ITT Instituto de Investigación y Transferencia en Tecnología

UNNOBA Universidad Nacional Noroeste ▪ Buenos Aires

Desempeño de tareas de preprocesamiento sobre texto en español extraído de Facebook.

Introducción

El lenguaje utilizado en redes sociales, en general, difiere del que podríamos encontrar en otros medios, contiene errores ortográficos, palabras de uso cotidiano que no están formalmente aceptadas, emoticones, urls y otras construcciones que no suelen estar presentes en el lenguaje formal. Esto podría afectar el desempeño de los clasificadores de texto basados en aprendizaje automático.

Para este trabajo se utilizaron comentarios extraídos de Facebook. Dichos comentarios corresponden a diversas publicaciones pertenecientes a los medios de comunicación más consumidos en el país.

Objetivos

El presente trabajo busca determinar si normalizar las características previamente señaladas conduciría necesariamente a mejorar la exactitud de la clasificación.

Metodología

Considerando las características encontradas en los textos se propusieron diferentes tareas de preprocesamiento para normalizarlos, como eliminación de emoticones, urls, hashtags, nombres de usuario, caracteres y sílabas redundantes, tags HTML, elementos que no tuvieran letras; corrección de acentos; procesamiento de abreviaturas y acrónimos; palabras incorrectamente unidas por un punto, entre otras.

Se midió la cantidad de palabras incorrectas (OOV - Out of Vocabulary Words) antes y después de aplicarse cada tarea, de manera tal de determinar el porcentaje de reducción de OOVs.

Con los textos normalizados, obtenidos de aplicar cada tarea, se realizó un proceso 10-Fold Cross Validation utilizando como algoritmo de base “Naive Bayes”.

Autores

Leonardo Esnaola; Juan Pablo Tessore.

{leonardo.esnaola, juanpablo.tessore}@itt.unnoba.edu.ar

Instituto de Investigación y Transferencia en Tecnología (ITT) - Centro Asociado a la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC).

Resultados

A continuación se muestra una tabla que resume los resultados obtenidos. En cada caso siempre se compara con una línea base, es decir, con el texto original sin aplicar ninguna tarea de normalización:

Tarea de pre procesamiento	Reduction of OOV tokens (%)	Mejora en la clasificación (%)
Procesamiento de acronismos y abreviaturas	0,2016	0,1100
Eliminación de tags HTML	0,0007	0,0000
Eliminación de nombres de usuario	0,0030	-0,0006
Correción de palabras acentadas	7,9288	-0,0067
Eliminación de palabras sin letras	40,9195	-0,0250
Separación de palabras unidas incorrectamente	3,8822	0,0017
Procesamiento de emoticones	0,5204	-0,0089
Procesamiento de caracteres redundantes	2,4183	0,1961
Procesamiento de hastags	0,5995	-0,0533
Procesamiento de simbolos redundantes	-1,4148	0,0000
Eliminación de URL's	0,0795	-0,0139
Eliminación de silabas redundantes	0,9030	0,0044
Convertir el texto a mayusculas	0,1460	-0,0006

Conclusiones

Los resultados muestran que, a pesar de que determinadas tareas reducen significativamente el porcentaje de palabras OOV, esto no se traduce en una mejora en la precisión de la clasificación alcanzada. Por ejemplo, observando el caso de la tarea “Eliminación de palabras sin letras” puede verse que el porcentaje de OOV se reduce en más del 40% pero la exactitud del clasificador respecto de la línea base de hecho empeora.