Logotipo

Descripción generada automáticamente

Sistemas de Información no estructurada

Trabajo del Máster Universitario en Ingeniería Informática

Memoria Practica Final

Victor Fagúndez Poyo

E.T.S de Ingeniería Informática

Curso 2021/2022

|  |  |  |
| --- | --- | --- |
| Historial de Revisiones | | |
| **Versión** | **Fecha** | **Descripción** |
| 0.0.0 | 29.03.22 | Creación del documento |
| 1.0.0 | 03.05.22 | Finalización del documento |

**Contenido**

[1. Enunciado 1](#_Toc102487019)

[2. Consideraciones iniciales 3](#_Toc102487020)

[3. Algoritmo base 4](#_Toc102487021)

[3.1. Descripción de la implementación realizada 4](#_Toc102487022)

[3.1.1. Problemas encontrados 4](#_Toc102487023)

[3.2. Descripción de los resultados obtenidos 5](#_Toc102487024)

[3.2.1. Evaluación usando 10 cross-fold validation 5](#_Toc102487025)

[4. Limpiando los datos. Todo a minúsculas 6](#_Toc102487026)

[4.1. Descripción de la implementación realizada 6](#_Toc102487027)

[4.2. Descripción de los resultados obtenidos 6](#_Toc102487028)

[4.2.1. Evaluación usando 10 cross-fold validation 6](#_Toc102487029)

[5. Limpiando los datos. Eliminar signos de puntuación 7](#_Toc102487030)

[5.1. Descripción de la implementación realizada 7](#_Toc102487031)

[5.2. Descripción de los resultados obtenidos 8](#_Toc102487032)

[5.2.1. Evaluación usando 10 cross-fold validation 8](#_Toc102487033)

[6. Limpiando los datos. Eliminando Stopwords 9](#_Toc102487034)

[6.1. Descripción de la implementación realizada 9](#_Toc102487035)

[6.1.1. Problemas encontrados 9](#_Toc102487036)

[6.2. Descripción de los resultados obtenidos 10](#_Toc102487037)

[6.2.1. Evaluación usando 10 cross-fold validation 10](#_Toc102487038)

[7. Limpiando los datos. Eliminando nombres de usuarios 12](#_Toc102487039)

[7.1. Descripción de la implementación realizada 12](#_Toc102487040)

[7.2. Descripción de los resultados obtenidos 12](#_Toc102487041)

[7.2.1. Evaluación usando 10 cross-fold validation 12](#_Toc102487042)

[8. Limpiando los datos. Eliminando enlaces 14](#_Toc102487043)

[8.1. Descripción de la implementación realizada 14](#_Toc102487044)

[8.2. Descripción de los resultados obtenidos 14](#_Toc102487045)

[8.2.1. Evaluación usando 10 cross-fold validation 14](#_Toc102487046)

[9. Limpiando los datos. Eliminando tildes 16](#_Toc102487047)

[9.1. Descripción de la implementación realizada 16](#_Toc102487048)

[9.2. Descripción de los resultados obtenidos 16](#_Toc102487049)

[9.2.1. Evaluación usando 10 cross-fold validation 16](#_Toc102487050)

[10. Generación de la matriz tfif a partir de bi-gramas 18](#_Toc102487051)

[10.1. Descripción de la implementación realizada 18](#_Toc102487052)

[10.2. Descripción de los resultados obtenidos 18](#_Toc102487053)

[10.2.1. Evaluación usando 10 cross-fold validation 18](#_Toc102487054)

[11. Generación de la matriz tfif a partir de tri-gramas 20](#_Toc102487055)

[11.1. Descripción de la implementación realizada 20](#_Toc102487056)

[11.2. Descripción de los resultados obtenidos 20](#_Toc102487057)

[11.2.1. Evaluación usando 10 cross-fold validation 20](#_Toc102487058)

[12. Conclusiones 22](#_Toc102487059)

[12.1. Impacto de las variables en los resultados 22](#_Toc102487060)

[12.2. Combinaciones de variables 22](#_Toc102487061)

[12.3. Consideraciones finales 23](#_Toc102487062)

[13. Bibliografía 24](#_Toc102487063)

Índice de Ilustraciones

[Ilustración 1. Error no encuentra el fichero de entrenamiento 4](#_Toc102486989)

[Ilustración 2. Comprobación de la existencia de los datasets 4](#_Toc102486990)

