

Los resultados muestran que la labor es compleja, con los mejores valores apenas arriba de 60 % para ambas métricas utilizando abreviaturas y máquinas de vectores de soporte o DT. Debido a esto, se determina que es necesario continuar con la investigación para encontrar otras características o modelos que mejoren el desempeño.

El resto del artículo está organizado de la siguiente forma: primero se presenta una revisión de los trabajos relevantes en la literatura; enseguida, se describe la metodología, incluyendo la descripción del conjunto de datos utilizado y de la experimentación; después, se muestran los resultados obtenidos y, por último, se resumen las conclusiones y se establecen algunas ideas para trabajos futuros.

Revisión de la literatura

En el estudio del problema de perfilado de autor se han desarrollado varios trabajos a lo largo de los años. Entre los más reconocidos se encuentran los presentados en los eventos PAN labs (*Plagiarism Analysis, Authorship Identification and Near Duplica-*

1 https://bit.ly/3f4o29K
2 https://bit.ly/3eXA18W

te Detection), que se han realizado como parte de las conferencias CLEF desde el 2013 (Rangel *et al*., 2013; Rangel *et al*., 2014, 2015, 2016, 2017 y 2018). Las tareas de perfilado que se han propuesto en estos eventos han tratado sobre la predicción de edad, género, idioma nativo y rasgos de personalidad. De entre ellos, las de la edad y el género han sido las tareas más populares. En estos eventos se ha utilizado una serie de conjuntos de datos extraídos de tuits en varios idiomas, siendo el inglés el dominante. De manera adicional, hasta el 2018, estos contenían de forma exclusiva texto; a partir de ese año, se ha incluido el uso de imágenes para reforzar la tarea.

En los trabajos presentados en los eventos de PAN@CLEF se han empleado muchos enfoques para resolver la tarea de perfilado de autor. Lo anterior, incluye el uso de varias características extraídas del texto, como diccionarios de frases; palabras en general, de sentimiento, de opinión o específicas asociadas con un grupo de edad o género; lemas; categorías gramaticales; etiquetas HTML; emoticonos; atributos de segundo orden; vectores singulares o de palabras; entre otros. Además, se exploraron diferentes modelos de predicción basados en aprendizaje automático. Entre los más populares se encuentran las máquinas de vectores de soporte, la regresión logística, los clasificadores bayesianos y las redes neuronales profundas (*Deep Learning*).

Para el análisis de información proveniente de *Twitter*, existen varias tareas que se han estudiado; por ejemplo, el análisis de sentimientos, que consiste en determinar si un tuit, usuario o grupo tiene una opinión positiva, negativa o neutra sobre un tema o tópico o en un momento determinado (Desai, 2018; Vashishtha & Susan, 2019). En la detección de temas de tendencia se intenta clasificar el impacto de uno de discusión actual, catalogándolo como algo disruptivo, popular o rutinario (Indra *et al*., 2019). Por otro lado, la predicción de cambios en el mercado de valores utilizando *Twitter* es un campo de investigación de interés actual. Algunos estudios han concluido que el estado de ánimo público recopilado de esta red social bien puede estar correlacionado con el Índice Dow Jones (Pagolu *et al*., 2016). Por último, en la predicción de resultados políticos se desea predecir al ganador de una contienda electoral con base en la popularidad de los contendientes en *Twitter* (Wang & Gan, 2018).

Para todas estas tareas, se han estudiado varias características textuales, como palabras, *n*-gramas, emoticonos, *hashtags* (Jianqiang *et al*., 2018), *BN*-gramas (Indra *et al*., 2019), diccionarios de sentimientos (Desai, 2018), vectores *GloVe*, patrones de difusión de sentimientos (Wang & Yu, 2018) y estadísticas, por ejemplo, el número de menciones a un individuo o su número de seguidores (Serban *et al*., 2014). De igual forma, se han utilizado diversos modelos basados en aprendizaje automático con enfoques supervisados y no supervisados.

En cuanto a la predicción del nivel educativo de usuarios, los trabajos en la literatura se orientan en la detección del nivel de aprendizaje en entornos educativos virtuales (en los cuales se tiene más control sobre las etapas del proceso de aprendizaje y del vocabulario empleado en el mismo, con menor cantidad de ruido que en una red social abierta), enfocándose en tópicos específicos y siguiendo ciertas taxonomías de clasificación (Echeverría *et al*., 2013; Pincay & Ochoa, 2013). De forma adicional, hay trabajos donde se estudia cómo los estudiantes usan las redes sociales para integrarlas en su proceso de aprendizaje (Tess, 2013).

Metodología

La utilizada en este trabajo está compuesta por tres fases: la adquisición de los datos, su procesamiento y la experimentación. Esta última incluye, a su vez, dos procesos: la construcción de los modelos para la predicción y la evaluación de los mismos.

Adquisición de datos

Para esto, primero se creó un *script* en *Python* que realizó un muestreo aleatorio de más de 500 usuarios activos ubicados en México utilizando la API na-

tiva de *Twitter*. A esta muestra inicial se le preguntó directamente a su cuenta por su género y su nivel educativo, obteniendo respuesta de 260 de ellos. Se les etiquetó en dos clases para nivel educativo: superior (positiva) y no superior (negativa).

