ISSN 1870-4069

Minería de opiniones centrada en tópicos usando textos cortos en español

José A. Reyes-Ortiz, Fabián Paniagua-Reyes, Leonardo Sánchez

Universidad Autónoma Metropolitana, Departamento de Sistemas, Ciudad de México, México

{jaro, ldsm}@correo.azc.uam.mx, al2112002241@alumnos.azc.uam.mx

Resumen. Los usuarios expresan sus sentimientos sobre una entidad de un tema específico de manera libre utilizando textos cortos en las redes sociales. El análisis de sentimientos, también conocido como minería de opiniones, se enfoca en examinar estos textos para determinar su polaridad. Este artículo presenta un enfoque para la minería de opiniones basada en tópicos a partir textos de Twitter en español. El objetivo principal es decidir la polaridad de un texto, determinando si el contenido tiene implicaciones positivas, negativas o neutras en la reputación de entidades para un tópico. En este trabajo se utiliza un enfoque supervisado para la clasificación de textos utilizando el modelo bolsa de palabras para la representación de características. La experimentación muestra resultados prometedores, aportando un recurso para el análisis de textos en español.

Palabras clave: Minería de opiniones, identificación de polaridad, análisis de sentimientos, recursos lingüísticos para el español.

Mining of Opinions Centered on Topics Using Short Texts in Spanish

Abstract. Users express their feelings about an entity of a specific topic in a free way using short texts on social networks. Sentiment analysis, also known as opinion mining, focuses on examining these texts to determine their polarity. This article presents an approach to the mining of opinions based on topics from Twitter texts in Spanish. The main objective is to decide the polarity of a text, determining if the content has positive, negative or neutral implications in the reputation of entities for a topic. In this paper a supervised approach is used for the classification of texts using the word bag model for the representation of characteristics. The experimentation shows promising results, providing a resource for the analysis of texts in Spanish.

Keywords: Mining of opinions, identification of polarity, analysis of feelings, linguistic resources for Spanish.

1. Introducción

La red social denominada Twitter es utilizada por diversos usuarios para expresar sus sentimientos sobre un tema, producto, servicio o entidad. Diariamente, miles de textos cortos, no mayores a 140 caracteres, son generados en esta red social. Esto representa una gran cantidad de información que no pueden ser procesadas de manera manual ya que involucraría una tarea costosa y que consume mucho tiempo. Es posible tener acceso a esta información y aplicar técnicas de minería de datos, aprendizaje automático y procesamiento de lenguaje natural para descubrir conocimiento útil. Este conocimiento puede ser acerca de la reputación de una marca, el nivel de aceptación sobre un producto, el sentimiento generado por un acto o evento donde participa un personaje famoso.

Existe una necesidad de contar con herramientas para el análisis de textos cortos, específicamente, para la minería de opiniones. La necesidad se incrementa en los recursos para el análisis de textos en español ya que existe muy poca investigación en este rubro. Con estas herramientas se pueden tomar decisiones oportunas y rápidas, por ejemplo, conocer hacia dónde se debe enfocar un nuevo producto o servicio, mejorar una marca de automóviles o corregir algún aspecto sobre un personaje famoso.

Se obtendrían resultados en tiempo real y las decisiones serían tomadas con datos frescos, es decir, producidos en ese momento o instante. Este es una ventaja de utilizar datos generados en las redes sociales.

Es por ello que, este artículo se centra en aportar un recurso de minería de opiniones utilizando datos de la red social Twitter, para disminuir la carencia de recursos para el idioma español. Estos datos son textos cortos, conocidos como *tweets.* En este trabajo se utiliza un léxico o diccionario obtenido a partir de un corpus mediante aprendizaje automático y utilizando modelos de clasificación. Los textos son clasificados por el tipo de polaridad (positiva, negativa o neutra) que contiene en mensaje en un tópico determinado, a saber: automóviles, bancario y artistas/músicos. Por lo tanto, el enfoque presentado en este artículo es de gran utilidad para la minería de opiniones sobre un tópico ya que identifica la polaridad de un mensaje utilizando una clasificación de textos. Por ejemplo, es posible determinar la polaridad general del tema de automóviles a partir del análisis de los textos generados por usuarios en la red social Twitter.