[Ilustración 3. El script indica que el dataset no existe en esa ruta y finaliza 5](#_Toc102486991)

[Ilustración 4. Se añade un punto a la ruta, para acceder a la carpeta anterior a la actual 5](#_Toc102486992)

[Ilustración 5. Resultados del algoritmo base 5](#_Toc102486993)

[Ilustración 6. Función inicial de cleanData 6](#_Toc102486994)

[Ilustración 7. Accuracy primera modificación 6](#_Toc102486995)

[Ilustración 8. Codigo Eliminar signos y puntuación 7](#_Toc102486996)

[Ilustración 9. Accuracy de la segunda modificación 8](#_Toc102486997)

[Ilustración 10. Código para eliminar stopwords 9](#_Toc102486998)

[Ilustración 11. Mensaje de error. Falta la librería NLTK 9](#_Toc102486999)

[Ilustración 12. Instrucción para instalar NLTK en nuestro entorno 10](#_Toc102487000)

[Ilustración 13. Ejecutamos la instrucción en nuestro entorno 10](#_Toc102487001)

[Ilustración 14. Accuracy de la tercera modificación 11](#_Toc102487002)

[Ilustración 15.Llamada a la función de filtrado 12](#_Toc102487003)

[Ilustración 16.Función de filtrado 12](#_Toc102487004)

[Ilustración 17. Accuracy de la cuarta modificación 13](#_Toc102487005)

[Ilustración 18. Restos de enlaces web tras el cleanData 14](#_Toc102487006)

[Ilustración 19. Llamada a la función para filtrar las url 14](#_Toc102487007)

[Ilustración 20. Función para filtrar URL 14](#_Toc102487008)

[Ilustración 21. Accuracy de la quinta modificación 15](#_Toc102487009)

[Ilustración 22. Función para reemplazar los acentos 16](#_Toc102487010)

[Ilustración 23. Accuracy de la sexta modificación 17](#_Toc102487011)

[Ilustración 24. Generación de bi-gramas 18](#_Toc102487012)

[Ilustración 25. Datos ordenados por bigramas 18](#_Toc102487013)

[Ilustración 26. Accuracy de la séptima modificación 19](#_Toc102487014)

[Ilustración 27. Generación de tri-gramas 20](#_Toc102487015)

[Ilustración 28. Datos ordenados por tri-igramas 20](#_Toc102487016)

[Ilustración 29. Accuracy de la séptima modificación 21](#_Toc102487017)

[Ilustración 30. Resultados mejor conjunto de variables 23](#_Toc102487018)

Índice de Tablas

[Tabla 1. Resultados obtenidos a partir de las diferentes modificaciones realizadas 22](#_Toc102476655)

# Enunciado

Se deberán realizar un mínimo de 4 modificaciones. En concreto, el objetivo de la práctica es mejorar el conjunto de variables utilizadas en el clasificador SVM, ya sea mediante la modificación de las variables propuestas o mediante la inclusión de nuevas (generalmente mediante la concatenación para cada vector). La idea de la práctica es jugar con diferentes tipos de variables, analizar su funcionamiento y rendimiento mediante la evaluación, así como compararlas, ya sea por separado o combinadas. Algunas sugerencias de modificaciones son:

* Modificar el tamaño del diccionario para el cálculo de tf-idf, ya sea mediante la eliminación de Stop Words, mediante la selección de las XX palabras más usadas, 9 o la elección de ciertos conjuntos gramaticales (verbos y adjetivos, o nombres y adjetivos).
* Generación de la matriz de tf-idf mediante otros elementos que no sean palabras, como pueden ser bi-gramas, tri-gramas o n-gramas, así como el uso de lemas en lugar de palabras, o mediante concatenación de pares como puede ser palabra/categoría-Gramatical.
* Uso de conceptos en lugar de palabras mediante algoritmos de desambiguación semántica.
* Inclusión de elementos como los vectores de Word Embeddings.
* Inclusión de categorías gramaticales más frecuentes.
* Uso de léxicos afectivos o de palabras agresivas para identificar palabras polares (SentiSense, Spanish Emotion Lexicon, HurtLex, etc.)

Otra opción de mejora es la modificación del algoritmo de entrenamiento, pudiendo utilizar otras familias de algoritmos como pueden ser los árboles de decisión, los sistemas basados en regla, o la combinación de varios mediante sistemas de votación.