Con otro *script* de *Python* se recopilaron los tuits de los usuarios identificados previamente, que fueron recuperados por fecha (del más reciente al más antiguo), con un límite máximo de 10 mil por cada uno. Con algunos usuarios hubo problemas en la captura de la información y se eliminaron, quedando, al final, 195. En total, se recopilaron 1 101 241 mensajes, para un promedio de 5 647.4 tuits por usuario; todos están en español, pero en varios se hacen mezclas de palabras en otros idiomas, sobre todo inglés.

En el cuadro 1 se observa la distribución de los usuarios en nuestro conjunto de datos por nivel educativo y género; se puede ver que tanto la distribución por género como por nivel educativo es similar a la encontrada en *Twitter* en su totalidad. En esta red se estima que 66 % son hombres,[3] mientras que 80 % tienen un nivel educativo universitario[4] (*college degree*). *Twitter* (junto con *LinkedIn*) es reconocida por ser popular entre las personas con alto nivel educativo y de ingresos.[5] Por lo tanto, se considera que nuestro conjunto de datos es representativo de la demografía de *Twitter*.

Procesamiento de los datos

De los tuits recopilados para extraer una serie de características relevantes, primero se limpiaron para remover etiquetas HTML, dejando solo el texto de cada uno; a continuación, se concatenaron todos los mensajes correspondientes a un mismo usuario en una sola cadena larga de texto y se pasó todo el contenido a minúsculas; después, se utilizaron expresiones regulares para extraer del texto seis características superficiales: palabras, *emojis*/emoticonos (*emoticons*), etiquetas (# o *hashtags*), menciones (@ o *ats*), ligas (*http links*) y abreviaturas comunes.

Las palabras se consideraron como una secuencia de letras, más los caracteres _ y -; de estas, se filtraron las vacías (*stopwords*), quedando una lista de 778 que se reunieron de diferentes fuentes en internet. En el caso de las abreviaturas, se recopiló de forma manual una lista de las 130 más utilizadas en *Twitter*. El cuadro 2 presenta una muestra de

3 https://www.omnicoreagency.com/twitter-statistics
4 https://www.pewinternet.org/fact-sheet/social-media/
5 https://blog.hootsuite.com/twitter-demographics/

Cuadro 1

Distribución de usuarios por género y nivel educativo

Nivel educativo	Hombres	Mujeres	Total
Superior	103	58	161 (82.5 %)
No superior	22	12	34 (17.5 %)
Total	**125 (64 %)**	**70 (36 %)**	**195 (100 %)**

Cuadro 2

Ejemplos de abreviaturas comunes encontradas en *Twitter*

Abr.	Significado	Abr.	Significado	Abr.	Significado
xd		*fb*	Facebook	*pq*	¿Por qué?
rt	*Retweet*	*omg*	*Oh my God*	*wtf*	What the fuck?
lol	*Laughing Out Loud*	*alv*	A la verga	*tqm*	Te quiero mucho

ellas con su significado; algunas provienen del inglés, pero su uso se ha difundido a otros idiomas, incluido el español.

Analizando la frecuencia del uso de abreviaturas y palabras para cada clase (superior y no superior), en los cuadros 3 y 4 se muestran las 15 más utilizadas. Las listas están ordenadas por frecuencia, primero por columna y luego por fila. En negritas se marcan las palabras o abreviaturas que son diferentes entre clases. Como puede observarse, los usuarios de ambas presentan ciertas diferencias en su uso.

En el cuadro 5 se muestran los promedios y las desviaciones estándar del número de características por tuit por clase. En promedio, un mensaje tiene poco menos de seis características, siendo las palabras las más abundantes y las abreviaturas, las menos. Los usuarios de la clase superior escriben mensajes más largos que contienen, en su mayoría, más palabras, etiquetas, menciones y ligas, mientras que los de la no superior lo hacen con más *emojis*/emoticonos y abreviaturas.

Una vez extraídas las características, se obtuvieron seis archivos (uno por cada una) con 195 líneas por archivo; cada línea representa las características de un usuario. Todos ellos se dividieron en 10 partes para generar el mismo número de conjuntos de entrenamiento y de prueba para poder aplicar una validación cruzada estratificada, manteniendo la proporción de cada clase en cada parte. En esta validación, los conjuntos de entrenamiento se encuentran formados por nueve partes (193 usuarios) y la sección restante se usa como conjunto de prueba (22 usuarios).