El resto de este artículo está organizado como sigue. En la Sección 2 se presentan los trabajos relacionados con la minería de opiniones con textos de redes sociales. La Sección 3 presenta el enfoque utilizado para la minería de opiniones que determina la polaridad de un mensaje centrado en un tópico. El conjunto de datos utilizado para el aprendizaje del lexicón y para la experimentación se describe en la Sección 4. Por su parte, la Sección 5, muestra los resultados de la experimentación con cuatro algoritmos (Naïve Bayes, *k*-vecino más cercano, máquinas de soporte vectorial y algoritmo basado en árboles de decisión). Finalmente, las conclusiones y el trabajo a futuro son presentados en la Sección 6.

2. Trabajo relacionado

El análisis de sentimientos y la minería de opiniones con datos de redes sociales han sido temas con gran interés en los últimos años. En estas tareas se encuentra la clasificación de opiniones en Twitter, la cual consiste en determinar la polaridad expresada por un texto corto, es decir, determinar la carga positiva o negativa que contiene dicho texto. En este contexto, la minería de opiniones con datos de redes sociales ha sido abordado desde diversas perspectivas en años recientes, como las que se describen a continuación.

Enfoques supervisados basados en análisis estadísticos han sido utilizados en [1], donde los autores utilizan los medios de comunicación social, específicamente, Twitter y Facebook para el análisis de polaridad sobre personas, entidades y marcas. Ellos presentan un marco de trabajo organizado por módulos, en el cual se puede experimentar con diversos clasificadores como Naïve Bayes, máquinas de soporte vectorial, árboles de decisión y *k*-vecino más cercano; donde el análisis de polaridad se centra en clasificar textos en tres categorías: positivo, negativo y neutro. En [2] se presenta un enfoque híbrido para la clasificación de textos extraídos de Twitter. El enfoque presentado considera la minería de opiniones utilizando una lista de símbolos y un análisis con *SentiWordNet*, además usan la frecuencia de palabras negativas y positivas; por último, la clasificación de los textos se centra en tres categorías: positivos, negativos y neutros. También, en [3] presentan un algoritmo de aprendizaje automático para clasificar la polaridad de los mensajes en español; los autores realizan un estudio de diferentes características como mensajes reenviados, menciones, ligas y etiquetas con el símbolo '#', los autores generan un corpus en español llamado COST y experimentaron con algoritmos de aprendizaje supervisado como Naïve Bayes (NB), máquinas de soporte vectorial y el algoritmo de Regresión Logística (LR).

Otro mecanismo de apoyo o soporte que se ha aplicado en la minería de opiniones son las ontologías. Como en [4] que utilizan una técnica basada en ontologías para la minería de mensajes generados en Twitter. La novedad del enfoque propuesto es que los mensajes se caracterizan con grados de sentimiento distintos por cada tema existente en el mensaje. Esto genera un análisis más detallado de las opiniones de los mensajes sobre un tema específico.

El uso de *n-gramas* y el modelo denominado bolsa de palabras se expone en los siguientes trabajos. En [5] se presenta un enfoque basado en una clasificación supervisada de textos utilizando unigramas, bigramas y trigramas con un análisis de frecuencias; el análisis estadístico genera un léxico específico y reducido para Twitter, el cual es utilizado para el análisis de sentimientos y está compuesto por 187 características que reduce la complejidad del modelo, mientras mantiene un alto grado de cobertura del corpus y, además, produce una mejor precisión para la clasificación de sentimientos.

En análisis de sentimientos para textos entre diferentes lenguajes ha sido propuesto por [6], quienes proponen un enfoque simple para múltiples lenguajes, basado en el modelo espacio vectorial, empleando características que pueden ser utilizadas entre tres lenguajes (español, inglés e italiano) y características independientes del lenguaje, entre las que destacan: operadores de negación, si es una palabra derivada o no, entre otros.