Como ya se ha expuesto, los puntos aquí planteados son sugerencias de posibles modificaciones, pero no es necesario restringirse a estos. Se pueden incluir cualesquiera originales, o ya vistos en la bibliografía o estado del arte de otros artículos científicos en el área de procesamiento del lenguaje o clasificación de texto.

Por último, se deberá realizar una memoria en la que se detalle el diseño de las modificaciones realizadas sobre el clasificador básico propuesto, su evaluación, análisis de resultados y conclusiones extraídas. En concreto la memoria deberá contener:

* Descripción de la implementación realizada, problemas encontrados, soluciones aportadas, etc.
* Descripción de resultados obtenidos para diferentes variables, y de todas las alternativas/combinaciones realizadas, con un análisis exhaustivo de los mismos en base a las distintas métricas utilizadas.
* Conclusiones que se puedan obtener a partir del análisis de resultados: cómo afectan las variables a los resultados, sus combinaciones, mejor conjunto de variables y por qué, etc.

# Consideraciones iniciales

Dado que los datasets de entrenamiento y test proporcionados para la práctica superaban con las 6000 y 4000 líneas respectivamente, se ha decidido, con el fin de acelerar las pruebas intermedias durante el desarrollo de las modificaciones del algoritmo, que se reducirán estos datasets a un tamaño de 100 líneas cada uno de las secciones en castellano. Una vez se hayan finalizado las modificaciones, y se tenga un prototipo de algoritmo viable, se realizará una ejecución con los datasets originales.

# Algoritmo base

## Descripción de la implementación realizada

Se realizará en primer lugar una ejecución del algoritmo base sin modificaciones a modo de referencia para el resto de las modificaciones que se implementen.

### Problemas encontrados

#### Problema 1

##### Descripción del problema

Cuando tratamos de ejecutar el script por primera vez, nos devuelve un mensaje de error, el cual se muestra en la ilustración 1.

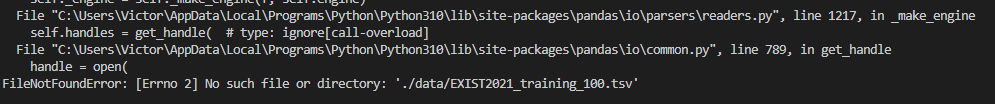


Ilustración 1. Error no encuentra el fichero de entrenamiento

##### Solución

Con el fin de realizar una comprobación de la existencia de los archivos de entrenamiento antes entrar a la función processEXISTTraining(), se ha optado por añadir una condición if que compruebe la existencia de los ficheros en la ruta proporcionada.

Texto

Descripción generada automáticamente

Ilustración 2. Comprobación de la existencia de los datasets

Si volvemos a ejecutar el script nos devuelve los siguiente:

Texto

Descripción generada automáticamente

Ilustración 3. El script indica que el dataset no existe en esa ruta y finaliza

En base a esto se interpreta que el script no esta consiguiendo acceder a la carpeta en la que se encuentran los datasets en la carpeta *data* del nivel anterior. Por ello, en la ruta descrita para pathTraining y pathTest se añade un punto en la parte anterior a la ruta ya descrita para que retroceda al nivel de jerarquía anterior. Por lo tanto, el código quedaría de la forma siguiente.

Texto

Descripción generada automáticamente

Ilustración 4. Se añade un punto a la ruta, para acceder a la carpeta anterior a la actual

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

Tras ejecutar el script obtenemos los resultados mostrados en la ilustración 5.

Calendario

Descripción generada automáticamente

Ilustración 5. Resultados del algoritmo base

Si observamos los resultados del algoritmo base tenemos que tiene una precisión o accuracy del 0.66.

# Limpiando los datos. Todo a minúsculas

Con el fin de homogeneizar los datos obtenidos a la hora de tokenizar las palabras obtenidas de los diferentes tweets, en esta primera aproximación nos centraremos en poner en minúsculas todas las palabras obtenidas tras el proceso de tokenización.

## Descripción de la implementación realizada

Para realizar esto, creamos una nueva función dentro del script a la que llamaremos cleanData(). Esta función podrá servirnos para futuras modificaciones del script a la hora de limpiar los datos obtenidos del algoritmo. De momento, para reducir a minúsculas se utilizará el siguiente código.