Cuadro 3

Listas de las 15 palabras más comunes por clase

Superior				No superior			
gracias	**bendiciones**	**dios**	mañana	gente	**juego**	años	**espero**
vida	the	**amo**	**feliz**	quiero	vida	mañana	**canal**
gente	años	**saludos**	**méxico**	**mierda**	the	alguien	**veo**
quiero	alguien	**mundo**		gracias	**cosa**	gusta	

Cuadro 4

Listas de las 15 abreviaturas más comunes por clase

Superior					No superior				
xd	**by**	**ptm**	alv	**bb**	xd	lol	**md**	**cm**	**wtf**
pa	omg	dm	**tqm**	**fb**	rt	pa	omg	**pls**	dm
rt	lol	cc	**tkm**	**gpi**	**yt**	**tl**	alv	**oc**	**rip**

Cuadro 5

Promedio del número de características por tuit por clase

	Promedio por tuit						
Nivel educativo	**Palabras**	***Emojis***	**Etiquetas**	**Menciones**	**Ligas**	**Abr.**	**Todas**
Superior	4.97 (4.49)	0.48 (1.63)	0.14 (0.55)	0.17 (0.55)	0.35 (0.51)	0.04 (0.22)	6.15 (4.93)
No superior	4.68 (4.12)	0.55 (1.97)	0.11 (0.48)	0.09 (0.41)	0.24 (0.44)	0.06 (0.28)	5.73 (4.58)
Promedio	4.83 (4.31)	0.52 (1.81)	0.13 (0.52)	0.13 (0.49)	0.30 (0.48)	0.05 (0.25)	5.94 (4.76)

En el cuadro 6 se muestran ejemplos de tuits del conjunto de datos con sus clases, antes de ser procesados y de las características extraídas. Debido a la poca presencia promedio de la mayoría de estas, no hay mensajes que contengan todas al mismo tiempo.

Para las características anteriores, de cada conjunto de entrenamiento se extrajo un vocabulario, formado por el grupo de particularidades únicas en ese conjunto. En el cuadro 7 se muestran los promedios y las desviaciones estándar de los tamaños de cada vocabulario. Las ligas tienen el más extenso, seguidas de las palabras, y las abreviaturas, el menor; esto indica que hay una gran cantidad de ligas y palabras únicas.

Usando el vocabulario correspondiente, se realizó un proceso de vectorización por usuario utilizando el método *term-document-inverse-document-frequency* (*tf-idf*), el cual está definido como $tfidf(t, d) = tf(t, d) \times idf(t)$. En esta ecuación, $tf(t, d)$ es la frecuencia de la característica t (palabra, *emoji*, liga, etc.) en el documento d que, para nuestro caso de estudio, corresponde a un usuario. La parte $idf(t)$ está definida como $idf(t) = log\left(\frac{1+n_d}{1+df(d,t)}\right) + 1$, donde $df(d, t)$ representa el número de usuarios que contienen la característica t, y n_d es la cantidad total de usuarios en el conjunto de entrenamiento. Para cada característica en el conjunto de entrenamiento se calculó el *idf* y se guardó para ser utilizado en la vectorización del conjunto de prueba.

Adicional a las características individuales, también se construyeron matrices con el conjunto combinado de todas ellas. Al utilizar el método *tf-idf*, cada usuario queda representado como un vector, cuyo tamaño corresponde con el tamaño de los vocabularios mostrados en el cuadro 7, dependiendo de la característica utilizada. El método *tf-idf* es común cuando se trabaja con texto para medir la importancia de una característica respecto a un documento (usuario) y a toda la colección de documentos (usuarios).

De igual forma, se construyeron matrices usando los promedios de uso de cada característica por tuit por usuario. De esta forma, un usuario queda representado por un vector de siete promedios (uno para cada una de las seis características y uno para el uso de todas combinadas).

Cuadro 6

Ejemplos de tuits con sus características extraídas

Clase	Tuit	Palabras	*Emojis*	Etiquetas	Menciones	Ligas	Abr.
Superior	Cuando irán a traer Promare a Latinoamérica 😔	irán traer promare latinoamérica	😔				
Superior	No sé, me reí por 10 minutos XD https://bit.ly/3iFL4EU	reí minutos				https://bit.ly/3iFL4EU	Xd
No superior	#Bowsette al estilo megaman rollo 8bits	estilo megaman rollo bits		bowsette			
No superior	Sopa misógina como @DieBatsuDie *https://bit.ly/3faA4Nx*	sopa misógina			diebatsudie	https://bit.ly/3faA4Nx	

Cuadro 7

Tamaños de los vocabularios por característica

Vocabulario						
Palabras	*Emojis*	Etiquetas	Menciones	Ligas	Abr.	Todas
233 607	1 222	32 122	31 356	305 730	108	594 469
(7 261)	(8)	(1 626)	(1 773)	(12 607)	(1)	(20 643)

Por último, se construyeron matrices usando como características los vectores de palabras *GloVe*, cuyo modelo mide estadísticas de coocurrencia entre palabras a partir de un conjunto de datos de entrenamiento. Para este trabajo, se utilizó un modelo preentrenado[6] sobre el conjunto de datos *Spanish Billion Word Corpus* (Cardellino, 2016), el cual contiene un diccionario de más de 800 mil palabras, cada una representada por un vector de 300 particularidades. Para las características *GloVe*, se calculó el vector promedio de todos los vectores de las palabras encontradas en los tuits de cada usuario. De esta forma, cada usuario queda representado por un vector promedio de 300 características densas. Se aplicó el proceso de vectorización para los usuarios en cada conjunto de entrenamiento y prueba.

