Ejercicio 1

Aplicaciones de la Lingüística Computacional – Curso 2023/2024

Iñaki Diez Lambies

# Ejercicio 5

## Resultados

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Language** | **Spanish** | | | | | | | | |
| **Preprocessing** | **basic** | | | **spacy** | | | **spacy-pos** | | |
| **Vectorising** | **clf1** | **clf2** | **w2v** | **clf1** | **clf2** | **w2v** | **clf1** | **clf2** | **w2v** |
| **MCC** | 0,626 | 0,621 | 0,316 | 0,611 | 0,637 | 0,489 | 0,64 | 0,622 | 0,481 |
| **Precision** | 0,818 | 0,834 | 0,726 | 0,804 | 0,846 | 0,803 | 0,819 | 0,838 | 0,786 |
| **Recall** | 0,808 | 0,789 | 0,61 | 0,807 | 0,792 | 0,698 | 0,82 | 0,786 | 0,702 |
| **F1-score critical** | 0,856 | 0,873 | 0,801 | 0,856 | 0,878 | 0,839 | 0,868 | 0,874 | 0,836 |
| **F1-score conspiracy** | 0,768 | 0,733 | 0,398 | 0,755 | 0,738 | 0,579 | 0,772 | 0,728 | 0,593 |
| **F1-score macro** | 0,812 | 0,803 | 0,6 | 0,805 | 0,808 | 0,709 | 0,82 | 0,801 | 0,714 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Language** | **English** | | | | | | | | |
| **Preprocessing** | **basic** | | | **spacy** | | | **spacy-pos** | | |
| **Vectorising** | **clf1** | **clf2** | **w2v** | **clf1** | **clf2** | **w2v** | **clf1** | **clf2** | **w2v** |
| **MCC** | 0,658 | 0,634 | 0,426 | 0,618 | 0,593 | 0,492 | 0,647 | 0,626 | 0,497 |
| **Precision** | 0,831 | 0,848 | 0,759 | 0,82 | 0,843 | 0,815 | 0,834 | 0,86 | 0,828 |
| **Recall** | 0,828 | 0,789 | 0,675 | 0,798 | 0,756 | 0,692 | 0,813 | 0,768 | 0,688 |
| **F1-score critical** | 0,883 | 0,885 | 0,833 | 0,875 | 0,875 | 0,851 | 0,884 | 0,884 | 0,852 |
| **F1-score conspiracy** | 0,775 | 0,73 | 0,54 | 0,739 | 0,679 | 0,564 | 0,759 | 0,698 | 0,553 |
| **F1-score macro** | 0,829 | 0,807 | 0,686 | 0,807 | 0,777 | 0,707 | 0,822 | 0,791 | 0,791 |

## Evaluación de resultados

Los resultados obtenidos revelan patrones interesantes y variaciones en el rendimiento de los clasificadores en función del idioma, la técnica de preprocesamiento de texto y el algoritmo de clasificación empleado. Es notable que, de manera consistente a través de las configuraciones, los modelos entrenados en datos en español e inglés presentan un desempeño variado, lo cual sugiere que las características lingüísticas específicas y las diferencias en las distribuciones de las clases pueden influir significativamente en la eficacia del modelo.

Los modelos que utilizan el enfoque *basic* para el preprocesamiento tienden a mostrar un rendimiento robusto en términos de las métricas MCC (*Matthews Correlation Coefficient*), precisión, *recall* y las puntuaciones F1 para las clases críticas y conspirativas. El MCC muestra que los clasificadores *clf1* (*Multinomial Naive Bayes*) y *clf2* (Regresión Logística) generalmente superan al enfoque basado en *word embeddings* (*w2v*) en ambos idiomas. Esto podría indicar que, para este conjunto de datos y tarea específica, los métodos tradicionales de vectorización (conteo de términos y *Tf-Idf*) aún son muy efectivos, posiblemente debido a la naturaleza distintiva del vocabulario asociado con el pensamiento conspirativo y crítico.

La precisión y el *recall* junto con las puntuaciones F1 que armonizan estas dos métricas, sugieren que el preprocesamiento avanzado (*spacy* y *spacy\_pos*) ofrece beneficios marginales en algunos casos, especialmente cuando se utiliza en combinación con *clf2*. Esto resalta la importancia de una selección cuidadosa de características y la normalización del texto para mejorar la calidad de las representaciones textuales ante tareas de clasificación.

Sin embargo, el enfoque basado en *word embeddings* ("w2v") consistentemente muestra un rendimiento inferior en comparación con los enfoques basados en *bag-of-words* y *Tf-Idf*, lo que se refleja en menores MCC, precisión, *recall* y puntuaciones F1. Esto puede deberse a que, a pesar de capturar el significado semántico de las palabras, el promedio de *embeddings* puede no preservar suficientes detalles distintivos sobre la estructura y el contenido específicos del texto necesario para distinguir efectivamente entre el pensamiento crítico y conspirativo.

Estos resultados subrayan dos consideraciones importantes en la clasificación de textos: primero, la elección de la técnica de preprocesamiento y vectorización tiene un impacto significativo en el rendimiento del modelo y debe ser considerada cuidadosamente en función de la tarea específica. Segundo, aunque los métodos basados en *embeddings* ofrecen una representación rica del lenguaje, no siempre son la mejor opción para todas las tareas de clasificación, especialmente cuando las características discretas y la presencia o ausencia de términos específicos son predictores importantes de la clase.

## Bag-of-words

El enfoque de bolsa de palabras, a pesar de su simplicidad, conserva una relevancia notable en el campo del procesamiento de lenguaje natural, ofreciendo ventajas y enfrentando desafíos específicos que definen su aplicabilidad en diversas situaciones. Una de sus principales ventajas radica en la simplicidad y facilidad de implementación, lo que permite transformar textos en una forma que es directamente utilizable por modelos de aprendizaje automático sin necesidad de un profundo conocimiento en lingüística computacional. Además, al tratar los documentos como colecciones desordenadas de palabras, facilita la detección de patrones de uso de palabras y temas recurrentes, lo cual es especialmente útil en tareas de clasificación de texto y análisis de sentimientos donde la presencia de ciertas palabras puede ser un fuerte indicador de la categoría o el sentimiento del texto.

Sin embargo, este enfoque también presenta limitaciones significativas. Al ignorar el orden de las palabras y la estructura gramatical, el modelo puede perder información contextual esencial que afecta el significado del texto, llevando a posibles ambigüedades o interpretaciones erróneas. Además, la representación de bolsa de palabras tiende a ser de alta dimensionalidad, especialmente con grandes corpus de texto, lo que puede conducir a problemas de eficiencia y a la maldición de la dimensionalidad, donde el volumen de características puede eclipsar la cantidad de datos de entrenamiento disponibles.

A pesar de estos desafíos, el enfoque de bolsa de palabras sigue siendo útil en la actualidad, particularmente en aplicaciones donde la complejidad del modelo es una preocupación secundaria frente a la velocidad y la simplicidad. Por ejemplo, en sistemas de filtrado de spam, donde la presencia de ciertas palabras clave puede ser suficiente para clasificar los correos electrónicos de manera efectiva, o en la categorización inicial de contenido web, donde se puede emplear para organizar rápidamente grandes volúmenes de texto en categorías generales antes de aplicar análisis más detallados.

Aunque el enfoque de bolsa de palabras tiene sus limitaciones, su simplicidad, eficacia en ciertos contextos y la base que ofrece para el aprendizaje y la experimentación lo mantienen como una herramienta valiosa en el repertorio de técnicas de procesamiento de lenguaje natural.