OpinionMOOC

La retroalimentación del profesor virtual

Alumno: José Antonio Ruiz Millán
Tutora: Maria Victoria Luzón García
Tutor: Eugenio Martínez Cámara

Índice

1. **MOOC**
2. Nuestro problema
3. Análisis
4. Diseño
5. Experimentos
6. Conclusiones

1. Definición MOOC

Un MOOC (acrónimo en inglés de Massive Open Online Course - curso online masivo abierto), **es un curso a distancia en línea** dirigido a un amplio número de participantes a través de Internet según el principio de educación abierta y masiva.

Estos cursos **almacenan la información de los usuarios** que los realizan para así mejorar y poder ofrecerle al usuario la atención adecuada y precisa dependiendo de sus necesidades en cada momento.

Índice

2. **Nuestro problema**
   a. **Definición**
   b. Hipótesis
   c. Objetivos

2.a. Definición del problema

El problema a resolver es la **clasificación de opinión** de los alumnos que realizan un determinado MOOC. Para ello partiremos de un xMOOC el cual proporciona una encuesta de satisfacción para que los alumnos puedan contestar mediante una pregunta abierta cómo se sienten respecto al curso.

El objetivo es **automatizar el proceso de clasificación** de cada uno de los comentarios, agilizar el proceso para que así el desarrollador del curso pueda obtener un resumen sobre cómo el curso está siendo aceptado por los alumnos **sin necesidad de revisar uno a uno todos los comentarios**.

Índice

2. **Nuestro problema**
   a. Definición
   b. Hipótesis
   c. Objetivos

2.b. Hipótesis

La hipótesis que da pie a este proyecto es la siguiente: *A partir de un conjunto de comentarios sobre un MOOC es posible el análisis y clasificación automática de la orientación de la opinión que los estudiantes tienen sobre dicho MOOC.*

Para ello, se desarrollarán las siguientes tareas:

1. Estudio de la bibliografía relacionada con el problema.
2. Compilación de datos.
3. Anotación de datos.
4. Propuesta de un clasificador.
5. Evaluación de la propuesta.
6. Análisis de errores.
7. Publicación de los resultados.

Índice

3. **Nuestro problema**
   a. Definición
   b. Hipótesis
   c. Objetivos

2.c. Objetivos

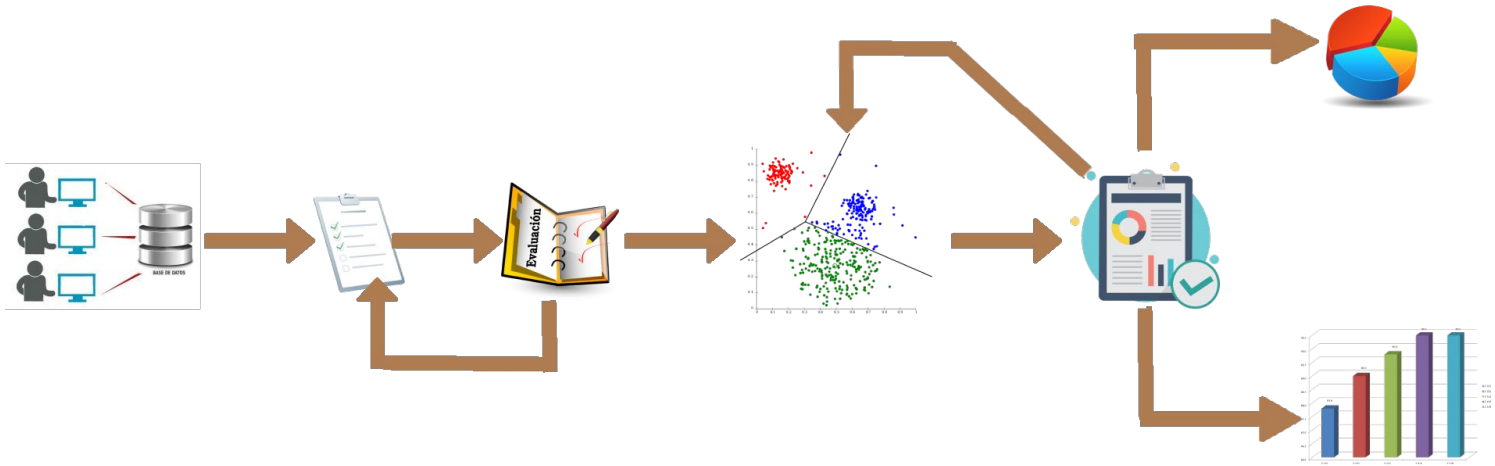
Los objetivos marcados para este proyecto son:

1. Estudiar el estado del arte de clasificación de la opinión, en especial la clasificación de la opinión en español.
2. Estudiar el proceso a seguir para anotar unos datos lingüísticos y cómo hacerlo.
3. Estudiar distintos tipos de métodos para evaluar la calidad de anotación de los datos, haciendo uso de alguno de ellos para evaluar nuestra propia anotación.
4. Desarrollar un clasificador de opiniones sobre un curso específico de la plataforma AbiertaUGR.
5. Visualizar de la forma más comprensible posible el resultado de la clasificación para facilitar la toma de decisiones.
6. Desarrollar técnicas de adaptación al dominio del sistema desarrollado.

Índice

3. **Análisis**
   a. **Descripción**
   b. AbiertaUGR
   c. Conjunto de datos
   d. Guía de anotación
   e. Ejercicio de anotación
   f. Conjunto de datos clasificados
   g. Diagrama de clases
   h. Análisis léxico

3.a Descripción



Índice

3. **Análisis**
   a. Descripción
   b. AbiertaUGR
   c. Conjunto de datos
   d. Guía de anotación
   e. Ejercicio de anotación
   f. Conjunto de datos clasificados
   g. Diagrama de clases
   h. Análisis léxico

3.b. AbiertaUGR

AbiertaUGR será la encargada de proporcionarnos los datos. Los datos son un conjunto de respuestas a comentarios abiertos donde los alumnos expresan su nivel de satisfacción antes de comenzar el curso y también expresan el nivel de satisfacción respecto a la encuesta en sí.

AbiertaUGR es el nombre de la plataforma de formación abierta de la Universidad de Granada que ofrece formación de coste gratuito y abierto a cualquier persona interesada. Ofrece una diversidad de cursos que podemos realizar cualquiera de nosotros. Estos cursos nos dan toda la documentación necesaria ya sea a través de ficheros (generalmente PDF) o ya sea través de vídeos.