Una vez obtenidas las matrices de entrenamiento y prueba, se aplicó una normalización euclidiana ($l2$) sobre cada vector de usuario de cada matriz para evitar problemas de escalamiento. La normalización euclidiana se define como $\vec{x}_N = \frac{\vec{x}}{\|\vec{x}\|_2} = \frac{\vec{x}}{\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}}$, donde $\vec{x}_N$ representa el vector $\vec{x}$ (un usuario) normalizado y $\|\vec{x}\|_2$ es la norma $l2$ del vector.

Experimentación

Después de la vectorización, para cada una de las características utilizadas para representar a los usuarios se tenían 10 matrices de entrenamiento y 10 más de prueba normalizadas, siguiendo la validación cruzada estratificada de 10 partes. Con las primeras se construyeron y probaron modelos de predicción usando máquinas de SVM, MNB y DT.

Para optimizar el hiperparámetro de regularización C de los modelos SVM, en cada iteración de la validación cruzada se probaron los siguientes valores de C: 0.1, 1.0, 10.0, 100.0, realizando una subvalidación cruzada de tres partes con cada conjunto de entrenamiento con cada posible valor de C. Una vez ajustado el parámetro, se construyó el modelo final con el mejor valor de C y con todo el conjunto de entrenamiento.

Para medir el desempeño de los modelos, se usaron dos métricas basadas en una *matriz de confusión*, la cual está formada por cuatro celdas (ver cuadro 8) con número de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Esta matriz relaciona las clases reales de los usuarios (columnas) con las predichas por el modelo (renglones). Se construye una matriz por cada clase, donde un usuario puede (positivo) o no (falso) pertenecer a la clase.

La primera métrica utilizada es macro $F1$, definida como $F1 = 2\left(\frac{Precision \cdot Recall}{Precision + Recall}\right)$, donde $Precision = \frac{TP}{TP+FP}$ y $Recall = \frac{TP}{TP+FN}$. $F1$ representan la media armónica de *precision* y *recall*. En su versión macro, se calcula de forma independiente para cada matriz de confusión de cada clase y después se promedia. De esta forma, cada clase tiene el mismo peso en el resultado de la métrica, independientemente del número de usuarios en cada una.

La curva *Receiver Operating Characteristic (ROC)* es una de probabilidad que grafica la razón de verdaderos positivos contra la de falsos positivos en varios umbrales. El área bajo la curva ROC (AUC) evalúa un grado de separabilidad, midiendo la probabilidad de que un modelo clasifique más alto a un usuario en la clase superior que a uno de la no superior, elegidos ambos de forma aleatoria.

6 Disponible en: https://github.com/dccuchile/spanish-word-embeddings

Cuadro 8

Matriz de confusión

	Salida del modelo: verdadera	Salida del modelo: falsa
Valor real: **verdadera**	**TP**	**FN**
Valor real: **falsa**	**FP**	**TN**

La métrica macro $F1$ toma valores entre 0 y 1, donde 0 representa una clasificación totalmente errónea y 1 una perfecta. AUC toma valores entre 0.5 y 1, donde 0.5 significa un clasificador aleatorio y 1 representa al perfecto.

Para efectos de comparación, se consideran tres modelos base: el primero es un MNB que usa palabras con un pesado binario. Este primer modelo crea un vector para cada usuario con 1 si una palabra aparece en el texto del usuario y 0, si no aparece. El segundo es totalitario, el cual asigna a todos los usuarios en el conjunto de prueba a la clase mayoritaria (superior). El tercero es aleatorio uniforme, el cual destina un usuario a una clase con probabilidad de 0.5. Para los modelos 2 y 3, se tienen valores para las métricas de 0.45 para macro $F1$ y 0.5 para AUC.

No se considera la métrica de *accuracy* (definida con la ayuda del cuadro 8 como $\frac{TP+TN}{TP+FP+FN+TN}$), muy popular en clasificación, ya que puede no representar de manera adecuada el desempeño de los modelos para conjuntos con datos no balanceados (como es el caso del presente trabajo). En estas situaciones, por lo general, una (o más) clase(s) domina(n) el conjunto de datos. Esto propicia que el modelo tienda a clasificar a todos los usuarios en la clase dominante, obteniendo un buen resultado en esta métrica. Sin embargo, tal resultado puede ser engañoso, pues los usuarios en la clase minoritaria son ignorados.

Todos los códigos se implementaron en *Python* utilizando las librerías NLTK, *scikit-learn*, *Beautiful Soup* y *emoji*. Los códigos y los datos utilizados para el procesamiento y la experimentación están disponibles en un repositorio en *GitHub*.[7]

Resultados

En el cuadro 9 se presentan los resultados de los experimentos con las características extraídas y los modelos de predicción. En los renglones 3 a 11 se indican las características: palabras, *emojis*/emoticonos, etiquetas, menciones, ligas, abreviaturas, la combinación de todas las anteriores, vectores *GloVe* y el número promedio de características en un tuit por usuario (longitud). Las columnas 2 a 7 indican las métricas separadas por modelo: SVM, MNB y DT. El renglón 12 (sombreado) muestra el promedio de la métrica de todas las características por modelo. Las columnas 8 y 9 (sombreadas) presentan los promedios de las métricas de todos los modelos por característica. Los renglones 13, 14 y 15 ofrecen los resultados de los modelos base: MNB con palabras y pesado binario (base 1), modelo totalitario (base 2) y modelo aleatorio uniforme (base 3). Se observa que el modelo base 1 tiene un desempeño similar a los otros modelos base. Los resultados mostrados son el promedio y la desviación estándar de las 10 iteraciones en la validación cruzada. En negritas se resaltan los valores más altos para cada métrica.

