Influencia del usuario en la clasificación de Tweets

Tweets Categorization: User Influence

J. Díaz Zamora[1], A. Fonseca Bruzón[2], R. Ortega Bueno[3]

[1]Facultad de Matemática y Computación, Universidad de Oriente, Santiago de Cuba, Cuba, juglar.diaz@cerpamid.co.cu
[2]Centro de Estudios de Reconocimiento de Patrones y Minería de Datos, Santiago de Cuba, Cuba, adrian@cerpamid.co.cu
[3]Centro de Estudios de Reconocimiento de Patrones y Minería de Datos, Santiago de Cuba, Cuba, reynier.ortega@cerpamid.co.cu

Abstract— **In this paper we analyze the impact of introduce user profile information in the task of tweets classification. For our study, we select a neighborhood based classification algorithm and build a user profile describing which topics users prefer to talk. This user profile modifies the voting phase in the classification algorithm. The results obtained in the experiments demonstrate that enhancing voting phase with the user information improve the classification results obtained.**

Keywords— **Tweets classification, user profile.**

I. Introducción

Con el reciente auge de las redes sociales como Twitter y Facebook, se ha producido un aumento de la cantidad de información disponible en textos cortos e informales, lo que ha motivado el interés de la comunidad científica en el desarrollo de técnicas de Minería de Textos sobre este tipo de información. Estos documentos plantean un reto en la comunidad científica por sus características específicas que los diferencian de los documentos tradicionales.

Una de las tareas de interés es la clasificación en tópicos de los mensajes producidos en la red social Twitter (tweets) [1,2,3]. Basándonos en que cada persona usualmente se interesa en un grupo de temas específicos, sobre los que suele escribir, buscar información y compartirla, si se desea clasificar un tweet, lo más probable es que pertenezca a alguno de los temas de interés de su autor. Este trabajo tiene por objetivo estudiar el impacto que tiene el considerar el perfil de los usuarios durante el proceso de categorización. Para conocer los temas sobre los que suelen escribir los usuarios con mayor frecuencia y utilizar esta información en la tarea de clasificación, construimos un perfil de usuario a partir de la muestra de entrenamiento y comprobamos cómo puede influenciar la clasificación de los tweets. Para nuestro estudio seleccionamos el algoritmo de clasificación αβ-NN [4], este algoritmo pertenece a la familia de los basados en vecindad [5,6]. Seleccionamos el αβ-NN por los buenos resultados reportados en la literatura, la facilidad de su implementación y la facilidad que nos brinda para incluir dentro del proceso de clasificación el perfil de usuario.

II. Materiales y métodos

Nuestro trabajo se enmarca en la categorización de los mensajes producidos en la red social Twitter. La red permite enviar mensajes de texto con un máximo de 140 caracteres, llamados tweets, que se muestran en la página principal del usuario. Los usuarios pueden subscribirse a los tweets de otros usuarios, a esto se llama "seguir", y al conjunto de usuarios subscritos a un usuario dado se les llama "seguidores". Por defecto los mensajes son públicos, pudiendo difundirse privadamente mostrándolos únicamente a unos seguidores determinados.

A continuación presentamos un tweet con algunas de sus características:

"Conozco a alguien (@jperez) q es adicto al drama!Jajaja te suena de algo!**#TV**"

Vemos algunas de las posibilidades que brinda la red como la mención de otro usuario (@jperez), el uso de etiquetas de tópico como (#TV), las cuales son utilizadas por los usuarios como un indicio del tema sobre el que están escribiendo. Comúnmente se les conoce a estas etiquetas por el nombre de *Hashtags*. Los tweets se caracterizan por el elevado grado de informalidad que presentan los usuarios al crear sus mensajes. Por ejemplo la frase: "Jajaja". Esta informalidad dificulta en gran medida el procesamiento y posterior clasificación de los tweets.

A. Algoritmo de clasificación

Como habíamos mencionado con anterioridad en nuestro trabajo utilizamos el algoritmo de clasificación αβ-NN. Este algoritmo pertenece a la familia de los algorit-

mos basados en vecindad [5,6]. Estos métodos basan su funcionamiento en la existencia de una medida de similitud entre los objetos a clasificar y están compuestos por tres pasos o fases principales: *construcción de la vecindad* a partir de la muestra de entrenamiento, *cálculo del voto*, fase en la cual cada categoría recibe un voto, y *regla de decisión*, donde se toma una decisión de acuerdo a los votos emitidos en la fase anterior. La familia de los algoritmos de clasificación basados en vecindad ha mostrado buenos resultados en tareas de clasificación de documentos y se caracterizan por su simplicidad conceptual, facilidad de implementación, que pueden ser construidos con pocos ejemplos de entrenamiento y permiten que exista solapamiento entre las categorías. Los pasos básicos del algoritmo αβ-NN [4] se muestran en el Algoritmo 1.