Índice

3. **Análisis**
   a. Descripción
   b. AbiertaUGR
   c. Conjunto de datos
   d. Guía de anotación
   e. Ejercicio de anotación
   f. Conjunto de datos clasificados
   g. Diagrama de clases
   h. Análisis léxico

3.c. Conjunto de datos

1. **Pregunta/entidad Q13:** Conjunto compuesto de **226 elementos** que refleja la respuesta de los alumnos en la que expresan su nivel de satisfacción antes de realizar el curso y así poder ver como se encuentran cada uno de ellos.
2. **Pregunta/entidad Q14:** Conjunto compuesto de **68 elementos** que refleja la respuesta de los alumnos en la que valoran la encuesta realizada, es decir, expresen su nivel de satisfacción respecto a la propia encuesta que acaban de realizar.

3.c. Conjunto de datos

Tanto Q13 como Q14 tienen el mismo formato, compuesto por los siguientes atributos:

- **id:** Identificador único del comentario indicando al curso al que pertenece la encuesta.
- **label:** Texto escrito por el alumno en respuesta a la pregunta concreta.
  - **Q13:** “Muy contenta y con ganas de comenzar el curso”
  - **Q14:** “Corta y sencilla, así da gusto.”
- **y**: Valor de clasificación del comentario después de haber seguido un proceso de anotación.

Índice

3. Análisis

a. Descripción
b. AbiertaUGR
c. Conjunto de datos
d. Guía de anotación
e. Ejercicio de anotación
f. Conjunto de datos clasificados
g. Diagrama de clases
h. Análisis léxico

3.d. Guía de anotación

Tenemos 6 grados de satisfacción, enumerados desde el 5 al 0, siendo el **5 el valor que indica el máximo nivel de satisfacción** y el **1 el que indica el menor nivel de satisfacción**. Hacemos uso de la etiqueta **0 para catalogar los comentarios que no indican un nivel de satisfacción**. Mostraremos ejemplos sobre cada uno de los casos, indicando a la entidad a la que pertenecen siendo **E.1** la entidad que se refiere al estado de ánimo del usuario antes del curso (Q13) y **E.2** la entidad que se refiere a la encuesta (Q14):

- **Grado (clase) 5**: Comentarios totalmente positivos, mostrando un entusiasmo alto respecto a su estado actual.
  - **E.1** - “Muy contenta y con ganas de comenzar el curso”
  - **E.2 -** “Muy sencilla y fácil”

3.d. Guía de anotación

- **Grado (clase) 4**: Comentarios positivos que pueden tener algún aspecto no positivo pero el comentario generalizado es positivo. Comentarios positivos sin mostrar un excesivo entusiasmo.
  - **E.1 -** “Animada por aprender cosas nuevas”
  - **E.2 -** “Me parece curiosa cuanto menos”
- **Grado (clase) 3:** Comentarios neutros, tienen parte positiva y negativa a partes iguales. No se puede definir con exactitud si muestran una satisfacción positiva o negativa.
  - **E.1 -** “Pues con ilusión aunque creo que va a ser muy duro”
  - **E.2 -** “Parece interesante, aunque en algunos aspectos se hace pesada.”
- **Grado (clase) 2:** Comentarios negativos sin excesiva seguridad sobre ellos. En el comentario se muestra negatividad pero con duda. No es una respuesta tajante negativa.
  - **E.1 -** “Perdida y agobiada”
  - **E.2 -** “No le veo gran utilidad.”

3.d. Guía de anotación

- **Grado (clase) 1:** Comentarios totalmente negativos, todo respecto a su estado de ánimo es negativo y se muestra una actitud totalmente desconforme.
  - **E.1 -** “Muy cansado y deprimido”
  - **E.2 -** “No sirve para nada”
- **Grado (clase) 0:** Comentarios que no indican un grado de satisfacción. No aporta nada al problema que estamos estudiando.
  - **E.1 -** “Es el primer año que estudio en la Universidad”
  - **E.2 -** “Nada que comentar.”

Índice

3. Análisis

a. Descripción
b. AbiertaUGR
c. Conjunto de datos
d. Guía de anotación
e. Ejercicio de anotación
f. Conjunto de datos clasificados
g. Diagrama de clases
h. Análisis léxico

3.e. Ejercicio de anotación

Basándonos en la guía anterior, se ha realizado un ejercicio de anotación c**ompuesto por 3 integrantes**, los cuales se ha escogido a **2 de ellos para etiquetar** cada una de las entradas en el conjunto de datos y así poder comprobar mediante distintas métricas el nivel de concordancia. **El tercer anotador se utiliza para decidir la etiqueta final** en los casos donde los 2 primeros anotadores no coinciden.

Para **evaluar el ejercicio de anotación**, decidimos utilizar el **acuerdo observado** y el **acuerdo esperado** y no simplemente el acuerdo observado para así intentar acercarnos algo más a la realidad.

3.e. Ejercicio de anotación

- **Acuerdo observado:** Expresa el porcentaje de acuerdo entre los observadores, es decir, en qué medida hubo coincidencia en la clasificación entre los observadores en relación al total de elementos examinados.
- **Acuerdo esperado:** Es la probabilidad hipotética de acuerdo por azar.

Decidimos por lo tanto utilizar métricas que tengan en cuenta estos dos elementos, por ello, escogimos ***k de Cohen*** y ***Π de Scott*** como métricas para evaluar el ejercicio de anotación.

3.e. Ejercicio de anotación

Vamos a definir brevemente las métricas seleccionadas.

$$k, \pi = \frac{A_0 - A_e}{1 - A_e}$$

$$A_0 = \frac{1}{i} \sum_{i \in I} agr_i$$

$$A_e^{\pi} = A_e^{k} = \sum_{k \in K} P(k|c_1) \cdot P(k|c_2)$$

Con la única diferencia de que en *π de Scott* asumimos la misma distribución de probabilidad para cada anotador, mientras que en *k de Cohen* se asume una distribución de probabilidad distinta para cada anotador

3.e. Ejercicio de anotación

Interpretación de los resultados:

Valores de kappa	Fuerza del acuerdo
< 0.0	Pobre
0.0-0.20	Leve
0.21-0.40	Justa
0.41-0.60	Moderada
0.61-0.80	Considerable
0.81-1.00	Perfecta

Obtuvimos un valor de **0.527389 de *k de Cohen*** y un valor de **0.52686 de *Π de Scott* en Q13**.
Obtuvimos un valor de **0.673273 de *k de Cohen*** y un valor de **0.672141 de *Π de Scott* en Q14**.

Fijándonos en la tabla de resultados y en los valores obtenidos, podemos concretar que para la pregunta **Q13** hemos obtenido un valor que está considerado de **concordancia *"Moderada"*** lo que nos permite aceptar el resultado y seguir con el proceso. Por otra parte, en la pregunta **Q14** hemos obtenido un valor de **concordancia *"Considerable"***, por lo tanto, aceptamos los resultados obtenidos.

