Procesamiento de Lenguaje Natural

Eric S. Téllez Mario Graff

Tabla de contenidos

Prefacio

El curso trata de ser auto-contenido, es decir, no debería de ser necesario leer otras fuentes para poder entenderlo y realizar las actividades. De cualquier manera es importante comentar que el curso está basado en los siguientes libros de texto:

- Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Third Edition draft. Daniel Jurafsky and James H. Martin. pdf
- Introduction to machine learning, Third Edition. Ethem Alpaydin. MIT Press.
- An Introduction to Statistical Learning with Applications in R. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Springer Texts in Statistics.
- All of Statistics. A Concise Course in Statistical Inference. Larry Wasserman. MIT Press.
- An Introduction to the Bootstrap. Bradley Efron and Robert J. Tibshirani. Monographs on Statistics and Applied Probability 57. Springer-Science+Business Media.
- Understanding Machine Learning: From Theory to Algorithms. Shai Shalev-Shwartz and Shai Ben-David. Cambridge University Press.

Notación

La Tabla 1 muestra la notación que se seguirá en este documento.

Tabla 1: Notación

Símbolo	Significado
x	Variable usada comunmente como entrada
y	Variable usada comunmente como salida
$\mathbb{R}$	Números reales
$\mathbf{x}$	Vector Columna $\mathbf{x} \in \mathbb{R}^d$
d	Dimensión
$\mathbf{w} \cdot \mathbf{x}$	Producto punto donde $\mathbf{w}$ y $\mathbf{x} \in \mathbb{R}^d$
$\mathcal{D}$	Conjunto de datos
$\mathcal{T}$	Conjunto de entrenamiento
$\mathcal{V}$	Conjunto de validación
$\mathcal{G}$	Conjunto de prueba
N	Número de ejemplos
K	Número de clases
$\mathbb{P}(\cdot)$	Probabilidad
$\mathcal{X}, \mathcal{Y}$	Variables aleatorías
$\mathcal{N}(\mu, \sigma^2)$	Distribución Normal con parámetros μ y σ^2
$f_{\mathcal{X}}$	Función de densidad de probabilidad de $\mathcal{X}$
$\mathbb{1}(e)$	Función para indicar; 1 only if e is true
Ω	Espacio de búsqueda
$\mathbb{V}$	Varianza
$\mathbb{E}$	Esperanza

Licencia

1 Introducción

El **objetivo** de la unidad es

2 Manejando Texto

El **objetivo** de la unidad es

Paquetes usados

```
from microtc.params import OPTION_GROUP, OPTION_DELETE,\
                           OPTION_NONE
from microtc.textmodel import SKIP_SYMBOLS
from b4msa.textmodel import TextModel
from b4msa.lang_dependency import LangDependency
from nltk.stem.snowball import SnowballStemmer
import unicodedata
import re
```

2.1 Introducción

Se podría suponer que el texto se que se analizará está bien escrito y tiene un formato adecuado para su procesamiento. Desafortunadamente, la realidad es que en la mayoría de aplicaciones el texto que se analiza tiene errores de ortográficos, errores de formato y además no es trivial identificar la unidad mínima de procesamiento que podría ser de manera natural, en el español, las palabras. Por este motivo, esta unidad trata técnicas comunes que se utilizan para normalizar el texto, esta normalización es un proceso previo al desarrollo de los algoritmos de PLN.

La Figura 2.1 esquematiza el procedimiento que se presenta en esta unidad, la idea es que se un texto pasa primeramente a un proceso de normalización (Sección 2.2 y Sección 2.3), para después ser segmentado (ver Sección 2.4) y el resultado es lo que se utiliza para modelar el lenguaje.

Las normalizaciones y segmentaciones descritas en esta unidad se basan principalmente en las utilizadas en los siguientes artículos científicos.

Figura 2.1: Diagrama de Pre-procesamiento

1. An automated text categorization framework based on hyperparameter optimization (Tellez et al. (2018))
2. A simple approach to multilingual polarity classification in Twitter (Tellez, Miranda-Jiménez, Graff, Moctezuma, Suárez, et al. (2017))
3. A case study of Spanish text transformations for twitter sentiment analysis (Tellez, Miranda-Jiménez, Graff, Moctezuma, Siordia, et al. (2017))

2.2 Normalización de Texto Sintáctica

La descripción de las normalizaciones empieza presentando las que se puede aplicar a nivel de caracteres, sin la necesidad de conocer el significado de las palabras. También se agrupan en este conjunto aquellas transformaciones que se realizan mediante expresiones regulares o su búsqueda en una lista de palabras previamente definidas.