Algoritmo 1. αβ-NN

1. Sea *d* el objeto a clasificar.
2. Construir la vecindad N_β= { d_j/sim (*d*, d_j) ≥β}.
3. Sea *max* la similitud entre *d* y su vecino más cercano.
4. Sea $N_{\alpha\beta}$ la vecindad de *d*, $N_{\alpha\beta}$ =∅.
5. Para cada $d \in N_{\alpha\beta}$
   a) Si sim(*d*, d_j) ≥*max*−α:
      i) Anadir d_j a $N_{\alpha\beta}$
6. Calcular el voto *V*(c_i,*d)* para cada categoría c_i.
7. Para cada c_i:
   a) Si *V*(c_i,*d)*≥γ:
      i) Asignar *d* a c_i

El voto para cada clase *V*(c_i,*d)* es calculado mediante la siguiente fórmula:

$$V(d, c_i) = \frac{\sum_{d_j \in N_{\alpha\beta}(c_i)} sim(d, d_j)}{\sum_{d_j \in N_{\alpha\beta}} sim(d, d_j)} \quad (1)$$

Debido a la poca cantidad de caracteres en cada tweet, es poco probable que la frecuencia de cada palabra sea mayor que uno. Por esta razón representamos a los tweets como un conjunto de términos, la cual es más adecuada según la naturaleza propia de estos mensajes. Para medir el grado de similitud entre dos tweets seleccionamos el coeficiente de Jaccard [7], que se define como la relación entre el tamaño de la intersección de dos conjuntos y el tamaño de su unión:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

En la expresión, A y B representan los conjuntos de términos empleados en la representación de los tweets. Esta expresión obtiene un mayor valor en la medida en que la cantidad de términos en común que compartan los conjuntos A y B, sin importar el orden de las palabras en los tweets.

B. Perfil de usuario

Basándonos en el hecho de que cada usuario suele interesarse en un grupo específico de temas sobre los que escribe con mayor frecuencia, consideramos que incluir en el proceso de clasificación la información sobre los temas de interés para el usuario autor del tweet puede significar una mejora en los resultados.

Partiendo de la muestra de entrenamiento calculamos el perfil de usuario, el cual otorga un voto a cada clase. En nuestro caso modelamos el perfil del usuario como la frecuencia relativa con la cual el usuario escribe sobre cada uno de los temas de interés. Para ello empleamos la siguiente fórmula:

$$UP(u, c_i) = \frac{cantTweets(u, c_i) + 1}{cantTweets(u) + clases(u)} \quad (2)$$

Donde $cantTweets(u, c_i)$ es la cantidad de tweets del usuario *u* en la colección que pertenecen a la clase c_i y $cantTweets(u)$ es la cantidad de tweets total emitidos por el usuario *u*. Para modelar el hecho de que eventualmente un usuario pudiera escribir un tweet relacionado con alguna temática nueva, empleamos en la expresión un suavizado "*adding one*", por ello *clases(u)* es la cantidad de clases posibles en el problema.

La información del perfil de usuario es introducida en el algoritmo de clasificación durante la etapa del cálculo del voto para cada una de las clases. Los pasos básicos del algoritmo de clasificación αβ-NN incluyendo el perfil de usuario quedaría como se muestra en el Algoritmo 2.

Algoritmo 2. αβ-NN con perfil de usuario

1. Sea *d* el tweet a clasificar.
2. Construir la vecindad N_β= { d_j / Jaccard (*d*, d_j) ≥β}.
3. Sea *max* la similitud entre *d* y su vecino más cercano.
4. Sea $N_{\alpha\beta}$ la vecindad de *d*, $N_{\alpha\beta}$ =∅.
5. Para cada $d_j \in N_{\alpha\beta}$
   a) Si Jaccard (*d*, d_j) ≥*max*−α:
      i) Anadir d_j a $N_{\alpha\beta}$
6. Sea *u* el usuario autor de *d*, calcular el voto para cada categoría mediante la siguiente fórmula:

   $$VotoUP(c_i, d) = V(c_i, d) * UP(u, c_i)$$

7. Asignar *d* a la categoría con mayor voto.

Al multiplicar la Fórmula 1 por el voto del usuario, el voto final siempre disminuye y las clases sobre las que el autor del tweet escribe poco son afectadas en mayor medida, favoreciendo a las clases sobre las que escribe con mayor frecuencia.