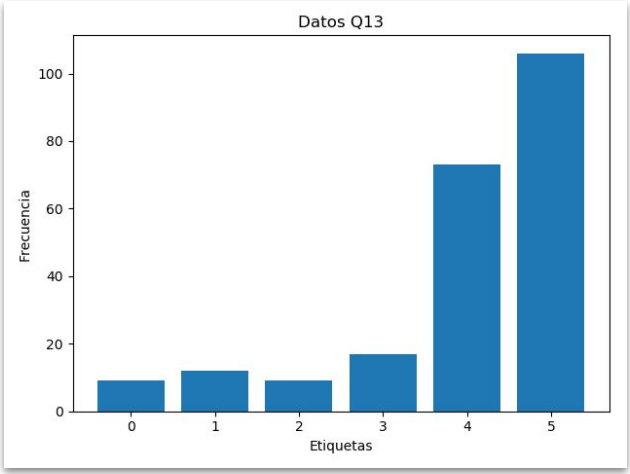
Índice

3. Análisis

a. Descripción
b. AbiertaUGR
c. Conjunto de datos
d. Guía de anotación
e. Ejercicio de anotación
f. Conjunto de datos clasificados
g. Diagrama de clases
h. Análisis léxico

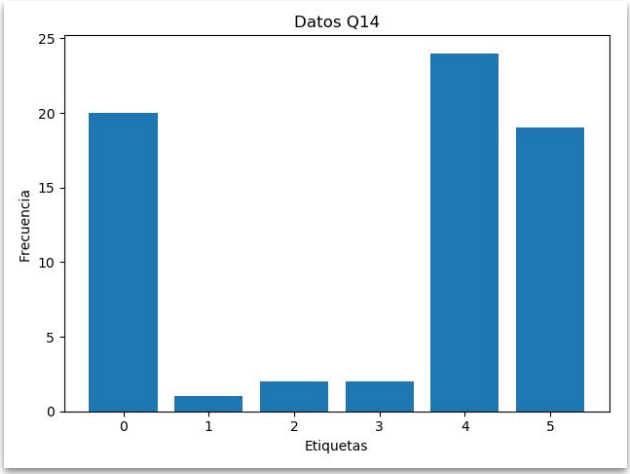
3.f. Conjunto de datos clasificados

Conjunto de datos Q13:



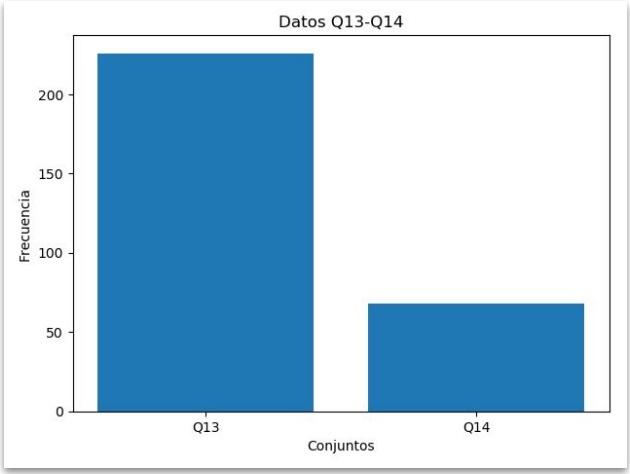
3.f. Conjunto de datos clasificados

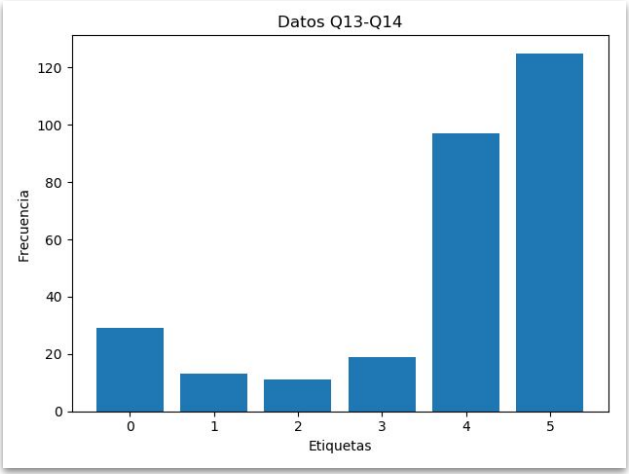
Conjunto de datos Q14:



3.f. Conjunto de datos clasificados

Conjunto de datos Q13 + Q14:

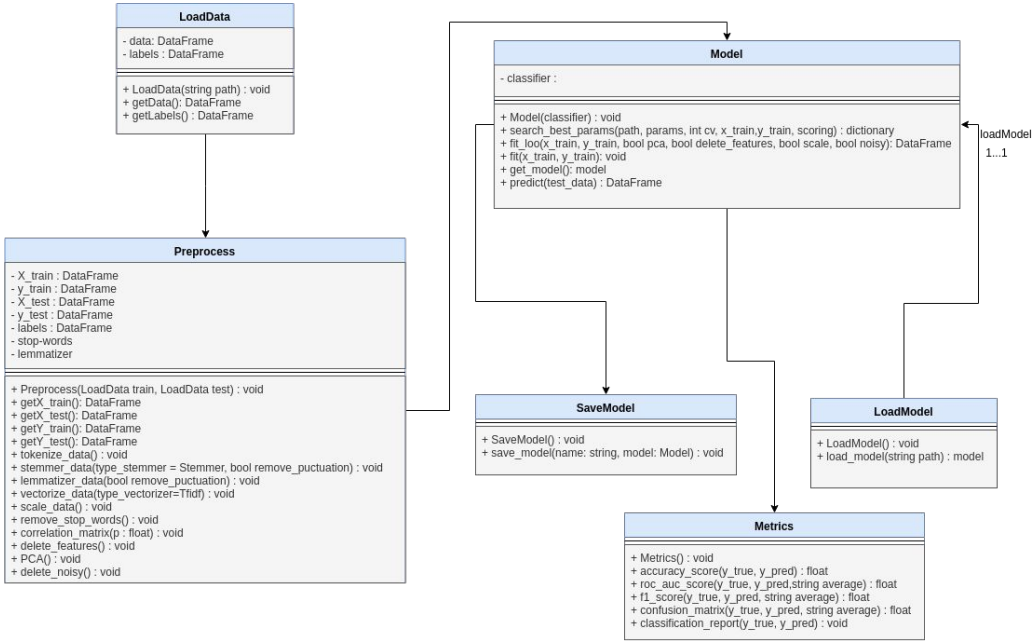




Índice

3. **Análisis**
    a. Descripción
    b. AbiertaUGR
    c. Conjunto de datos
    d. Guía de anotación
    e. Ejercicio de anotación
    f. Conjunto de datos clasificados
    g. Diagrama de clases
    h. Análisis léxico