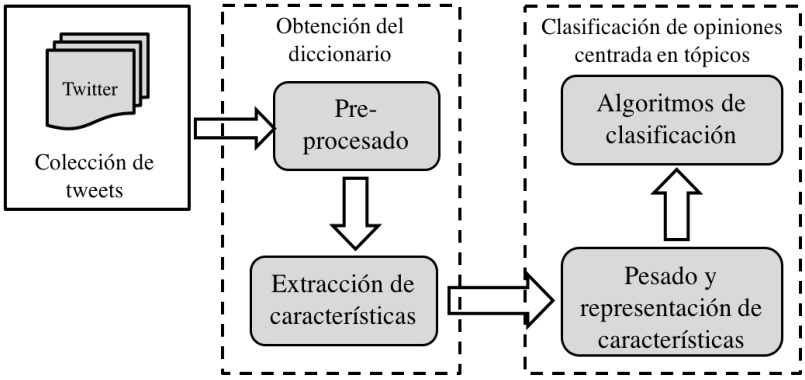

Fig. 1. Arquitectura del enfoque para la minería de opiniones.

Finalmente, en algunos casos, lexicones o diccionarios son combinados en un enfoque basado en aprendizaje supervisado. Estos enfoques aportan una solución al problema de la clasificación de sentimientos u opiniones expresadas en mensajes de Twitter o Facebook; como se presenta en [7], donde se propone un método que adopta un lexicón para llevar a cabo el análisis de sentimientos a nivel de entidades, logrando una alta precisión pero con una baja sensibilidad; en [8] se propone un método basado en un lexicón para el análisis de sentimientos utilizando mensajes de Facebook en vietnamita, los autores construyen un diccionario de emociones para el vietnamita incluyendo cinco sub-diccionarios: sustantivo, verbo, adjetivo, adverbio y un diccionario adaptado del inglés; en [9] presentan los pasos detallados para la construcción de los componentes que componen un enfoque basado en lexicones para el análisis de sentimientos.

La mayoría de trabajos del estado del arte se centran en el idioma inglés. Sin embargo, existe una necesidad de recursos para la minería de opiniones en español. Se ha detectado una carencia de enfoques para el español con las perspectivas descritas anteriormente. Con respecto a esto, este trabajo centra su aportación en reducir esta falta de recursos para la minería de opiniones, presentando un enfoque supervisado para la identificación de polaridad centrada en tópicos utilizando textos generadas en Twitter en idioma español.

La minería de opiniones consiste en determinar si un mensaje expresa una polaridad positiva, negativa o neutra sobre un tópico específico, tales como: automóviles, bancos y artistas/músicos.

3. Minería de opiniones centrada en tópicos

En esta sección se presentan los componentes del enfoque para la minería de opiniones, es especifico, la identificación de una polaridad mediante una clasificación de textos cortos centrada en tópicos para el español. Estos componentes se observan en la Figura 1, los cuales incluyen diversas tareas como el pre-procesado de los textos, la

Tabla 1. Normalización de la risa.

Patrón	Frase	Risa normalizada
(ja)+	ja	jaja
(je)+	jeje	jaja
(jo)+	jojojo	jaja
(ji)+	jijijiji	jaja
lol	lol	jaja

Tabla 2. Forma enraizada de palabras para el español.

Palabra	Raíz
buenos	buen
rebajarán	rebaj
fueron	ir
llegaría	lleg

extracción de características o la obtención del lexicón/diccionario, la ponderación de las características considerando la frecuencia de aparición de los términos y finalmente, los algoritmos de clasificación supervisada que se centran en tres tópicos: automóviles, bancos y música.

3.1. Pre-procesado de textos

La primera tarea para la obtención de un diccionario o lexicón de palabras, es la limpieza de los textos, para ello se realiza una segmentación por palabras (*tokens*) y la eliminación de caracteres especiales. Después, las unidades léxicas son filtradas eliminando las ligas (*url*) a sitios web externos, menciones de usuarios en Twitter (@), entidades nombradas para cada dominio, ya sea directamente o mediante *hashtag* (#). Por ejemplo, para el dominio automóviles fueron eliminadas menciones a entidades como: #bmw, #volvo, #ferrari, entre otras.