Texto

Descripción generada automáticamente

Ilustración 6. Función inicial de cleanData

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

Si ejecutamos nuevamente el script con la modificación, obtenemos una leve mejoría en el accuaracy del script, pasando de una precisión del 0.66 al 0.68, como podemos observar en la ilustración siguiente.

Calendario

Descripción generada automáticamente

Ilustración 7. Accuracy primera modificación

# Limpiando los datos. Eliminar signos de puntuación

El siguiente paso para limpiar los datos obtenidos al tokenizar los mensajes del dataset, es eliminar los signos de puntuación y espacios en blanco.

## Descripción de la implementación realizada

Para realizar esto, dentro de la función que creamos anteriormente llamada cleanData() añadimos una nueva sección destinada a la eliminación de símbolos de puntuación y espacios en blanco. Como se muestra en la imagen a continuación.

Texto

Descripción generada automáticamente con confianza media

Ilustración . Codigo Eliminar signos y puntuación

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

Si ejecutamos nuevamente el script con la segunda modificación, obtenemos una leve mejoría en el accuracy del script, pasando de una precisión del 0.68 al 0.70, como podemos observar en la ilustración siguiente.

Calendario

Descripción generada automáticamente con confianza media

Ilustración . Accuracy de la segunda modificación

# Limpiando los datos. Eliminando Stopwords

El siguiente paso para limpiar los datos obtenidos al tokenizar los mensajes del dataset, es eliminar las stopwords o palabras vacías, estas son palabras que carecen de sentido propio cuando se escriben solas o sin la palabra clave o keyword.[1]

## Descripción de la implementación realizada

Para realizar esto, dentro de la función que creamos anteriormente llamada cleanData() añadimos una nueva sección destinada a la eliminación de stopwords. Para ello necesitamos importar la librería nltk con herramientas de gestión del lenguaje.

Para ello utilizamos el siguiente código de la ilustración siguiente.

Texto

Descripción generada automáticamente

Ilustración . Código para eliminar stopwords

Esto permitirá en primer lugar filtrar todas las palabras que no aparten información que están en la lista de stopwords de la librería nltk. Y en segundo lugar eliminamos todos los palabras vacías o ‘’ que se hayan quedado en la lista despues de la limpieza y normalización de los datos.

### Problemas encontrados

#### NLTK no disponible

##### Descripción del problema

Dado que usaremos la librería NLTK (Natural Lenguage Toolkit) [2] necesitamos no solo importarla en nuestro script, si no que también debemos instalarla en nuestro entorno de desarrollo.

Es por ello que cuando tratamos de ejecutar el script nos devuelve el siguiente mensaje de error que se muestra en la ilustración 10.

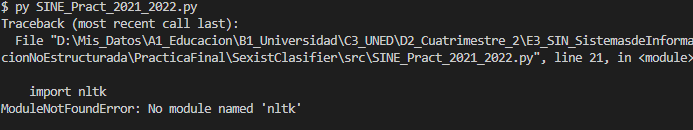


Ilustración . Mensaje de error. Falta la librería NLTK

##### Solución

Para ello si acudimos a la documentación oficial de la librería en su parte de instalación [2]encontraremos las instrucciones para instalar la librería en nuestro entorno[3].

Interfaz de usuario gráfica

Descripción generada automáticamente

Ilustración . Instrucción para instalar NLTK en nuestro entorno

Ejecutamos la instrucción en nuestra consola, y ya la tendríamos instalada.

Texto

Descripción generada automáticamente

Ilustración . Ejecutamos la instrucción en nuestro entorno

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

Si ejecutamos nuevamente el script con la segunda modificación, obtenemos un leve empeoramiento en el accuracy del script, pasando de una precisión del 0.7 al 0.66, como podemos observar en la ilustración siguiente. Sin embargo, la desviación estándar se ha reducido de un 0.10 a un 0.06.

Calendario

Descripción generada automáticamente

Ilustración 14. Accuracy de la tercera modificación

# Limpiando los datos. Eliminando nombres de usuarios

Con el fin de evitar sesgos a determinados usuarios, se ha obtado por eliminar toda referencia a nombres de usuarios de los comentarios utilizados.

## Descripción de la implementación realizada

Para realizar esto, buscaremos todas las palabras que empiecen por un @ y las eliminaremos de la lista dentro de nuestra función cleanData(). Para ello, usamos el código que mostramos a continuación.