3.g. Diagrama de clases



Índice

3. **Análisis**
   a. Descripción
   b. AbiertaUGR
   c. Conjunto de datos
   d. Guía de anotación
   e. Ejercicio de anotación
   f. Conjunto de datos clasificados
   g. Diagrama de clases
   h. Análisis léxico

3.h. Análisis léxico

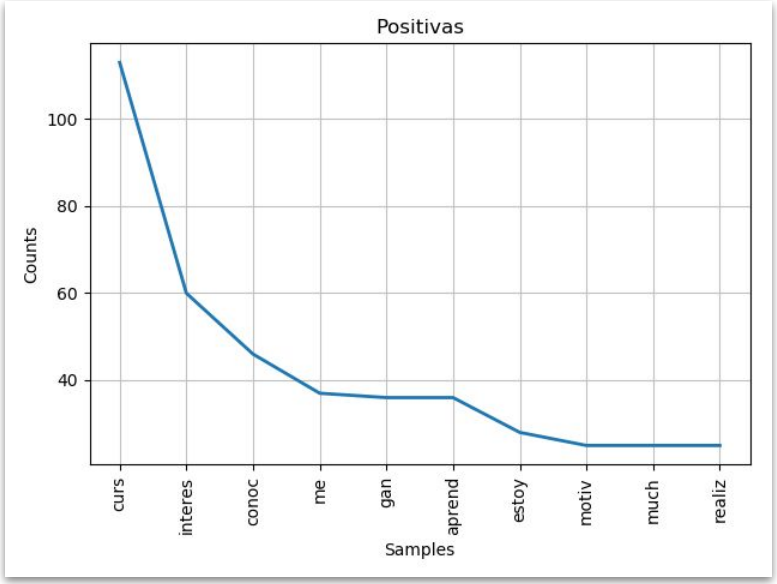
Decidimos hacer un análisis léxico para facilitar y entender los datos, haciendo una agrupación en la que podamos a parte de entenderlos, sacar conclusiones de ellos.

Para ello, hemos creado 4 grupos:

- **Clase positiva**: Etiquetas 4 y 5.
- **Clase negativa**: Etiquetas 1 y 2.
- **Clase neutra**: Etiqueta 3.
- **Clase nula**: Etiqueta 0.

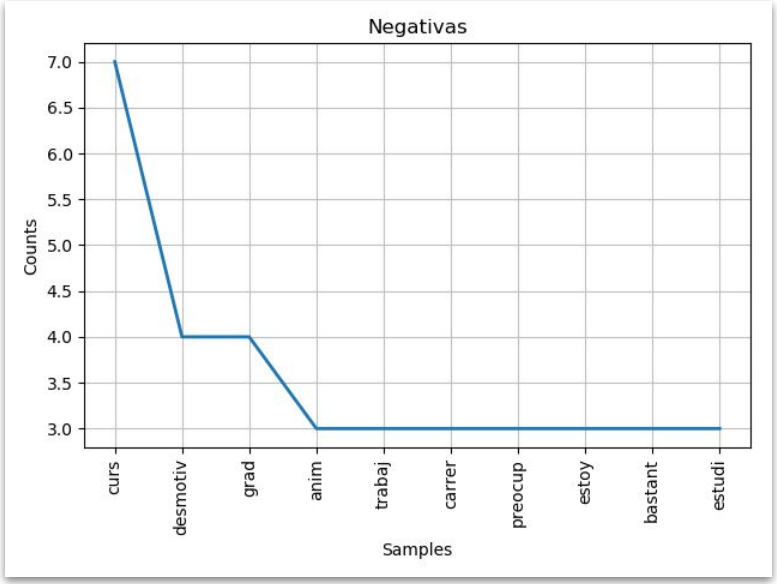
3.h. Análisis léxico - positivas

Palabra	Frecuencia	Frecuencia/Total
curs	113	0.17
interes	60	0.092
conoc	46	0.07
me	37	0.057
gan	36	0.055
aprend	36	0.055
estoy	28	0.043
motiv	25	0.038
much	25	0.038
realiz	25	0.038



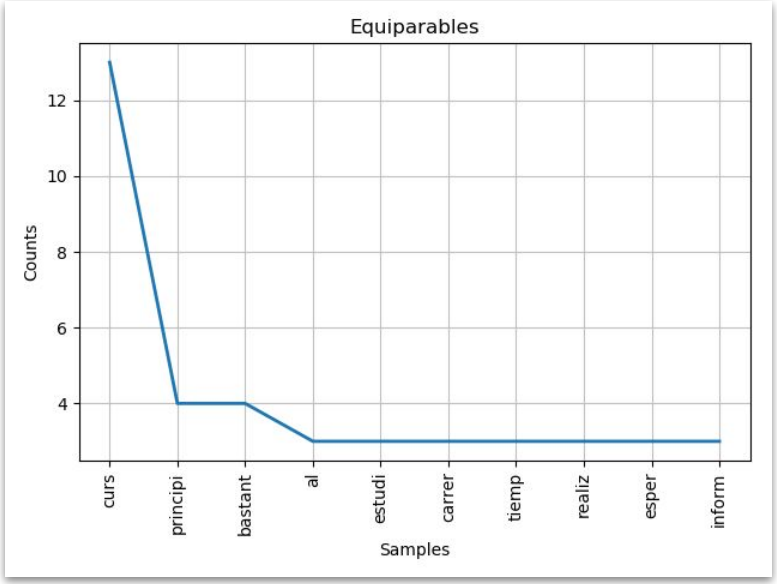
3.h. Análisis léxico - negativas

Palabra	Frecuencia	Frecuencia/Total
curs	7	0.068
desmotiv	4	0.039
grad	4	0.039
anum	4	0.039
trabaj	3	0.029
carrer	3	0.029
preocup	3	0.029
estoy	3	0.029
bastant	3	0.029
estudi	3	0.029