También, en esta fase se lleva a cabo una normalización de las unidades léxicas resultantes a minúsculas y se eliminan las *stopwords*, palabras que no aportan significado y por lo tanto, no son funcionales para la identificación de polaridad. Esta lista de palabras contiene artículos (un, la, los), preposiciones (a, con, de, para), verbos no funcionales (ser, estar), entre otros. Sin embargo, se descartan de esta lista palabras de negación (no, ni) o afirmación (si), al ser consideradas como funcionales para la identificación de la polaridad manifestada por un mensaje (*tweet*).

Para evitar redundancia en la forma de expresar una risa en los textos, en esta etapa del pre-procesado, se considera un paso de normalización. Para el cual, se aplican reglas y se sustituyen las diversas formas de expresar risa por un término en común, como se muestra en la Tabla 1, donde el símbolo (+) significa una o más ocurrencias, es decir que una secuencia, por ejemplo "*ja*", se puede repetir una o más veces en los textos.

Finalmente, en esta fase, para cada palabra del léxico obtenido hasta este punto, se lleva a cabo un proceso *Stemming*, el cual consiste en reducir una palabra a su raíz, es

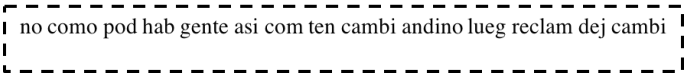

Fig. 2. Algunas palabras del diccionario (bolsa de palabra) como características.

decir, eliminar los sufijos o flexiones de las palabras. Esto permite agrupar todas las palabras con la misma raíz en una sola representación en el lexicón. Para esta tarea se utiliza el, bien conocido, algoritmo de *Porter – Snowball Stemmer* [10, 11], el cual tiene soporte para el español. En la Tabla 2, se muestra un ejemplo de palabras en español como aparecen en los textos y con su forma derivada (raíz) que es generada por el algoritmo.

3.2. Extracción de características

El diccionario o lexicón es obtenido como las unidades léxicas de todos los textos sin repeticiones. Este lexicón constituye las características para la etapa de clasificación y se utiliza la representación mediante bolsa de palabras (*bag-of-words*).

Con esta tarea de extracción se obtiene un diccionario de palabras normalizado y reducido, el cual será utilizado para representar cada instancia (*tweet*) mediante una ponderación, para su posterior clasificación en positivo, negativo o neutro. Un total de 6037 palabras fueron obtenidas a partir del conjunto de datos (corpus), las cuales conforman el vocabulario o diccionario.

La Figura 2 muestra una lista de algunas palabras del diccionario extraído, el cual funciona como el conjunto de características para la tarea denominada identificación de polaridad mediante clasificación textual. Las características mostradas en la Figura 2 son ponderadas con la métrica denominada "Frecuencia de los términos- Frecuencia inversa en los documentos (TF-IDF)", la cual se describe en la Sección 3.3.

3.3. Ponderación de las características

En este trabajo, confiamos en una ponderación de las características basada en la importancia de los términos en un mensaje (*tweet*) enfocada en tópicos. De manera específica, la ponderación de los términos con la finalidad de clasificar los textos está centrada en tres tópicos: automóviles, bancos, y artistas/música.

Existen diferentes enfoques para obtener la importancia o ponderación de los términos del vocabulario sobre un texto corto. Este vocabulario es representado mediante el modelo espacio vectorial con la representación llamada bolsa de palabra (*BoW* por sus siglas en inglés) [12], el cual consiste en una colección de textos y su vocabulario de términos (características). Cada *tweet* es representado como un vector $S_j = (w_{1j}, w_{2j} \dots w_{nj})$, donde cada componente w_{ij} expresa la importancia que produce la característica i, palabra del vocabulario, en el mensaje j.

Para el pesado de los términos o palabras, es decir, determinar su importancia en un *tweet*, se utiliza el pesado basado en la frecuencia de aparición del término dentro de una colección de textos para un tópico (TF-IDF).