III. Resultados y discusión

El algoritmo de clasificación no puede utilizarse directamente sobre el texto de los tweets, por lo que aplicamos un método de indexación para obtener una representación de cada tweet como un conjunto de términos. El método de indexación utilizado reconoce como términos las menciones de usuarios, las etiquetas de tópico (*Hashtags*), los nombres de entidades [8] y los lemas de los sustantivos y formas verbales [9].

Durante la experimentación se utilizó la colección desarrollada para la competición TASS-2013 [10]. El TASS-2013 es un taller de análisis de sentimiento en los medios sociales, concretamente en la red social Twitter. El taller consta de cuatro tareas: análisis de sentimientos a nivel global, clasificación en tópicos, análisis de sentimientos a nivel de entidad e identificación de tendencia política. Nuestros experimentos fueron realizados sobre la colección para la tarea de clasificación en tópicos de interés. A continuación presentamos el tweet visto anteriormente en formato XML en la forma en que aparecen los tweets en la colección sobre la que se realizaron los experimentos. El nombre de cada etiqueta XML indica el contenido que aparece dentro de ella:

```
<tweet>
<tweetid>000000000000</tweetid>
<user>usuario0</user>
    <content><![CDATA['Conozco a alguien (@jperez) q es adicto al drama! Jajaja te suena de algo!#TV']]>
    </content>
<date>2011-12-02T00:03:32</date>
<lang>es</lang>
<topics>
<topic>entretenimiento</topic>
</topics>
</tweet>
```

Tabla 1: Distribución por tópicos en las muestras de entrenamiento y de prueba.

Tópicos	Cantidad de tweets (Entrenamiento)	Cantidad de tweets (Prueba)
Política	3 120 (33%)	30 067 (43%)
Otros	2 337 (24%)	28 191 (40%)
Entretenimiento	1 678 (17%)	5 421 (8%)
Economía	942 (10%)	2 549 (3%)
Música	566 (6%)	1 498 (2%)
Fútbol	252 (3%)	823 (1%)
Películas	245 (3%)	596 (1%)
Tecnología	217 (2%)	287 (0%)
Deportes	113 (1%)	135(0%)
Literatura	103 (1%)	93(0%)
TOTAL	9 573	69 660

La colección cuenta con 68017 tweets escritos en español por personalidades del mundo de la política, el entretenimiento, los medios de comunicación y la cultura. Está dividida en 7219 tweets de entrenamiento y 60798 de prueba. En la Tabla 1 se muestra la frecuencia de tweets por cada tópico en el conjunto de entrenamiento y de prueba, en cada caso la suma total de las frecuencias absolutas de los tweets en cada clase supera el número total de tweets en el conjunto (de entrenamiento o de prueba) debido al solapamiento existente entre las diferentes categorías.

Para evaluar y comparar los resultados obtenidos se utilizaron las medidas de calidad *precisión*, *relevancia* y F_1 [11] calculadas de la siguiente manera:

$$presición = \frac{N(clasificaciones\ correctas)}{N(clasificaciones)},.$$

$$relevancia = \frac{N(clasificaciones\ correctas)}{N(documentos)}$$

$$F_1 = \frac{2 * presicion * relevancia}{presicion + relevancia}$$

En la Tabla 2 se muestran los resultados obtenidos en la medida F_1 considerando (o no) la información del perfil de usuario en la clasificación para combinaciones de α y β con resultados bajos, medios y altos.

Tabla 2: Resultados de la clasificación con perfil de usuario y sin perfil de usuario.

Mejor combinación αβ	Sin perfil de usuario	Con perfil de usuario
α=0.1 y β=0.1	**0.5969**	0.6818
α=0.08 y β=0.08	0.5960	**0.6819**
α=0.12 y β=0.09	0.5670	0.6600
α=0.08 y β=0.06	0.5742	0.6750
α=0.12 y β=0.03	0.5063	0.6185
α=0.08 y β=0.02	0.4978	0.6148
Promedio	**0.5563**	**0.6553**

En cada una de las posibles combinaciones de α y β la inclusión del perfil de usuario significó una mejora significativa en la medida F_1. Como promedio se mejoró en un 0.0989 y comparando el mejor resultado para cada una de las variantes se mejoró la calidad de la clasificación en un 0.085.