Ilustración .Llamada a la función de filtrado

Texto

Descripción generada automáticamente

Ilustración .Función de filtrado

Para determinar si una palabra es un nombre de usuario buscamos si contiene el símbolo @.

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

Si ejecutamos nuevamente el script con la segunda modificación, obtenemos un leve empeoramiento en el accuracy del script, pasando de una precisión del 0.66 al 0.65, como podemos observar en la ilustración siguiente. Sin embargo, la desviación estándar se ha mantenido en un 0.06.

Calendario

Descripción generada automáticamente

Ilustración . Accuracy de la cuarta modificación

# Limpiando los datos. Eliminando enlaces

Si observamos los datos despues del proceso de tokenización y despues de pasar por cleanData() como se muestra en la ilustración siguiente, observamos que hay enlaces que no aportan información relevante.

Imagen en blanco y negro

Descripción generada automáticamente con confianza media

Ilustración . Restos de enlaces web tras el cleanData

Por ello, se procederá en este apartado a eliminar estos enlaces.

## Descripción de la implementación realizada

Para ello, buscaremos todas las palabras que empiecen o contengan https. A continuación se muestra el código utilizado para filtrar las URL.



Ilustración . Llamada a la función para filtrar las url

Texto

Descripción generada automáticamente

Ilustración . Función para filtrar URL

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

Si ejecutamos nuevamente el script con la quinta modificación, obtenemos que el accuracy del script no vario, permanece a la precisión de 0.65, como podemos observar en la ilustración siguiente. Y la desviación estándar se ha mantenido en un 0.06.

Calendario

Descripción generada automáticamente

Ilustración 21. Accuracy de la quinta modificación

# Limpiando los datos. Eliminando tildes

Para normalizar aun mas los datos obtenidos, vamos a eliminar todas las tildes, sustituyendo esos caracteres por sus equivalentes sin tildes.

## Descripción de la implementación realizada

Para ello, buscaremos todas los caracteres con acentos sustituyéndolos por sus equivalentes sin acentos como se muestra en el código mostrado en la siguiente ilustración.

Pantalla de computadora con letras

Descripción generada automáticamente con confianza media

Ilustración . Función para reemplazar los acentos

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

Si ejecutamos nuevamente el script con la sexta modificación, obtenemos que el accuracy del script se incrementó del 0.65 anterior a un 0.66, como podemos observar en la ilustración siguiente. Y la desviación estándar se ha reducido de un 0.06 a un 0.05.

Calendario

Descripción generada automáticamente

Ilustración 23. Accuracy de la sexta modificación

# Generación de la matriz tfif a partir de bi-gramas

En este apartado probaremos a generar la matriz de tf-idf mediante otros elementos que no sean palabras, como pueden ser los bi-gramas, es decir, pares de palabras.

## Descripción de la implementación realizada

Para ello, despues de haber limpiado los datos con la función que creamos en apartados anteriores cleanData() usaremos la función bigrams() de la librería nltk para formar los bigramas como se muestra en el código de la ilustración siguiente.



Ilustración . Generación de bi-gramas

De esta forma los datos quedarían colocados por pares de la siguiente forma:

Texto

Descripción generada automáticamente

Ilustración . Datos ordenados por bigramas

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

Si ejecutamos nuevamente el script con esta modificación, obtenemos que el accuracy del script se reduce del 0.66 anterior a un 0.57, como podemos observar en la ilustración siguiente. Y la desviación estándar se ha reducido de un 0.05 a un 0.03.

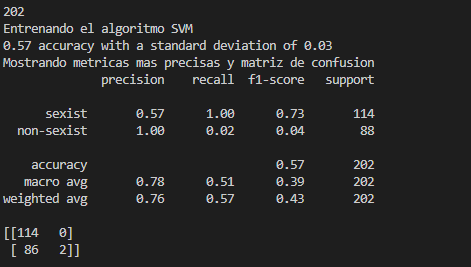


Ilustración 26. Accuracy de la séptima modificación

# Generación de la matriz tfif a partir de tri-gramas

En este apartado probaremos a generar la matriz de tf-idf mediante tri-gramas, es decir, conjuntos de tres palabras.

## Descripción de la implementación realizada

Para ello, despues de haber limpiado los datos con la función que creamos en apartados anteriores cleanData() usaremos la función trigrams() de la librería nltk para formar los trigramas como se muestra en el código de la ilustración siguiente.