Esta ponderación utiliza la frecuencia de aparición para los términos del vocabulario en un texto, la cual consiste en el número de veces que un término (*t*) del vocabulario aparece en un *tweet* (*S*), ver Ecuación 1, y la frecuencia inversa que determina si el término es común en la colección (Ecuación 2). Esta información se utiliza, entonces, para calcular el valor final de *TF-IDF* (Ecuación 3):

$$TF(t_i, S_j) = \mathrm{f}\,(t_i, S_j), \tag{1}$$

$$IDF(t_i, S_j) = \log \frac{|S|}{1 + |s \in S \,:\, t_i \in s|}, \tag{2}$$

$$w_{ij} = TF(t_i, S_j) \times IDF(t_i, S_j). \tag{3}$$

3.4. Identificación de polaridad centrada en tópicos

La identificación de una polaridad, se centra en una tarea típica de clasificación supervisada, la cual se basa en el vector ponderado de los términos del vocabulario con respecto a cada mensaje. La ponderación de los términos se centra en la métrica *TF-IDF* presentada previamente y la clasificación en un enfoque supervisado. Estos vectores son la entrada para la etapa de clasificación, la cual ha sido desempeñada con diversos algoritmos para decidir la polaridad de los textos.

El objetivo de esta fase es construir un clasificador textual capaz de predecir la polaridad de un mensaje en tres posibles categorías: positivo, negativo y neutro. Para ello, es necesario dividir los textos en dos subconjuntos: entrenamiento y prueba.

Como ya se ha mencionado anteriormente, el clasificador está centrado en tópicos, es decir, se obtiene un mismo vocabulario para el todos los mensajes. Sin embargo, la ponderación y clasificación está focalizada en tres tópicos, a saber: automóviles, bancos, y artistas/música.

La tarea de clasificación centrada en los tópicos se lleva a cabo mediante cuatro algoritmos: el clasificador Naïve-Bayes (NB) que se basa en el teorema de Bayes y su función es encontrar la hipótesis más probable que describa los vectores que representan a los textos de prueba, con esto obtiene la probabilidad para que dado los valores que describen a un *tweet*, éste pertenezca a una clase dada [13]; las máquinas de soporte vectorial (SVM) [14] que construyen un conjunto de hiperplanos en un espacio n-dimensional con los textos de entrenamiento, estos hiperplanos son utilizados para predecir la clase de los nuevos textos; algoritmo basado en árboles de decisión (C4.5) es un algoritmo que realiza la inducción a partir de ejemplos preclasificados generando un árbol de decisión con los datos, mediante particiones realizadas recursivamente [15]; y, el algoritmo del k-vecino más cercano (kNN) [16] que estima la función de densidad para los pares a predecir por cada clase basándose en el conjunto de entrenamientos y prototipos.

La idea es evaluar la tarea de clasificación, combinando los cuatro algoritmos (NB, SVM, C4.5 y kNN) con la ponderación de los términos (TF-IDF), para encontrar la mejor solución en cuanto a precisión y cobertura. La implementación de los algoritmos de clasificación se ha llevado a cabo mediante la herramienta WEKA [17].

Tabla 3. Distribución de los textos por tópicos.

Tópico	Textos de entrenamiento	Textos de prueba	Total
Automóviles	1388	716	**2104**
Bancario	1215	626	**1841**
Artistas/Músicos	1993	1027	**3020**
Total	**4596**	**2369**	**6965**

Tabla 4. Resultados para la identificación de polaridad centrada en tópicos utilizando el algoritmo Naïve Bayes.

Tópico	Automóviles			Bancos			Artistas/Músicos		
	P	R	F	P	R	F	P	R	F
Positiva	0.51	0.59	0.55	0.69	0.74	0.72	0.68	0.60	0.64
Negativa	0.58	0.56	0.57	0.68	0.66	0.67	0.67	0.63	0.65
Neutra	0.51	0.54	0.52	0.61	0.70	0.66	0.52	0.63	0.57

Tabla 5. Resultados para la identificación de polaridad utilizando el algoritmo C4.5 (basado en árboles de decisión).