También, se observa que las características de etiquetas, menciones y ligas tienen un desempeño similar a los modelos base, por lo que no tienen poder discriminatorio. Solo las etiquetas con el DT o

7 https://github.com/jcgcarranza/education_twitter

el SVM presentan resultados arriba de los modelos base. Considerando que los vocabularios de estas características son grandes (ver cuadro 7) y que su presencia en los tuits de los usuarios es baja (ver cuadro 5), se estima que ellas están distribuidas sobre los usuarios y no sobre las clases; es decir, que cada usuario introduce nuevas características de manera aislada. Por otro lado, las etiquetas pueden tener mejor desempeño porque se refieren a temas (tendencias), los cuales estarían relacionados con intereses de un determinado nivel educativo.

En el caso de las palabras, a pesar de tener una mayor presencia en los tuits y un vocabulario más alto, tienen un desempeño inferior a los *emojis*, cuya presencia y vocabulario son más limitados. Esto nos indica que hay una mayor superposición de las distribuciones de uso de palabras entre las clases. Por otro lado, los emojis están más presentes en los tuits que las etiquetas, menciones o ligas (ver cuadro 5), por lo que su uso indica, en cierta medida, que las expresiones emocionales pictográficas, además de ser más populares, cambian entre los niveles educativos.

Más interesante es el uso de las abreviaturas, cuyo vocabulario es muy pequeño (en promedio, 108, con desviación estándar de 1) y tienen la menor presencia en los tuits (con solo 0.05 en promedio por mensaje, con desviación estándar de 0.25), pero que presentan el mejor desempeño de todas las características. Esto nos indica que las distribuciones del uso de abreviaturas entre las clases presentan diferencias, donde los usuarios de los dos niveles educativos las utilizan de forma distinta para comunicarse.

Cuadro 9

Resultados de la predicción del nivel educativo utilizando las diferentes características

	SVM		MNB		DT		Promedio	
Caract.	**Macro *F*1**	**AUC**	**Macro *F*1**	**AUC**	**Macro *F*1**	**AUC**	**Macro *F*1**	**AUC**
Palabras	0.52 (0.15)	0.54 (0.11)	0.45 (0.01)	0.50 (0.00)	0.55 (0.13)	0.57 (0.14)	0.51 (0.11)	0.54 (0.10)
Emojis	0.57 (0.16)	0.57 (0.15)	0.45 (0.01)	0.50 (0.00)	0.60 (0.09)	0.60 (0.09)	0.54 (0.10)	0.56 (0.10)
Etiquetas	0.45 (0.02)	0.49 (0.03)	0.45 (0.01)	0.50 (0.00)	0.53 (0.11)	0.53 (0.10)	0.48 (0.06)	0.51 (0.06)
Menciones	0.45 (0.01)	0.50 (0.00)	0.45 (0.01)	0.50 (0.00)	0.48 (0.08)	0.51 (0.07)	0.46 (0.05)	0.50 (0.04)
Ligas	0.45 (0.01)	0.50 (0.00)	0.45 (0.01)	0.50 (0.00)	0.45 (0.01)	0.50 (0.00)	0.45 (0.01)	0.50 (0.00)
Abr.	**0.62 (0.18)**	**0.65 (0.17)**	0.45 (0.01)	0.50 (0.00)	0.60 (0.15)	0.62 (0.15)	0.56 (0.13)	**0.59 (0.13)**
Todas	0.56 (0.15)	0.56 (0.13)	0.45 (0.01)	0.50 (0.00)	0.59 (0.11)	0.61 (0.12)	0.53 (0.11)	0.56 (0.10)
GloVe	0.60 (0.14)	0.61 (0.13)	--	--	0.53 (0.12)	0.53 (0.12)	**0.57 (0.13)**	0.57 (0.12)
Longitudes	0.51 (0.10)	0.54 (0.07)	0.45 (0.01)	0.50 (0.00)	0.57 (0.12)	0.58 (0.12)	0.51 (0.09)	0.54 (0.08)
Promedio	0.53 (0.12)	0.55 (0.10)	0.45 (0.01)	0.50 (0.00)	**0.54 (0.11)**	**0.56 (0.11)**		
Base 1	0.45 (0.01) 0.50 (0.00)							
Base 2	0.45 0.50							
Base 3	0.45 0.50							

Para ilustrar la diferencia en las distribuciones del uso de las abreviaturas, en la gráfica 1 se presenta la frecuencia relativa (dividida entre la frecuencia máxima) de cada una en cada clase. Algunas se utilizan con frecuencia similar (p. ej. *xd*, que es la más común en ambas clases, o + 1, también usada con frecuencia), mientras que otras se usan de manera distinta (p. ej. *amix* se utiliza más en la clase superior y *yt*, en la no superior). Se realizó la prueba X^2 de Pearson para homogeneidad sobre ambas distribuciones, con la hipótesis nula de que ambas son iguales. El resultado para X^2 de la prueba fue de 6 236.93. Comparando este valor con el cuadro de referencia con una significancia de 95 % (p = 0.05) y 129 grados de libertad (valor en tabla de 156.508), se observa que es mayor, por lo que se rechaza la hipótesis nula y podemos decir que ambas distribuciones son distintas.