2.2.1 Entidades

La descripción de diferentes técnicas de normalización empieza con el manejo de entidades en el texto. Algunas entidades que se tratarán serán los nombres de usuario, números o URLs mencionados en un texto. Por otro lado están las acciones que se realizarán a las entidades encontradas, estas acciones corresponden a su borrado o remplazo por algún otro toquen.

2.2.1.1 Usuarios

En esta sección se trabajará con los nombres de usuarios que siguen el formato usado por Twitter. En un tuit, los nombres de usuarios son aquellas palabras que inician con el caracter @ y terminan con un espacio o caracter terminal. Las acciones que se realizarán con los nombres de usuario encontrados serán su borrado o reemplazo por una etiqueta en particular.

El procedimiento para encontrar los nombres de usuarios es mediante expresiones regulares, en particular se usa la expresión `@\S+`, tal y como se muestra en el siguiente ejemplo.

```
text = 'Hola @xx, @mm te está buscando'
re.sub(r"@\S+", "", text)
```

```
'Hola   te está buscando'
```

La segunda acción es reemplazar cada nombre de usuario por una etiqueta particular, en el siguiente ejemplo se reemplaza por la etiqueta `_usr`.

```
text = 'Hola @xx, @mm te está buscando'
re.sub(r"@\S+", "_usr", text)
```

```
'Hola _usr _usr te está buscando'
```

2.2.1.2 URL

Los ejemplos anteriores se pueden adaptar para manejar URL; solamente es necesario adecuar la expresión regular que identifica una URL. En el siguiente ejemplo se muestra como se pueden borrar las URLs que aparecen en un texto.

```
text = "puedes verificar que http://google.com esté funcionando"
re.sub(r"https?://\S+", "", text)
```

```
'puedes verificar que  esté funcionando'
```

2.2.1.3 Números

The previous code can be modified to deal with numbers and replace the number found with a shared label such as `_num`.

```
text = "acabamos de ganar 10 M"
re.sub(r"\d\d*\.?\d*|\d*\.\d\d*", "_num", text)
```

```
'acabamos de ganar _num M'
```

2.2.2 Ortografía

El siguiente bloque de normalizaciones agrupa aquellas modificaciones que se realizan a algún componente de tal manera que aunque impacta en su ortografía puede ser utilizado para reducir la dimensión y se ve reflejado en la complejidad del algoritmo.

2.2.2.1 Mayúsculas y Minúsculas

La primera de estas transformaciones es convertir todas los caracteres a minúsculas. Como se puede observar esta transformación hace que el vocabulario se reduzca, por ejemplo, las palabras *México* o *MÉXICO* son representados por la palabra *méxico*. Esta operación se puede realizar con la función `lower` tal y cómo se muestra a continuación.

```
text = "México"
text.lower()
```

```
'méxico'
```

2.2.2.2 Signos de Puntuación

Los signos de puntuación son necesarios para tareas como la generación de textos, pero existen otras aplicaciones donde los signos de puntuación tienen un efecto positivo en el rendimiento del algorithm, este es el caso de tareas de categorización de texto. El efecto que tiene el quitar los signos de puntuación es que el vocabulario se reduce. Los símbolos de puntuación se pueden remover teniendo una lista de los mismos, esta lista de signos de puntuación se encuentra en la variable `SKIP_SYMBOLS` y el siguiente código muestra un procedimiento para quitarlos.

```
text = "¡Hola! buenos días:"
output = ""
for x in text:
    if x in SKIP_SYMBOLS:
        continue
    output += x
output
```

```
'Hola buenos días'
```

2.2.2.3 Símbolos Diacríticos

Continuando con la misma idea de reducir el vocabulario, es común eliminar los símbolos diacríticos en las palabras. Esta transformación también tiene el objetivo de normalizar aquellos textos informales donde los símbolos diacríticos son usado con una menor frecuencia, en particular los acentos en el caso del español. Por ejemplo, es común encontrar la palabra *México* escrita como *Mexico*.

El siguiente código muestra un procedimiento para eliminar los símbolos diacríticos.

```
text = 'México'
output = ""
for x in unicodedata.normalize('NFD', text):
    o = ord(x)
    if 0x300 <= o and o <= 0x036F:
        continue
    output += x
output
```

```
'Mexico'
```

2.3 Normalización Semántica

Las siguientes normalizaciones comparten el objetivo con las normalizaciones presentadas hasta este momento, el cual es la reducción del vocabulario; la diferencia es que las siguientes utilizan el significado o uso de la palabra.