Tópico	Automóviles			Bancos			Artistas/Músicos		
	P	R	F	P	R	F	P	R	F
Positiva	0.69	0.80	0.74	0.77	0.79	0.78	0.77	0.70	0.74
Negativa	0.78	0.72	0.75	0.83	0.82	0.82	0.79	0.83	0.81
Neutra	0.69	0.63	0.66	0.69	0.77	0.73	0.73	0.74	0.73

4. Conjunto de datos

La evaluación del enfoque para la identificación de polaridad centrada en tópicos fue realizada con el conjunto de datos proporcionado por la competencia RepLab [18] para la tarea específica denominada "polaridad de la reputación", cuyo objetivo es decidir si el contenido de un mensaje (*tweet*) en español tiene implicaciones positivas o negativas para la reputación de una entidad, tal como: marca automotriz, entidad financiera, institución educativa o persona famosa en la música.

El conjunto original de datos consta de cuatro tópicos: automóviles, bancario, universidades y artistas/músicos. Sin embargo, en este artículo solo se toma el conjunto de datos de tres tópicos, debido a que el conjunto para el tópico universidades no está balanceado con respecto a los otros tópicos, ya que contiene únicamente 223 mensajes efectivos a diferencia de 3020 para el tópico artistas.

A partir del conjunto de datos seleccionado, se obtuvieron 6965 textos efectivos, para los cuales fue posible obtener su contenido de Twitter y que, además, estaban clasificados manualmente con su etiqueta o categoría para la polaridad (Positivo, Negativo, Neutro). Este conjunto de datos representa un excelente marco de referencia

Tabla 6. Resultados de la experimentación utilizando el algoritmo kNN (*k*-vecino más cercano).

Tópico	Automóviles			Bancos			Artistas/Músicos		
	P	R	F	P	R	F	P	R	F
Positiva	0.54	0.66	0.60	0.75	0.87	0.81	0.61	0.73	0.67
Negativa	0.47	0.51	0.49	0.85	0.73	0.79	0.63	0.67	0.65
Neutra	0.39	0.49	0.44	0.74	0.62	0.68	0.47	0.55	0.51

para la evaluación de algoritmos de minería de opiniones. El conjunto de datos finales se divide en tres tópicos, que a su vez son divididos en dos conjuntos, 66% para el entrenamiento y 34 % para la evaluación, quedando distribuido de la forma como se muestra en la Tabla 3.

5. Experimentación y resultados

La experimentación consiste en utilizar cada uno de los algoritmos con la ponderación TF-IDF, utilizando el mismo conjunto de datos centrado en cada tópico para el entrenamiento y pruebas.

La evaluación de todos los experimentos se realiza utilizando las métricas *Precisión (P)*, *Cobertura (R)* y *medida F*, las cuales han sido ampliamente utilizadas en cualquier tarea de clasificación textual. Estas métricas comparan los resultados del clasificador a evaluar con los valores externos de confianza (texto preclasificado), utilizando los siguientes valores: a) *Verdadero Positivo (VP)* es el número de predicciones correctas del clasificador que corresponden al juicio externo de confianza (texto preclasificado); *Verdadero Negativo (VN)* es el número de predicciones correctas del clasificador de opiniones que no corresponden al juicio externo de confianza; *Falso Positivo (FP)* corresponde al número predicciones incorrectas del clasificador que corresponden al juicio externo de confianza; y, finalmente *Falso Negativo (FN)* es el número de predicciones incorrectas del clasificador que no corresponden al juicio externo de confianza.

Bajo estos criterios, se emplea la *Precisión (P)* para evaluar los algoritmos en términos de los valores de predicciones positivas, la cual se define, en la Ecuación 4, como:

$$P = \frac{VP}{VP + FP}. \tag{4}$$

También, se utiliza el *Cobertura (R)* para expresar la tasa de correspondencias correctas con las opiniones de textos preclasificados de manera externa con una alta confianza (Ecuación 5):

$$R = \frac{VP}{VP + FN}. \tag{5}$$

Tabla 7. Resultados para la identificación de polaridad utilizando SVM (máquinas de soporte vectorial).