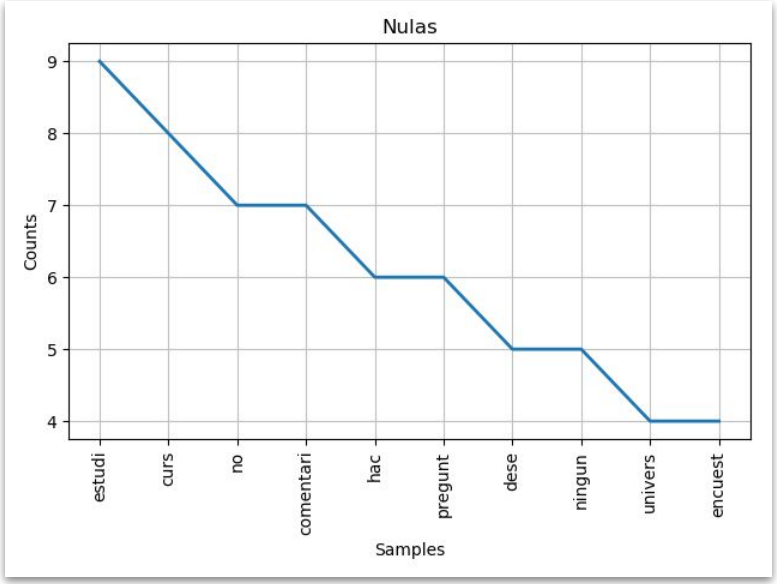
3.h. Análisis léxico - neutras

Palabra	Frecuencia	Frecuencia/Total
curs	13	0.082
principi	4	0.025
bastant	4	0.025
al	3	0.019
estudi	3	0.019
carrer	3	0.019
tiemp	3	0.019
realiz	3	0.019
esper	3	0.019
inform	3	0.019



3.h. Análisis léxico - nulas

Palabra	Frecuencia	Frecuencia/Total
estudi	9	0.05
curs	8	0.044
no	7	0.039
comentari	7	0.039
hac	6	0.033
pregunt	6	0.033
dese	5	0.028
ningun	5	0.028
univers	4	0.022
escuest	4	0.022



Índice

4. **Diseño**
   a. **Caso base**
   b. Caso techo
   c. Casos de diseño
   d. Procesado de texto
   e. Preprocesado de datos
   f. Algoritmos seleccionados

4.a. Caso base

Hemos creado un caso base simple para poder compararlo con los diferentes diseños que realizaremos posteriormente.

Este caso se basa en el **algoritmo OneR**, el cual clasifica siempre todos los elementos con la **etiqueta más común**, es decir, la etiqueta que más asignaciones tiene en el conjunto de datos.

Por lo tanto, éste caso clasificará todos los elementos como **“5” en Q13** y como **“4” en Q14** ya que como hemos visto anteriormente son las clases que más apariciones tienen en cada uno de los conjuntos.

Índice

4. **Diseño**
   a. Caso base
   b. Caso techo
   c. Casos de diseño
   d. Procesado de texto
   e. Preprocesado de datos
   f. Algoritmos seleccionados

4.b. Caso techo

Creamos este caso para **comparar los resultados de los algoritmos respecto al porcentaje de acierto humano**.

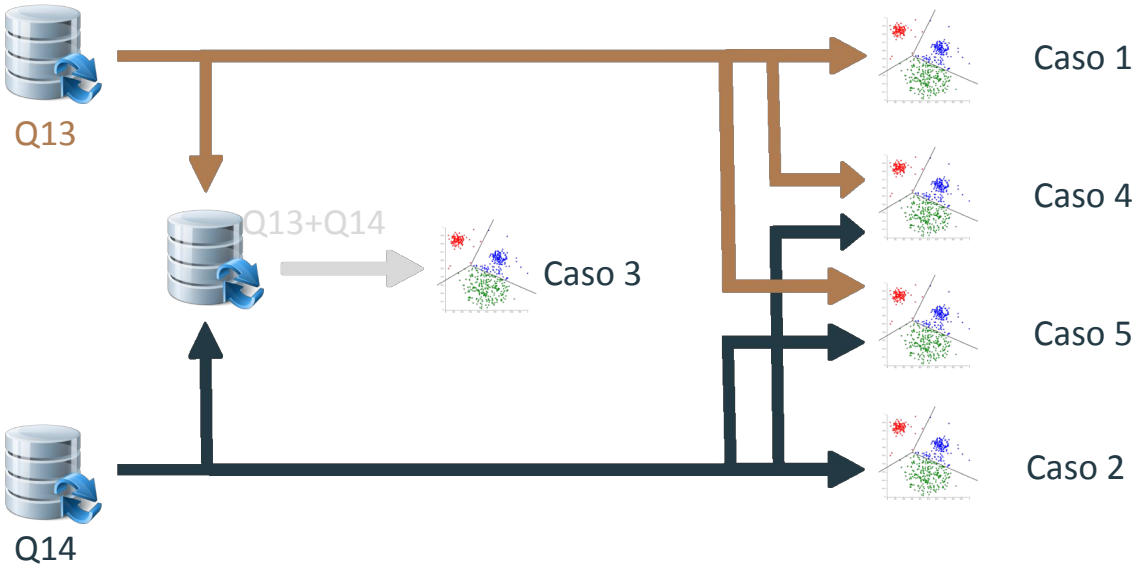
Por lo tanto, utilizaremos este caso para ver si el algoritmo ha sido capaz de superar el porcentaje de acuerdo que han sido capaz de obtener dos personas.

Haciendo el experimento, los **dos anotadores humanos** tuvieron un acuerdo observado de un **0.681 en Q13** y de **0.764 en Q14.** Claramente si un humano ha tenido un porcentaje de acierto con estos valores, podemos esperar que la máquina como mucho, iguale o se acerque por abajo a estos datos.

Índice

4. **Diseño**
   a. Caso base
   b. Caso techo
   c. Casos de diseño
   d. Procesado de texto
   e. Preprocesado de datos
   f. Algoritmos seleccionados

4.c. Casos de diseño



Índice

4. **Diseño**
   a. Caso base
   b. Caso techo
   c. Casos de diseño
   d. Procesado de texto
   e. Preprocesado de datos
   f. Algoritmos seleccionados

4.d. Procesado de texto

Para el procesado de texto se ha utilizado el paquete **ntlk** de Python el cual nos proporciona una enorme lista de funciones para trabajar con textos en castellano. Los procesos que hemos seguido para ésta transformación son los siguientes:

- **Tokenización:** Separación de frases en tokens.
- **Eliminar "stop-words":** Eliminar palabras sin información en nuestro lenguaje.
- **Aplicar un Stemmer:** Permite resumir palabras con la misma raíz en una misma palabra. En particular, para nuestro problema utilizaremos la función "*SnowBallStemmer*".
- **Vectorizar:** Nos permite transformar el texto el números. En nuestro caso, utilizaremos *Tfidf* para realizar esta transformación.

$$tidf(t,d) = tf(t,d) \cdot idf(t)$$
$$idf(t) = \log(n/df(t)) + 1$$

Índice

4. **Diseño**
   a. Caso base
   b. Caso techo
   c. Casos de diseño
   d. Procesado de texto
   e. Preprocesado de datos
   f. Algoritmos seleccionados

4.e. Preprocesado de datos

Una vez tenemos los textos transformados, decidimos hacer un preprocesado de datos en algunos de los casos diseñados. El proceso que seguimos es el siguiente:

- **Eliminación de ruido:** Se utiliza una estrategia *"Ensamble filter".*
- **Algoritmo PCA:** Nos permite eliminar características sin perder información.
- **Selección de características:** Selección de características utilizando algoritmos que computen la importancia de las características. En nuestro caso hemos utilizado "*Random Forest".*
- **Escalado:** Al realizar los procesos anteriores, necesitamos escalar los datos [0-1] para que los algoritmos utilizados funcionen correctamente.

