Conference: Congreso Interdisciplinario de Energías Renovables - Mantenimiento Industrial - Mecatrónica e Informática

Booklets

RENIECYT Registro Nacional de Instituciones y Empresas Científicas y Tecnológicas 2015-20795 CONACYT

RENIECYT - LATINDEX - Research Gate - DULCINEA - CLASE - Sudoc - HISPANA - SHERPA UNIVERSIA - E-Revistas - Google Scholar

DOI - REBID - Mendeley - DIALNET - ROAD - ORCID

Title: Revisión de técnicas de pre-procesamiento de textos para la clasificación automática de tweets en español

Author: Jesús Fidencio GARCÍA AMARO

Editorial label ECORFAN: 607-8324
BCIERMIMI Control Number: 2017-02
BCIERMIMI Classification (2017): 270917-0201

Pages: *22*
Mail: *jfgarcia@upfim.edu.mx*
RNA: 03-2010-032610115700-14

Contenido

1. Introducción
2. Clasificación automática de documentos
3. Metodología
4. Pre-procesamiento de *tweets*
5. Resultados
6. Conclusiones y trabajo futuro

1. Introducción

- El uso de redes sociales a jugado un papel muy importante en el intercambio de información.
- Twitter contribuyó a aumentar la conciencia mundial sobre los ataques terroristas en la India en 2008 (Barash & Golder, 2011).
- Al realizar búsquedas mediante palabras clave en Twitter, generalmente solo encuentra *tweets* donde hay coincidencia de palabras.

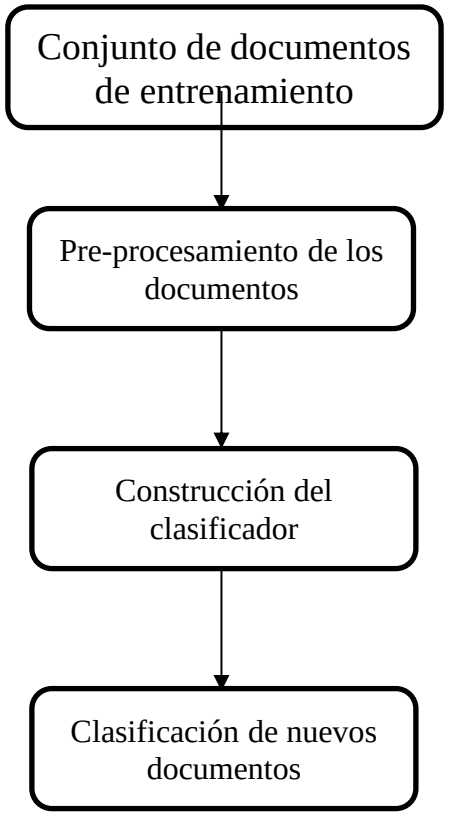
2. Clasificación automática de documentos

- La clasificación o categorización automática de documentos, tiene como finalidad la etiquetación de textos en lenguaje natural en categorías temáticas previamente definidas (Abelleira & Carolina A., 2010).
- Este proceso es posible, mediante la aplicación de técnicas de aprendizaje automático (Montes, 2001).

3. Metodología

La clasificación automática de textos ofrece una arquitectura de tres etapas: pre-procesamiento de los documentos, construcción del clasificador y clasificación de nuevos documentos (Abelleira & Carolina A., 2010).

Figura 1. Arquitectura para la clasificación de documentos



4. Pre-procesamiento de *tweets*

Existen múltiples técnicas para el pre-procesamiento de texto y cada una de ellas se debe escoger dependiendo el objetivo que se desee alcanzar.

De acuerdo a la revisión de la literatura, se identificaron dos grandes ramas: limpieza y normalización de textos.

4. Pre-procesamiento de *tweets*

Limpieza de textos

Se comprende la eliminación de palabras, símbolos o características superfluas contenidas en un *tweet* y que no aportan beneficio para el proceso de clasificación (García, Ramírez, Villatoro, & Jiménez, 2014).

A continuación, se mencionan componentes que son considerados en el proceso de limpieza aplicado a tweets en español.

4. Pre-procesamiento de *tweets*

Limpieza de textos

Direcciones web

- Todas las URL's en Twitter se acotan bajo el dominio de http://t.co/.
- Generalmente se eliminan por no generar valor en el proceso de clasificación.

4. Pre-procesamiento de *tweets*

Limpieza de textos

Sustitución o eliminación de emoticones

- Los emoticones pueden ser una característica importante cuando se analizan sentimientos (Jasso Hernández, Pinto, & Vilari, 2014) .
- Cuando este no sea el objetivo, los emoticones pueden ser eliminados.

4. Pre-procesamiento de *tweets*

Limpieza de textos

Caracteres especiales y signos de puntuación

- Los signos de admiración podrían ayudar a enfatizar ciertas características para el análisis de sentimientos (Martis & Alfaro, 2010) (Guevara, 2011).