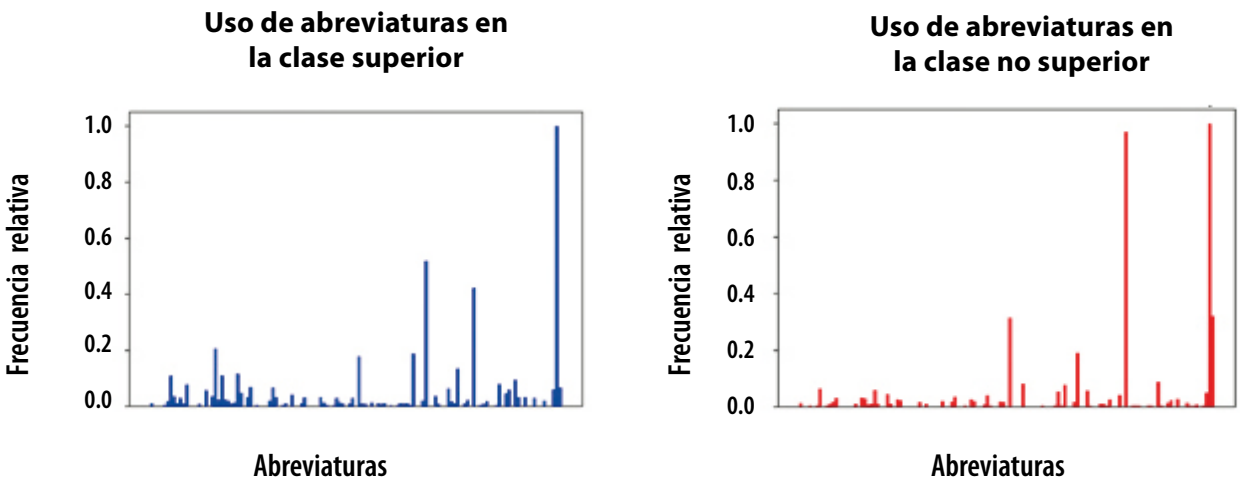
De forma adicional, se aplicó un análisis de componentes principales (PCA) a las matrices de los usuarios representados con características de palabras y abreviaturas. La gráfica 2 muestra el resultado de la proyección de los usuarios de cada clase al espacio de los dos primeros PCA. Se observa que los usuarios representados con palabras tienen una mayor superposición entre clases, lo cual complica la clasificación, mientras que cuando se ha hecho con abreviaturas, se distingue una mayor separación. No obstante, a pesar de tener una mayor separación entre clases con las abreviaturas, los usuarios de la clase no superior están agrupados en áreas diferentes, con lo que la clasificación no es tan directa y, por ello, se observan valores aún bajos para las métricas de clasificación en esta tarea.

Por otro lado, cuando se consideran todas las características anteriores en conjunto (lo que sería equivalente a tomar todo el contenido del tuit), el desempeño es un poco mejor al uso de solo las palabras, pero no al de emplear solo los *emojis* o las abreviaturas. Esto indica que agregar el resto de las características a las abreviaturas o los *emojis* modifica la distribución de los usuarios en los espacios de las características, creando una mayor superposición y limitando la discriminación entre las clases.

En el caso de los vectores de palabras *GloVe*, estos presentan un buen desempeño con el modelo SVM (casi similar a las abreviaturas), pero más bajo con el DT (inferior al uso de palabras). Esto significa que los usuarios representados en el espacio de