Índice

4. **Diseño**
   a. Caso base
   b. Caso techo
   c. Casos de diseño
   d. Procesado de texto
   e. Preprocesado de datos
   f. Algoritmos seleccionados

4.f. Algoritmos seleccionados

Dados los conjuntos de datos que tenemos y las características del problema, hemos decidido utilizar **3 algoritmos distintos**. En realidad utilizamos 2 algoritmos, pero de 1 de ellos utilizamos 2 modelos distintos.

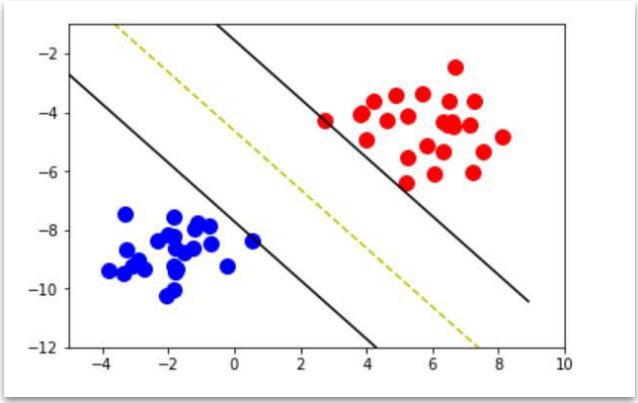
Para resolver el problema y crear los distintos modelos para cada uno de los casos diseñados anteriormente, hemos decidido utilizar el **algoritmo SVM** y el a**lgoritmo de Naive Bayes**.

Dentro del algoritmo de Naive Bayes, utilizaremos la versión de **Naive Bayes Multinomial** y la versión de **Naive Bayes Complement**.

4.f. Algoritmos seleccionados - SVM

Ventajas:

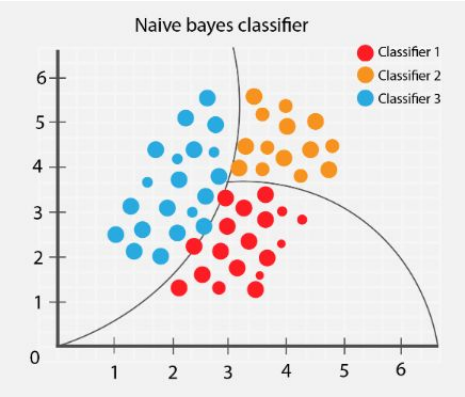
- Es un algoritmo efectivo en grandes dimensiones.
- Sigue siendo efectivo en casos donde el número de dimensiones es mayor que el número de muestras.
- Utiliza un subconjunto de puntos los cuales se encuentran en la frontera y ésto hace que el algoritmo sea eficiente en memoria.



4.f. Algoritmos seleccionados - Naive Bayes

Ventajas:

- Necesita pocos datos para poder estimar la predicción medianamente bien.
- Al suponer las caracteriísticas independientes, pueden calcularse como una distribución unidimensional y asií eliminar el problema de el número de dimensiones elevado
- Es extremadamente rápido en comparación con otros métodos sofisticados.



4.f. Algoritmos seleccionados - Naive Bayes Multinomial

Esta modificación claramente **sigue la estructura del algoritmo básico** que acabamos de explicar, no obstante, este algoritmo ha sido diseñado para trabajar con **variables numéricas enteras**.

No obstante, se ha comprobado en la práctica que el algoritmo funciona correctamente cuando las **variables numéricas son reales** en vez de enteras. Se suele utilizar en problemas donde los datos son textos.

Por ello, este algoritmo ha sido escogido para nuestros experimentos.

4.f. Algoritmos seleccionados - Naive Bayes Complement

Al igual que el anterior, este algoritmo trabaja **siguiendo los principios del algoritmo básico** explicado anteriormente. Es más, esta modificación **se basa en Naive Bayes Multinomial** explicado anteriormente.

Lo que los diferencia es que éste algoritmo tiene en cuenta también la distribución de las clases, es decir, t**iene en cuenta si los datos están balanceados o no**.

Por ello, también decidimos escoger esta modificación del algoritmo Naive Bayes.

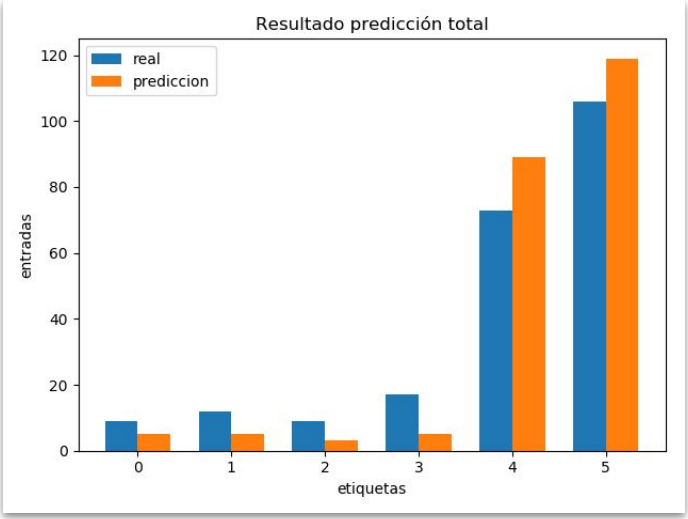
Índice

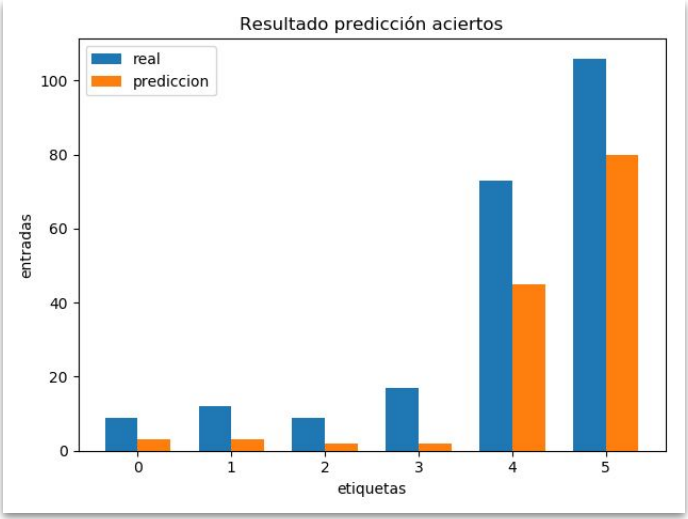
5. **Experimentos**
   a. **Pregunta Q13**
   b. Pregunta Q14
   c. Pregunta Q13+Q14
   d. Train Q13 - Test Q14
   e. Train Q14 - Test Q13