Eliminación de múltiples espacios y saltos de línea

- Los espacios duplicados, los saltos de línea y retorno de carro son reemplazados por espacios sencillos (Guevara, 2011).

4. Pre-procesamiento de *tweets*

Limpieza de textos

Eliminación de palabras repetidas

- En ocasiones las personas agregan letras demás en las palabras o múltiples palabras repetidas seguidas para enfatizar sentimientos dentro de los textos, tales como ira, felicidad, éxtasis, etc. (Go, Bhayani, & Huang, 2009). Por ejemplo:

  - Te amooooo
  - Tengo mucha mucha hambre

4. Pre-procesamiento de *tweets*

Limpieza de textos

Eliminación de acentos

- Es recomendable la eliminación de acentos, ya que en el proceso de clasificación podría hacer diferencia entre la misma palabra, por ejemplo:

$$Capitán \neq Capitan$$
$$Árbol \neq Arbol$$

4. Pre-procesamiento de *tweets*

Normalización de textos

La normalización está basada en la estandarización del texto en un formato específico. Algunas técnicas son las siguientes:

Conversión de palabras a minúsculas

- Los usuarios suelen suelen intercalar mayúsculas y minúsculas indiscriminadamente.
- Es recomendable normalizar el texto, convirtiendo todas las publicaciones en minúsculas.

4. Pre-procesamiento de *tweets*

Normalización de textos

Eliminación de palabras vacías (stop words)

- El uso de esta técnica ayuda con la eliminación de palabras que no tienen significados relevantes como lo son: artículos, pronombres, preposiciones, etc. (Delgado, 2014) .

4. Pre-procesamiento de *tweets*

Normalización de textos

Extractores de raíces de palabras (stemming)

- Es un método utilizado para reducir una palabra a su raíz canónica o a un stem o lema (Bográn, Alonso, & García, 2013) .
- La aplicación de esta técnica ayuda a encontrar las palabras recurrentes en los documentos mediante su raíz. Por ejemplo:

$$Analizar, análisis, analizador => anali$$

4. Pre-procesamiento de *tweets*

Normalización de textos

Aplicación de diccionarios SMS

- Los tweets están inspirados en el servicio de mensajes de textos cortos conocidos como SMS, y las personas aún suelen usar estas abreviaturas: tales como: tkm, tmb, ntc, etc.

Diccionarios de corrección ortográfica

- En redes sociales, muchos usuarios no se preocupan por aplicar una correcta escritura entre sus publicaciones, o pueden recurrir a errores de escritura.

5. Resultados

- Los resultados arrojados en este punto de la investigación, difícilmente pueden ser expresados de manera cuantitativa, porque no existen métricas para definir con exactitud la efectividad de las técnicas aplicadas. En cambio, sí es posible expresarlas de forma cualitativa, observando el texto.

5. Resultados

Eliminación de caracteres especiales, signos de puntuación, acentos y emoticones

Tweet original	Tweet pre-procesado
¡El tiempo pasa, las cosas cambian, pero la esencia permanece! ;) ¿Cuál fue tu primera #PlayStation? :o #CosasDeGamer	El tiempo pasa las cosas cambian pero la esencia permanece Cual fue tu primera #PlayStation #CosasDeGamer

5. Resultados

Eliminación de palabras vacías y conversión de palabras a minúsculas

Tweet original	Tweet pre-procesado
La Selva Lacandona es el gran pulmón de México, representa 50% de las selvas tropicales húmedas del país y tiene amplia diversidad biológica https://t.co/nleGgNm0ZY	selva lacandona pulmón méxico, representa 50% selvas tropicales húmedas país amplia diversidad biológica

5. Resultados

Aplicación de extractores de palabras raíz y eliminación de direcciones web

Tweet original	Tweet pre-procesado
Chatear mientras maneja equivale a conducir después de tomar entre 15 y 20 cervezas http://bit.ly/2vyRgJL	chatear mientr mane equiva a conducir despu de tomar ent 15 y 20 cervez

5. Resultados

Implementación de múltiples técnicas

Tweet original	Tweet pre-procesado
Volver a #NuevaYork y tocar para toda #MiGente fue un hermoso regalo de la vida! Descarga mi app para ver más fotos: https://t.co/Wbj9wXQXo5	volver #nuevayork tocar #migen hermo rega vida descar app fot

6. Conclusiones y trabajo futuro

Conclusiones

- A pesar de que existen múltiples técnicas de pre-procesamiento de tweets, no todas son aplicables, y esto depende del objetivo que se desee alcanzar y la temática escogida.
- Al escoger incorrectamente una técnica, se corre el riesgo de eliminar atributos considerados como importantes.

6. Conclusiones y trabajo futuro

Trabajo futuro

- Se contempla la realización del estado del arte, para conocer los algoritmos de clasificación automática de textos que mejor trabajan con tweets en español.
- Desarrollar una aplicación que implemente un algoritmo de clasificación automática.

ECORFAN®