Tópico	**Automóviles**			**Bancos**			**Artistas/Músicos**		
	P	R	F	P	R	F	P	R	F
Positiva	0.79	0.83	0.81	0.88	0.86	0.87	0.84	0.86	0.85
Negativa	0.81	0.83	0.82	0.87	0.87	0.87	0.87	0.85	0.86
Neutra	0.79	0.80	0.79	0.81	0.89	0.85	0.84	0.83	0.83

Tabla 8. Resumen de resultados de los algoritmos por tópico

Tópico	Automóviles		Bancos		Artistas/Música		Promedio	
Algoritmo	C	I	C	I	C	I	C	I
NB	54.9	45.1	68.3	31.7	63.2	36.8	62.1	37.9
C4.5	71.5	28.5	77.8	22.2	76.1	23.9	75.2	4.8
kNN	55.3	44.7	74.2	25.8	65.4	34.6	64.9	35.1
SVM	82.3	17.7	87.3	12.7	84.6	15.4	84.7	15.3

Finalmente, la *medida F* que representa la media armónica entre *Precisión* y *Cobertura*, la cual tiene como fundamento el obtener un valor único ponderado entre ellas (Ecuación 6):

$$medida\ F = 2 * \frac{P * R}{P + R}. \tag{6}$$

Todos los experimentos utilizan la ponderación de palabras mediante la métrica TF-IDF. Por su parte, los resultados de esta experimentación están centrados en los tres tópicos mencionados y se analizan por cada categoría (Positiva, Negativa, Neutra) para cada algoritmo de clasificación.

La Tabla 4 muestra los resultados, por tópico, de los experimentos utilizando el algoritmo Naïve Bayes.

La Tabla 5 muestra los resultados de los experimentos utilizando el algoritmo C4.5 (basado en árboles de decisión) por tópico.

La Tabla 6 muestra los resultados de los experimentos utilizando el algoritmo kNN (*k*-vecino más cercano) por tópico.

La Tabla 7 muestra los resultados de los experimentos utilizando el algoritmo SVM (máquinas de soporte vectorial) por tópico.

Finalmente, la Tabla 8 hace un resumen de los resultados para cada algoritmo por tópico y en términos prácticos para el análisis: porcentaje de instancias clasificadas correctamente (C) para las tres categorías y el porcentaje de instancias clasificadas incorrectamente (I).

La Tabla 8, que representa un resumen de los resultados, hace notar que el algoritmo de clasificación llamado Máquinas de Soporte Vectorial clasifica el conjunto de datos con el mejor porcentaje (84.7 en promedio), mostrando, los mejores resultados para el

tópico bancario con un 87.3 de instancias clasificadas correctamente, comportamiento que se presente en todos los algoritmos.

6. Conclusiones y trabajo futuro

Este artículo ha presentado un enfoque para la minería de opiniones centrada en tópicos, la cual consiste en identificar la polaridad de un mensaje (*tweet*). Para ello, se ha presentado un esquema para la clasificación de mensajes en tres categorías: positivo, negativo y neutro. La clasificación se lleva a cabo mediante cuatro algoritmos (NB, SVM, C4.5 y kNN). Para la clasificación se ha obtenido un vocabulario a partir de un conjunto de textos en español y posteriormente, este vocabulario ha sido ponderado mediante la métrica TF-IDF que se basa en la frecuencia de aparición del término en una colección de textos de un tópico.

A partir de los experimentos, se ha notado un mejor resultado en el algoritmo SVM en cuanto a precisión y cobertura para la identificación de polaridad utilizando los mensajes de Twitter, alcanzando un porcentaje promedio cercano a 85% de instancias clasificadas correctamente.

Adicionalmente, debido a que el objetivo de este trabajo es centrar la clasificación en tópicos, se puede aportar que la mejor identificación de polaridad se presenta en el tópico "Bancario" obteniendo una los promedios de cobertura más altos, para las tres categorías, en todos los algoritmos.