5.a. Pregunta Q13

	Accuracy	ROC	F1	Precision	Recall	Tiempo (s)
SVM	0.597	0.675	0.576	0.59	0.6	13.42
Caso base	0.469	0.5	0.299	0.22	0.47	0
Caso techo	0.681					

5.a. Pregunta Q13





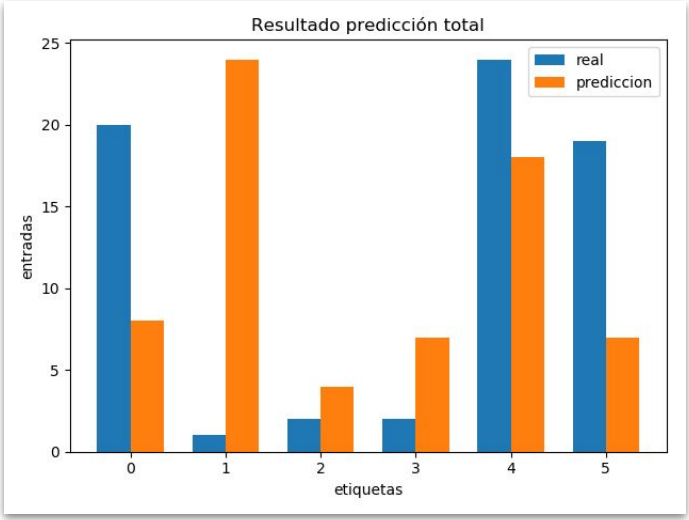
Índice

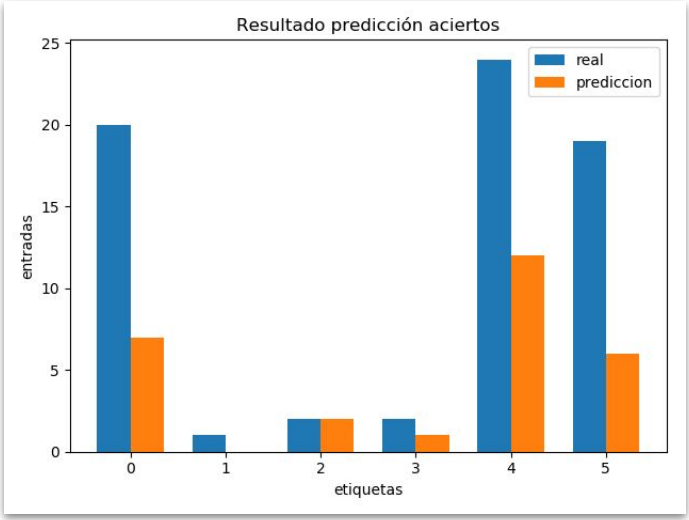
5. **Experimentos**
   a. Pregunta Q13
   b. Pregunta Q14
   c. Pregunta Q13+Q14
   d. Train Q13 - Test Q14
   e. Train Q14 - Test Q13

5.b. Pregunta Q14

	Accuracy	ROC	F1	Precision	Recall	Tiempo (s)
NB Complement	0.412	0.6715	0.504	0.75	0.41	0.04
Caso base	0.279	0.5	0.122	0.08	0.28	0
Caso techo	0.764					

5.b. Pregunta Q14





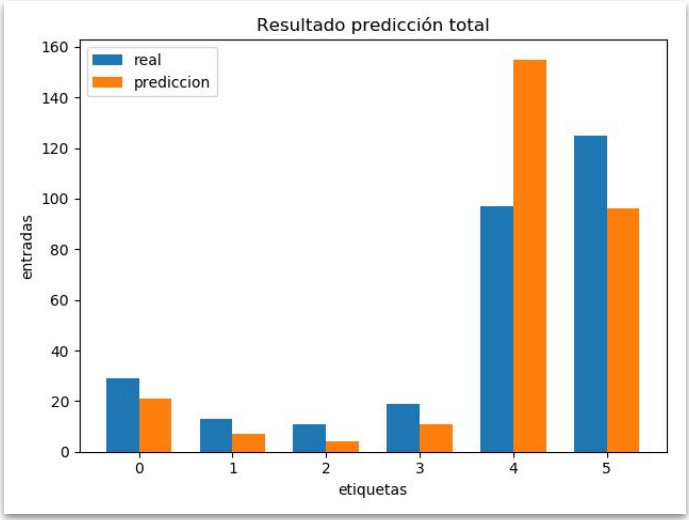
Índice

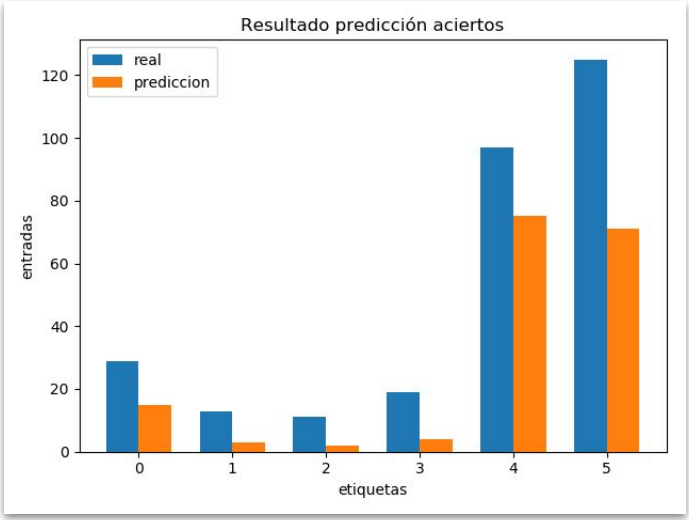
5. **Experimentos**
   a. Pregunta Q13
   b. Pregunta Q14
   c. Pregunta Q13+Q14
   d. Train Q13 - Test Q14
   e. Train Q14 - Test Q13

5.c. Pregunta Q13+Q14

	Accuracy	ROC	F1	Precision	Recall	Tiempo (s)
SVM	0.578	0.688	0.569	0.61	0.58	30.42
Caso base	0.425	0.5	0.254	0.18	0.43	0