2.3.1 Palabras Comunes

Las palabras comunes (*stop words*) son palabras utilizadas frecuentemente en el lenguaje, las cuales son necesarias para comunicación, pero no aportan información para discriminar un texto de acuerdo a su significado.

The stop words are the most frequent words used in the language. These words are essential to communicate but are not so much on tasks where the aim is to discriminate texts according to their meaning.

Las palabras vacías se pueden guardar en un diccionario y el proceso de identificación consiste en buscar la existencia de la palabra en el diccionario. Una vez que la palabra analizada se encuentra en el diccionario, se procede a quitarla o cambiarla por un token particular. El proceso de borrado se muestra en el siguiente código.

```
lang = LangDependency('spanish')

text = '¡Buenos días! El día de hoy tendremos un día cálido.'
output = []
for word in text.split():
    if word.lower() in lang.stopwords[len(word)]:
        continue
    output.append(word)
```

```
output = " ".join(output)
output
```

```
'¡Buenos días! día hoy día cálido.'
```

2.3.2 Lematización y Reducción a la Raíz

La idea de lematización y reducción a la raíz (*stemming*) es transformar una palabra a su raíz mediante un proceso heurístico o morfológico. Por ejemplo, las palabras *jugando* o *jugaron* se transforman a la palabra *jugar*.

El siguiente código muestra el proceso de reducción a la raíz utilizando la clase `SnowballStemmer`.

```
stemmer = SnowballStemmer('spanish')

text = 'Estoy jugando futbol con mis amigos'
output = []
for word in text.split():
    w = stemmer.stem(word)
    output.append(w)
output = " ".join(output)
output
```

```
'estoy jug futbol con mis amig'
```

2.4 Segmentación

Una vez que el texto ha sido normalizado es necesario segmentarlo (*tokenize*) a sus componentes fundamentales, e.g., palabras o gramas de caracteres (q-grams) o de palabras (n-grams). Existen diferentes métodos para segmentar un texto, probablemente una de las más sencillas es asumir que una palabra está limitada entre dos espacios o signos de puntuación. Partiendo de las palabras encontradas se empiezan a generar los gramas de palabras, e.g., bigramas, o los gramas de caracteres si se desea solo generarlos a partir de las palabras.

2.4.1 Gramas de Palabras (n-grams)

El primer método de segmentación revisado es la creación de los gramas de palabras. El primer paso es encontrar las palabras las cuales se pueden encontrar mediante la función `split`; una vez que las palabras están definidas estás se pueden unir para generar los gramas de palabras del tamaño deseado, tal y como se muestra en el siguiente código.

```
text = 'Estoy jugando futbol con mis amigos'
words = text.split()
n = 3
n_grams = []
for a in zip(*[words[i:] for i in range(n)]):
    n_grams.append("~".join(a))
n_grams
```

```
['Estoy~jugando~futbol',
 'jugando~futbol~con',
 'futbol~con~mis',
 'con~mis~amigos']
```

2.4.2 Gramas de Caracteres (q-grams)

La segmentación de gramas de caracteres complementa los gramas de palabras. Los gramas de caracteres están definidos como la subcadena de longitud q. Este tipo de segmentación tiene la característica de que es agnóstica al lenguaje, es decir, se puede aplicar en cualquier idioma; contrastando, los gramas de palabras se pueden aplicar solo a los lenguajes que tienen definido el concepto de palabra, por ejemplo en el idioma chino las palabras no se pueden identificar como se pueden identificar en el español o inglés. La segunda característica importante es que ayuda en el problema de errores ortográficos, siguiendo una perspectiva de similitud aproximada.

El código para realizar los gramas de caracteres es similar a la presentada anteriormente, siendo la diferencia que el ciclo está por los caracteres en lugar de la palabras como se había realizado. El siguiente código muestra una implementación para realizar gramas de caracteres.

```
text = 'Estoy jugando'
q = 4
q_grams = []
for a in zip(*[text[i:] for i in range(q)]):
    q_grams.append("".join(a))
q_grams
```

```
['Esto',
 'stoy',
 'toy ',
 'oy j',
 'y ju',
 ' jug',
 'juga',
 'ugan',
 'gand',
 'ando']
```

2.5 TextModel

Habiendo descrito diferentes tipos de normalización (sintáctica y semántica) y el proceso de segmentación es momento para describir la librería B4MSA (Tellez, Miranda-Jiménez, Graff, Moctezuma, Suárez, et al. (2017)) que implementa estos procedimientos; específicamente, el punto de acceso de estos procedimientos corresponde a la clase `TextModel`. El método `TextModel.text_transformations` es el que realiza todos los métodos de normalización (Sección 2.2 y Sección 2.3) y el método `TextModel.tokenize` es el encargado de realizar la segmentación (Sección 2.4) siguiendo el flujo mostrado en la Figura 2.1.