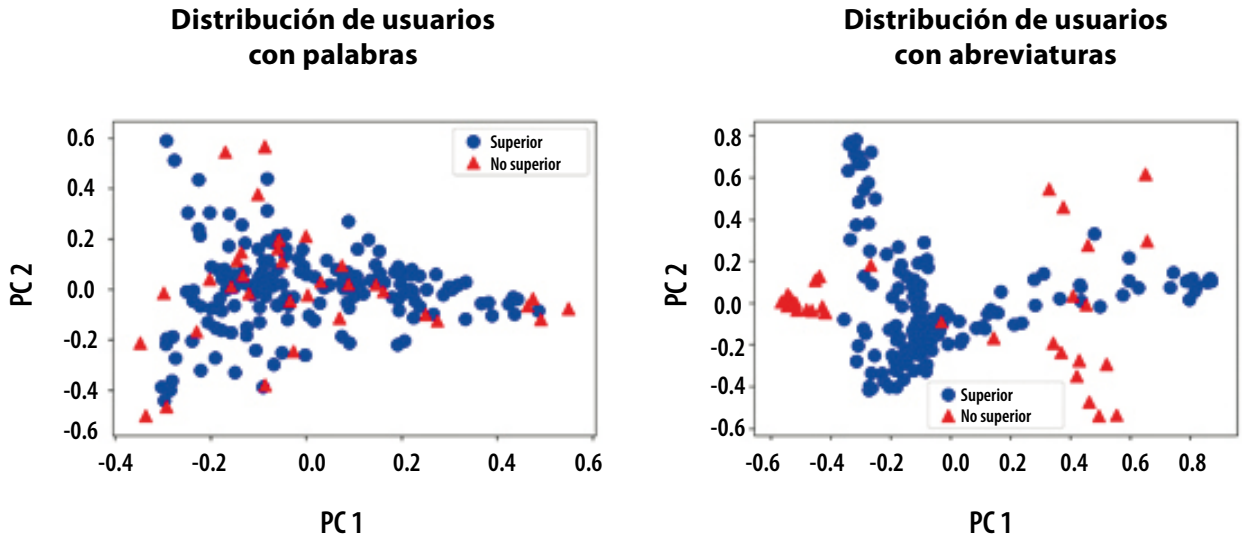
Gráfica 1

Frecuencia de uso de las abreviaturas entre clases



Gráfica 2

Proyección de los usuarios con PCA utilizando palabras y abreviaturas



características de estos vectores pueden ser mejor separados utilizando el hiperplano del modelo SVM, que con las reglas de decisión del DT.

Por último, cuando se utiliza el número promedio de características por tuit, se alcanza un desempeño similar al del uso de las palabras. Esto indica que el tamaño promedio de los mensajes que escribe un usuario refleja de manera parcial una diferencia entre las clases (ver cuadro 5), donde se muestra que los usuarios de la clase superior tienden a escribir tuits un poco más largos que los de la otra clase. No obstante, las diferencias no son suficientes para una discriminación completa.

Respecto a los modelos de predicción utilizados, se observa en el cuadro 9 que el MNB tiene un comportamiento igual al de los modelos base, por lo que las probabilidades condicionales de las características son similares entre clases y el modelo no puede discriminar entre ellas. Por otro lado, cuando se promedian sus resultados sobre todas las características, los modelos SVM y DT tienen un comportamiento similar, con ambos presentando su mejor desempeño con el uso de abreviaturas. El primero también tiene buen comportamiento con los vectores de palabras *GloVe*, mientras que el segundo lo tiene con los *emojis*. En general, el mejor desempeño se alcanza con el uso de SVM y abreviaturas. Esto indica que los usuarios representados en el espacio de las abreviaturas pueden separarse de una mejor manera utilizando un hiperplano.

Conclusiones y trabajo futuro

En este trabajo se presentó un estudio sobre la predicción automática del nivel educativo en usuarios de *Twitter* en México. Se extrajo una serie de características del contenido textual de los tuits publicados por los usuarios y se utilizaron para construir tres tipos de modelos basados en aprendizaje automático: SVM, MNB y DT. La predicción fue de tipo binaria, donde los modelos determinaban si un usuario tenía estudios universitarios o no. Se probaron las características y los modelos con un conjunto de datos extraído de forma directa del sitio, compuesto por más de un millón de tuits en español, correspondientes a 195 usuarios ubicados en México.

De acuerdo con los experimentos para la predicción automática del nivel educativo de usuarios de *Twitter* en México, se concluye lo siguiente:

- Algunas características individuales, como las ligas, menciones y etiquetas, son demasiado específicas (haciendo referencia a sitios en la red, usuarios o temas particulares) y no aportan información suficiente para distinguir entre las clases.
- Otras, como las palabras, aportan un mejor contexto de los usuarios sobre temas de interés propios de la clase, lo cual ayuda a una mejor discriminación que las características previas.
- El uso de *emojis* es más popular que el de ligas, etiquetas o menciones. Además, estas expresiones emocionales pictográficas cambian entre los niveles educativos y pueden ayudar a separar entre clases con el empleo de reglas de decisión.
- Las abreviaturas presentan el mejor desempeño para separar entre usuarios de los dos niveles educativos. Esto indica que los usuarios de distintas clases usan las abreviaturas de formas diferentes.
- La combinación de todas las características introduce ruido, generando una mayor superposición entre las distribuciones de uso de estas y en la representación de los usuarios en el espacio de características.
- Los vectores de palabras *GloVe* capturan el contexto de cada palabra, lo que permite representar a los usuarios en un espacio donde pueden ser separados de forma parcial con un hiperplano.
- El número promedio de características en los tuits de un usuario es un indicador parcial de las clases, considerando que los de la clase superior tienden a escribir mensajes más largos.

Algunas ideas por explorar para trabajos futuros incluyen el uso de otros modelos de clasificación, como las redes neuronales, las cuales pueden funcionar de manera adecuada con características densas, como los vectores de palabras *GloVe*. De igual forma, podemos considerar el uso de otros tipos de vectores de palabras, como *FastText*, que consideran de manera diferente el contexto de estas. Por último, se puede estudiar el uso de métodos de reducción de dimensión, como *Latent Dirichlet Allocation*, *Latent Semantic Indexing*, *Principal Component Analysis*, *Biased Discriminant Analysis* y *Non Negative Matrix Factorization*, entre otros, que se encargan de calcular asociaciones entre palabras para agruparlas en tópicos o temas.