Las principales contribuciones de este trabajo son: a) un enfoque para la minería de opiniones que representa un recurso valioso para el análisis en textos extraídos de las redes sociales para el idioma español; b) la comparativa de cuatro algoritmos bajo el mismo escenario y configuración de los experimentos para la predicción de polaridad en los textos.

Como trabajo futuro, es posible realizar una clasificación basada en entidades ya que el conjunto de datos o corpus original incluye mensajes de texto para 61 entidades divididas en los cuatro tópicos mencionados. Un análisis de sentimientos combinando pares con la configuración *tópico: entidad* resultaría de gran utilidad para determinar la reputación de una entidad en un tópico específico.

Agradecimientos. Los autores agradecen al PRODEP-SEP, SNI-CONACyT y a la Universidad Autónoma Metropolitana Azcapotzalco por las facilidades proporcionadas para la realización de este artículo.

Referencias

1. Lima, A. C. E., de Castro, L. N., Corchado, J. M. A.: Polarity analysis framework for Twitter messages. Applied Mathematics and Computation, Vol. 270, pp. 756−767 (2015)
2. Khan, F. H., Bashir, S., Qamar, U.: TOM: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems, Vol. 57, pp. 245−257 (2014)

3. Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A., Mitkov, R.: Polarity classification for Spanish tweets using the COST corpus. Journal of Information Science (2015)
4. Kontopoulos, E., Berberidis, C., Dergiades, T., Bassiliades, N.: Ontology-based sentiment analysis of twitter posts. Expert systems with applications, Vol. 40, No. 10, pp. 4065−4074 (2013)
5. Ghiassi, M., Skinner, J., Zimbra, D.: Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with applications, Vol. 40, No. 16, pp. 6266−6282 (2013)
6. Tellez, E. S., Jiménez, S. M., Graff, M., Moctezuma, D., Suárez, R. R., Siordia, O. S.: A Simple Approach to Multilingual Polarity Classification in Twitter. CoRR abs/1612.05270 (2016)
7. Khan, A. Z., Atique, M., Thakare, V. M.: Combining lexicon-based and learning-based methods for Twitter sentiment analysis. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), 89 (2015)
8. Trinh, S., Nguyen, L., Vo, M., Do, P.: Lexicon-based sentiment analysis of Facebook comments in Vietnamese language. In: Recent Developments in Intelligent Information and Database Systems, pp. 263−276 (2016)
9. Abdulla, N. A., Ahmed, N. A., Shehab, M. A., Al-Ayyoub, M., Al-Kabi, M. N., Al-rifai, S.: Towards improving the lexicon-based approach for Arabic sentiment analysis. In Big Data: Concepts, Methodologies, Tools, and Applications, IGI Global, pp. 1970−1986 (2016)
10. Porter, M. F.: An algorithm for suffix stripping. Program, Vol. 14, No. 3, pp.130−137 (1980)
11. Karen, S. J., Peter, W.: Readings in Information Retrieval. San Francisco: Morgan Kaufmann (1997)
12. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys, 34 (1), pp. 1−47 (2002)
13. Aha, D. W., Kibler, D., Albert, M. K.: Instance-based learning algorithms. Machine Learning, Vol. 6, No. 1, pp. 37–66 (1991)
14. Chang, Ch., Lin, Ch.: LIBSVM - A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST), Vol. 2, No. 3, pp. 27−28 (2001)
15. John, G. H., Langley, P.: Estimating continuous distributions in Bayesian classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence (UAI'95), Montreal, Canada, pp. 338−345 (1995)
16. Quinlan, J. R.: C4.5: Programs for Machine Learning. San Mateo, Calif.: Morgan Kaufmann Publishers (1993)
17. Garner, S.R.: Weka: The Waikato environment for knowledge analysis. In: Proc. of the New Zealand Computer Science Research Students Conference, pp. 57−64 (1995)
18. Amigó, E., De Albornoz, J. C., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Spina, D.: Overview of replay 2013: Evaluating online reputation monitoring systems. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 333−352, Springer Berlin Heidelberg (2013)