2.5.1 Normalizaciones

El primer conjunto de parámetros que se describen son los que corresponden a las entidades (Sección 2.2.1). Estos parámetros tiene tres opciones, borrar (`OPTION_DELETE`), remplazar (`OPTION_GROUP`) o ignorar. Los nombres de los parámetros son:

- usr_option
- url_option
- num_option

que corresponden al procesamiento de usuarios, URL y números respectivamente. Adicionalmente, `TextModel` trata los emojis, hashtags y nombres, mediante los siguientes parámetros:

- emo_option
- hashtag_option
- ent_option

Por ejemplo, el siguiente código muestra como se borra el usuario y se reemplaza un hashtag; se puede observar que en la respuesta se cambian todos los espacios por el caracter `~` y se incluye ese mismo al inicio y final del texto.

```
tm = TextModel(hashtag_option=OPTION_GROUP,
               usr_option=OPTION_DELETE)
texto = 'mira @xyz estoy triste. #UnDiaLluvioso'
tm.text_transformations(texto)
```

```
'~mira~estoy~triste.~_htag~'
```

Siguiendo con las transformaciones sintácticas, toca el tiempo a describir aquellas que relacionadas a la ortografía (Sección 2.2.2) las cuales corresponden a la conversión a minúsculas, borrado de signos de puntuación y símbolos diacríticos. Estas normalizaciones se activan con los siguiente parámetros.

- lc
- del_punc
- del_diac

En el siguiente ejemplo se transforman el texto a minúscula y se remueven los signos de puntuación.

```
tm = TextModel(lc=True,
               del_punc=True,
               del_diac=False)
texto = 'Hoy está despejado.'
tm.text_transformations(texto)
```

```
'~hoy~está~despejado~'
```

Las normalizaciones semánticas (Sección 2.3) que se tienen implementadas en la librería corresponden al borrado de palabras comunes y reducción a la raíz; estás se pueden activar con los siguientes parámetros.

- stopwords
- stemming

Por ejemplo, las siguientes instrucciones quitan las palabras comunes y realizan una reducción a la raíz.

```
tm = TextModel(lang='es',
               stopwords=OPTION_DELETE,
               stemming=True)
texto = 'el clima es perfecto'
tm.text_transformations(texto)
```

```
'~clim~perfect~'
```

2.5.2 Segmentación

El paso final es describir el uso de la segmentación. La librería utiliza el parámetro `token_list` para indicar el tipo de segmentación que se desea realizar. El formato es una lista de número, donde el valor indica el tipo de segmentación. El número 1 indica que se realizará una segmentación por palabras, los número positivo corresponden a los gramas de caracteres y los números negativos a los gramas de palabras.

Por ejemplo, utilizando las normalizaciones que se tienen por defecto, el siguiente código segmenta utilizando gramas de caracteres de tamañan 4.

```
tm = TextModel(token_list=[4])
tm.tokenize('buenos días')
```

```
['q:~bue',
 'q:buen',
 'q:ueno',
 'q:enos',
 'q:nos~',
 'q:os~d',
 'q:s~di',
 'q:~dia',
 'q:dias',
 'q:ias~']
```

para poder identificar cuando se trata de un segmento que corresponde a una palabra o un grama de caracteres, a los últimos se les agrega el prefijo `q:`. Cabe mencionar que por defecto se remueven los símbolos diacríticos.

El ejemplo anterior, se utiliza para generar un grama de palabras de tamaño 2. Como se ha mencionado los gramas de palabras se especifican con números negativos siendo el valor absoluto el tamaño del grama.

```
tm = TextModel(token_list=[-2])
tm.tokenize('buenos días')
```

```
['buenos~dias']
```

Para completar la explicación, se combinan la segmentación de gramas de caracteres y palabras además de incluir las palabras en la segmentación.

```
tm = TextModel(token_list=[4, -2, -1])
tm.tokenize('buenos días')
```

```
['buenos~dias',
 'buenos',
 'dias',
 'q:~bue',
 'q:buen',
 'q:ueno',
 'q:enos',
 'q:nos~',
 'q:os~d',
 'q:s~di',
 'q:~dia',
 'q:dias',
 'q:ias~']
```