Ilustración 27. Generación de tri-gramas

De esta forma los datos quedarían colocados por conjuntos de tres palabras de la siguiente forma:

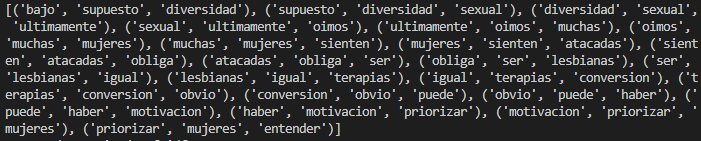


Ilustración 28. Datos ordenados por tri-igramas

## Descripción de los resultados obtenidos

### Evaluación usando 10 cross-fold validation

Si ejecutamos nuevamente el script con esta modificación, obtenemos que el accuracy del script no cambia con respecto al uso de bigramas, como podemos observar en la ilustración siguiente. Y la desviación estándar también se mantiene a un 0.03.

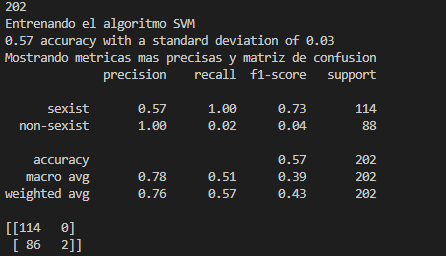


Ilustración 29. Accuracy de la séptima modificación

# Conclusiones

## Impacto de las variables en los resultados

A continuación, realizaremos una tabla en la que compararemos los distintos accuracy generados con cada modificación del algoritmo base para un dataset de 100 datos tanto para entrenamiento como para test.

Tabla . Resultados obtenidos a partir de las diferentes modificaciones realizadas

|  |  |  |
| --- | --- | --- |
| Modificación | Accuracy | Desviación Típica |
| Algoritmo Base | 0.66 | 0.08 |
| Todo a minúsculas | 0.68 | 0.10 |
| Eliminar signos de puntuación | 0.70 | 0.10 |
| Eliminar stopwords | 0.66 | 0.06 |
| Eliminar nombres de usuarios | 0.65 | 0.06 |
| Eliminar enlaces | 0.65 | 0.06 |
| Eliminar tildes | 0.66 | 0.05 |
| Matriz tfidf a partir de bigramas | 0.57 | 0.03 |
| Matriz tfidf a partir de trigramas | 0.57 | 0.03 |

Como podemos observar en los resultados obtenidos las modificaciones de pasar a minúsculas, eliminar los signos de puntuación y eliminar las tildes mejoraron la precisión o accuracy. Sin embargo, el paso a minúsculas y la eliminación de los signos de puntuación provoco un aumento de la desviación típica.

Por el contrario, eliminar las stopwords, los nombres de usuario, generar la matriz tfidf a partir de bigramas o trigramas provocaron una perdida de precisión o accuracy, aunque generaron una mejoría en la desviación típica.

## Combinaciones de variables

En base a estos datos, hemos realizado pruebas probando una combinación de estas modificaciones probemos a dejar solo las modificaciones siguientes:

* Todo a minúsculas
* Eliminación signos de puntuación
* Eliminar stopwords
* Eliminar nombres de usuario
* Eliminar enlaces
* Eliminar tildes

Que produce los siguientes resultados:

Calendario

Descripción generada automáticamente

Ilustración . Resultados mejor conjunto de variables

Podemos afirmar que, aunque eliminar las stopwords reduce la precisión produce una mejora relevante en cuanto a la desviación típica.

## Consideraciones finales

El código de esta practica puede encontrarse en [4].

# Bibliografía

[1] “¿Qué son las Stop Words y para qué sirven? – BlackBeast.” https://blackbeast.pro/diccionario/stop-words/ (accessed May 02, 2022).

[2] “NLTK :: Natural Language Toolkit.” https://www.nltk.org/index.html (accessed May 02, 2022).

[3] “nltk · PyPI.” https://pypi.org/project/nltk/ (accessed May 02, 2022).

[4] “vincitori-dev/MUNED\_SIN\_SexistClasifier: Codigo fuente del proyecto final de la asignatura Sistema de Información no estructurada del Master Universitario en Ingenieria Informática.” https://github.com/vincitori-dev/MUNED\_SIN\_SexistClasifier (accessed May 03, 2022).