5.c. Pregunta Q13+Q14





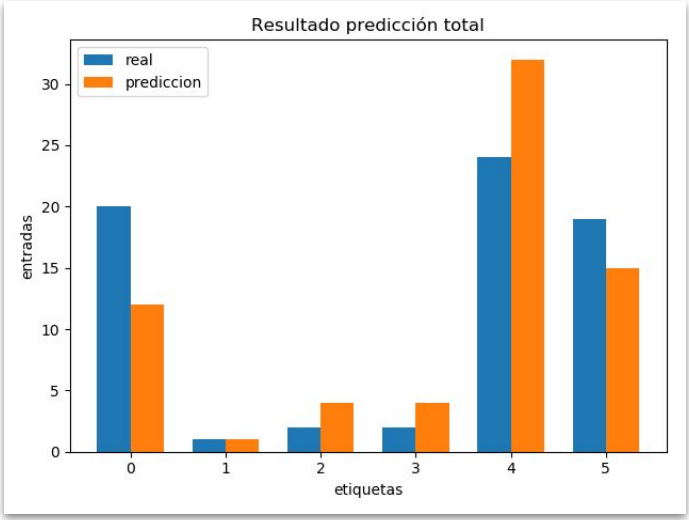
Índice

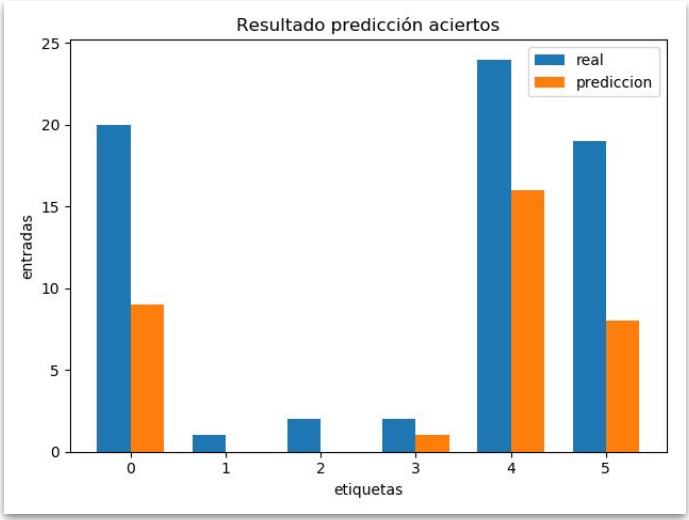
5. **Experimentos**
    a. Pregunta Q13
    b. Pregunta Q14
    c. Pregunta Q13+Q14
    d. Train Q13 - Test Q14
    e. Train Q14 - Test Q13

5.d. Train Q13 - Test Q14

	Accuracy	ROC	F1	Precision	Recall	Instancias	Características	Tiempo (s)
NB Multinomial	0.5	0.655	0.508	0.55	0.50	226	687	0.002
Caso Base	0.279	0.5	0.122	0.08	0.28	226	687	0

5.d. Train Q13 - Test Q14





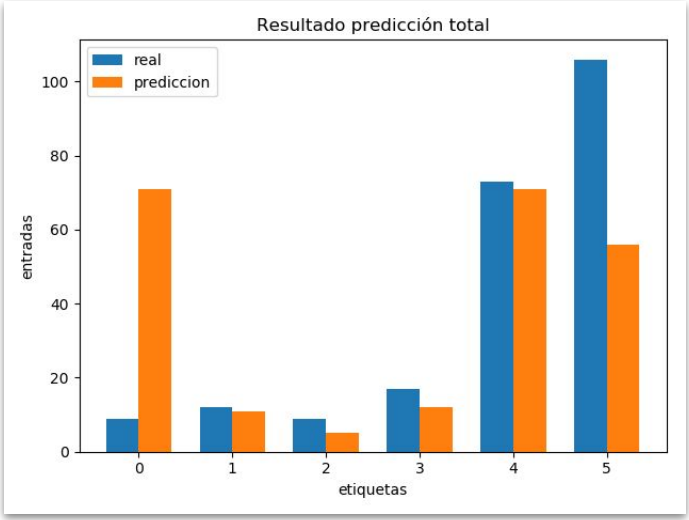
Índice

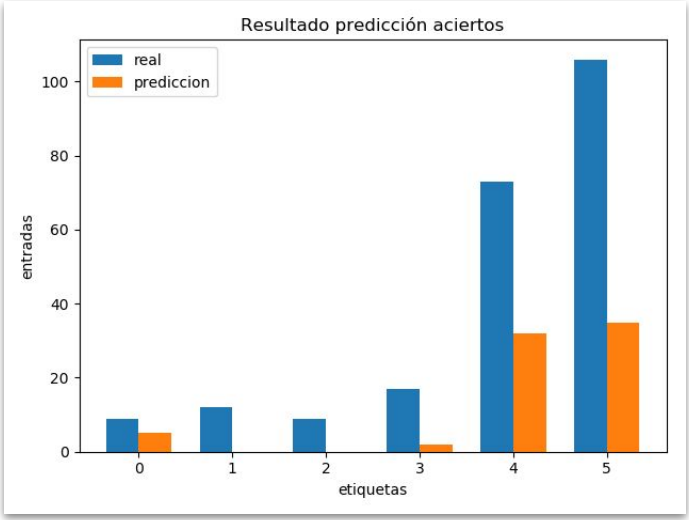
5. **Experimentos**
   a. Pregunta Q13
   b. Pregunta Q14
   c. Pregunta Q13+Q14
   d. Train Q13 - Test Q14
   e. Train Q14 - Test Q13

5.e. Train Q14 - Test Q13

	Accuracy	ROC	F1	Precision	Recall	Instancias	Características	Tiempo (s)
NB Complement	0.327	0.572	0.362	0.45	0.33	226	687	0.0006
Caso Base	0.469	0.5	0.299	0.22	0.47	226	687	0

5.e. Train Q14 - Test Q13





Índice

1. MOOC
2. Nuestro problema
3. Análisis
4. Diseño
5. Experimentos
6. **Conclusiones**

6. Conclusiones

Como conclusiones volvemos a los objetivos marcados en este trabajo/proyecto, con lo que hemos conseguido **estudiar el proceso para anotar datos lingüísticos y cómo hacerlo,** también hemos **estudiado los distintos tipos de métodos para evaluar la calidad e una anotación de datos**, hemos **desarrollado un clasificador de opiniones** sobre un curso específico de AbiertaUGR y hemos **visualizado los resultados de una forma fácil y comprensible** para su entendimiento.

Los objetivos marcados han sido superados aunque tenemos que tener en cuenta que debido a los datos de los que partíamos, y la cantidad de clases a seleccionar, podemos tener unos resultados que no pueden ser los deseados, no obstante, hemos conseguido aproximar, entender y clasificar las diferentes respuestas de los alumnos del curso.

Final - Cuestiones

Alumno: José Antonio Ruiz Millán

Tutora: Maria Victoria Luzón García

Tutor: Eugenio Martínez Cámara