Fuentes

Cardellino, C. *Spanish Billion Words Corpus and Embeddings*. 2016.

Desai, R. D. "Sentiment Analysis of Twitter Data", en: *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. Madurai, India, IEEE, 2018, pp. 114-117.

Echeverría, V.; J. C. Gómez & M. F. Moens. "Automatic labeling of forums using bloom's taxonomy", en: *International Conference on Advanced Data Mining and Applications*. Springer, 2013, pp. 517-528.

Kumar, S.; F. Morstatter & H. Liu. *Twitter Data Analytics*. New York, Springer Publishing Company, 2014.

Özgüven, N. & B. Mucan. "The relationship between personality traits and social media use", en: *Social Behavior and Personality: an International Journal*. 2013, vol. 41, no. 3, pp. 517-528.

Pagolu, V. S.; K. N. Reddy; G. Panda & B. Majhi. "Sentiment analysis of Twitter data for predicting stock market movements", en: *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*. IEEE, 2016, pp. 1345-1350.

Pal, N. R.; K. Pal; J. M. Keller & J. C. Bezdek. "A possibilistic fuzzy c-means clustering algorithm" en: *IEEE Transactions on Fuzzy Systems*. 2005, vol. 13, no. 4, pp. 517-530.

Pincay, J. & X. Ochoa. "Automatic classification of answers to discussion forums according to the cognitive domain of Bloom's taxonomy using text mining and a Bayesian classifier", en: *EdMedia+ Innovate Learning*. Association for the Advancement of Computing in Education (AACE), 2013, pp. 626-634.

Ramos, J. E. "Bases para la evaluación del dominio de las formas disciplinares de comunicación y de los usos lingüísticos especializados en el Espacio Europeo de Educación Superior (EEES)", en: *Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas*. Núm. 12, 2012, pp. 88-119.

Rangel Pardo, F. M.; F. Celli; P. Rosso; M. Potthast; B. Stein & W. Daelemans. "Overview of the 3rd Author Profiling Task at PAN 2015", en: *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*. 2015, pp. 1-8.

Rangel, F.; P. Rosso; I. Chugur; M. Potthast; M. Trenkmann; B. Stein (...) & W. Daelemans. "Overview of the 2nd author profiling task at PAN 2014", en: *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers*. UK, Sheffield, 2014, pp. 1-30.

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

Rangel, F.; P. Rosso; M. Montes-y-Gómez; M. Potthast & B. Stein. "Overview of the 6th author profiling task at PAN 2018: multimodal gender identification in Twitter", en: *Working Notes Papers of the CLEF*. 2018.

Rangel, F.; P. Rosso; M. Potthast & B. Stein. "Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in twitter", en: *Working Notes Papers of the CLEF*. 2017.

Rangel, F.; P. Rosso; B. Verhoeven; W. Daelemans; M. Potthast & B. Stein. "Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations", en: *Working Notes Papers of the CLEF 2016 Evaluation Labs*. CEUR Workshop Proceedings/Balog, Krisztian (edit.) *et al*. 2016, pp. 750-784.

Serban, I. V.; D. S. González & X. Wu. *Prediction of changes in the stock market using twitter and sentiment analysis*. 2014.

Tess, P. A. "The role of social media in higher education classes (real and virtual)-A literature review", en: *Computers in human behavior*. Vol. 29, no. 5, 2013, pp. A60-A68.

Vashishtha, S. & S. Susan. "Fuzzy rule based unsupervised sentiment analysis from social media posts", en: *Expert Systems with Applications*. Vol. 138, 2019, p. 112834.

Wang, J.; H. Zhao & Z. Liu. "Topic Propagation Prediction Based on Dynamic Probability Model", en: *IEEE Access*. Vol. 7, 2019, pp. 58685-58694.

Wang, L. & J. Q. Gan. "Prediction of the 2017 French Election Based on Twitter Data Analysis Using Term Weighting", en: *2018 10th Computer Science and Electronic Engineering (CEEC)*. IEEE, 2018, pp. 231-235.

Wang, L.; J. Niu & S. Yu. "SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis", en: *IEEE Transactions on Knowledge and Data Engineering*. 2019.

Wasserman, S. & K. Faust. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge, Cambridge University Press, 1994.

Winarko, E. & R. Pulungan. "Trending topics detection of Indonesian tweets using BN-grams and Doc-p", en: *Journal of King Saud University-Computer and Information Sciences*. Vol 31, no. 2, 2019, pp. 266-274.

Wisdom, V. & R. Gupta. *An introduction to Twitter Data Analysis in Python*. 2016.

Zhao, D. & M. B. Rosson. "How and why people Twitter: the role that micro-blogging plays in informal communication at work", en: *Proceedings of the ACM 2009 International Conference on Supporting group work*. New York, NY, USA, ACM, 2009, pp. 243-252.