Como mencionamos anteriormente los experimentos fueron realizados sobre la colección desarrollada para el TASS-2013 [9], por lo que consideramos conveniente compararnos con los trabajos presentados en dicho taller. A continuación se muestra en la Tabla 3 el mejor resultado obtenido por cada grupo que presentó trabajos en el TASS-2013, además del mejor resultado para nuestra propuesta.

Tabla 3: Comparación con los resultados obtenidos en la competición TASS-2013.

Grupo	Precisión	Relevancia	F1
LYS	0.804	0.804	0.804
UNED_LSI	0.777	0.184	0.298
UPV	0.756	0.756	0.756
αβ con perfil de usuario	**0.731**	**0.638**	**0.682**
FHC25_IMDEA	0.719	0.702	0.710
UNED_JRM	0.479	0.479	0.479
SINAI	0.161	0.159	0.160

La clasificación utilizando αβ-NN con información de usuario queda ubicada en el cuarto lugar en *precisión* y F_1. Una vez analizados los valores presentes en la Tabla 3 podemos decir que los resultados obtenidos son alentadores. Nótese que en nuestra propuesta no empleamos recursos externos construidos de forma manual como es el caso de algunos de los participantes ocupantes de las primeras posiciones. Al analizar el valor de *relevancia* que obtiene el algoritmo, en comparación con el resto, y realizando un análisis de la colección de entrenamiento pudimos constatar que nuestra propuesta se ve afectada por el hecho de no considerar la multiclasificación y en la colección un volumen grande de tweets pertenecen a más de una categoría.

IV. Conclusiones

En este trabajo se presentó un estudio realizado para medir la influencia de los usuarios en la clasificación de los mensajes producidos en la red social Twitter. Los experimentos realizados muestran que la inclusión de la información sobre los temas que suelen escribir los usuarios influye de manera positiva en los resultados. Se realizó una comparación con los trabajos presentados en el TASS-2013 [10], donde nuestra propuesta se ubicó en cuarto lugar en las medidas de *precisión* y F_1 .Como trabajo futuro nos planteamos adaptar el algoritmo para considerar la posibilidad de que un tweet pueda ser clasificado en más de una categoría. Adicionalmente estudiaremos otras formas de modelar el perfil del usuario que involucre además los términos comúnmente empleados por los autores de los tweets.

Referencias

1. Sankaranarayanan, J, Samet H., Teitler B, Sperling, J (2009) TwitterStand: news in tweets. In Proc.ACM GIS'09 (Seattle, Washington, Nov. 2009), 42-51.
2. Phan X.-H., Nguyen L.-M., Horiguchi, S (2008) Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proc. WWW (Beijing, China, Apr. 2008), 91-100.
3. Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas,M (2010) Short Text Classification in Twitter to Improve Information Filtering in Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 2010, pp. 841–842.
4. Gil R, Pons A (2006) A New Nearest Neighbor Rule for Text Categorization. Progress in Pattern Recognition, Image Analysis and Applications Lecture Notes on Computer Science 225 814-823.
5. Moreno F., Micó L, Oncina (2004) J A new classification rule based on nearest neighbor search. In: 17th International Conference on Pattern Recognition. Volume 4, IEEE Computer Society (2004) 408–411.
6. Sánchez J, Pla F, Ferri F (1997) On the use of neighbourhood-based non-parametric classifiers. Pattern Recognition Letters18(1997) 1179–1186.
7. Jaccard P (1901) Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547-579.
8. García L, Pons A, Ruiz L (2007) A proposal of a Morphological Tagger for Spanish Based on Cuban Corpora. In Proceedings of International Conference on Recent Advances in Natural Language Processing. pp. 210 - 214. September, 2007. ISBN: 978-954-91743-7-3.
9. Cruz Y, Anaya H, Gil R, C Y (2007) Un enfoque híbrido al Reconocimiento de Nombres de Entidades para el español. Actas del X Simposio Internacional de Comunicación Social, Vol. 1. Leonel Ruiz Miyares, Alex Muñoz Alvarado, Celia Álvarez Moreno (Eds). pp. 521 - 524. Enero, 2007. ISBN: 959-7174-08-1.
10. Díaz A, Alegría I, Villena J (2013) Proceedings of the TASS workshop at SEPLN 2013. Actas del XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural. IV Congreso Español de Informática. 17-20 September 2013, Madrid, Spain. (eds). ISBN: 978-84-695-8349-4.
11. Van Rijsbergen, C (1974). Foundation of Evaluation.Journal of Documentation, 30(4), 365–373.