3 Modelado de Lenguaje

El **objetivo** de la unidad es

4 Fundamentos de Clasificación de Texto

El **objetivo** de la unidad es

Paquetes usados

```
from microtc.utils import tweet_iterator, load_model, save_model
from b4msa.textmodel import TextModel
from EvoMSA.tests.test_base import TWEETS
from EvoMSA.utils import bootstrap_confidence_interval
from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import recall_score, precision_score, f1_score
from sklearn.naive_bayes import MultinomialNB
from scipy.stats import norm, multinomial, multivariate_normal
from scipy.special import logsumexp
from collections import Counter
from matplotlib import pylab as plt
from os.path import join
import numpy as np
```

4.1 Introducción

El problema de categorización (clasificación) de texto es una tarea de PLN que desarrolla algoritmos capaces de identificar la categoría de un texto de un conjunto de categorías previamente definidas. Por ejemplo, en análisis de sentimientos pertenece a esta tarea y su objetivo es el detectar la polaridad (e.g., positiva, neutral, o negativa) del texto. Cabe mencionar, que diferentes tareas de PLN pueden ser formuladas como problemas de clasificación, e.g., la tarea de preguntas y respuestas, vinculación de enunciados, entre otras.

El problema de clasificación de texto se puede resolver desde diferentes perspectivas; el camino que se seguirá corresponde a aprendizaje supervisado. Los problemas de aprendizaje supervisado comienzan con un conjunto de pares, donde el primer elementos del par corresponde a las entradas (variables independientes) y el segundo es la respuesta (variable

dependiente). Sea $\mathcal{D} = \{(\text{texto}_i, y_i) \mid i = 1, \ldots, N\}$ donde $y \in \{c_1, \ldots c_K\}$ y texto_i contiene el texto.

4.2 Teorema de Bayes

Una manera de modelar este problema es modelando la probabilidad de observar la clase $\mathcal{Y}$ dada la entrada, es decir, $\mathbb{P}(\mathcal{Y} \mid \mathcal{X})$. El Teorema de Bayes ayuda a expresa esta expresión en términos de elementos que se pueden medir de un conjunto de entrenamiento.

La probabilidad conjunta se puede expresar como $\mathbb{P}(\mathcal{X}, \mathcal{Y})$, esta probabilidad es conmutativa por lo que $\mathbb{P}(\mathcal{X}, \mathcal{Y}) = \mathbb{P}(\mathcal{Y}, \mathcal{X})$. En este momento se puede utilizar la definición de **probabilidad condicional** que es $\mathbb{P}(\mathcal{Y}, \mathcal{X}) = \mathbb{P}(\mathcal{Y} \mid \mathcal{X})\mathbb{P}(\mathcal{X})$. Utilizando estas ecuaciones el **Teorema de Bayes** queda como

$$\mathbb{P}(\mathcal{Y} \mid \mathcal{X}) = \frac{\mathbb{P}(\mathcal{X} \mid \mathcal{Y})\mathbb{P}(\mathcal{Y})}{\mathbb{P}(\mathcal{X})}, \tag{4.1}$$

donde al término $\mathbb{P}(\mathcal{X} \mid \mathcal{Y})$ se le conoce como **verosimilitud**, $\mathbb{P}(\mathcal{Y})$ es la probabilidad **a priori** y $\mathbb{P}(\mathcal{X})$ es la **evidencia**.

Es importante mencionar que la evidencia se puede calcular mediante la probabilidad total, es decir:

$$\mathbb{P}(\mathcal{X}) = \sum_{y \in \mathcal{Y}} \mathbb{P}(\mathcal{X} \mid \mathcal{Y} = y)\mathbb{P}(\mathcal{Y} = y). \tag{4.2}$$

4.3 Modelado Probabilistico (Distribución Categórica)

Se inicia la descripción de clasificación de texto presentando un ejemplo sintético que ejemplifica los supuestos que se realizan en el modelo. La distribución categórica modela el evento de seleccionar K eventos, los cuales pueden estar codificados como caracteres. Si esta selección se realiza ℓ veces se cuenta con una secuencia de eventos representados por caracteres. Por ejemplo, los K eventos pueden ser representados por los caracteres w, x, y y z. Utilizando este proceso se puede utilizar para ejemplificar el proceso de asociar una secuencia a una clase, e.g., positiva o negativa.

El primer paso es seleccionar los parámetros de dos distribuciones tal y como se muestra en las siguientes primeras dos líneas. Cada distribución se asume que es la generadora de una clase. El segundo paso es tomar una muestra de cada distribución, en particular se toman 1000 muestras con el siguiente procedimiento. En cada iteración se toma una muestra de una distribución Gausiana ($\mathcal{N}(15, 3)$), la variable aleatoria se guarda en la variable `length`. Esta

Tabla 4.1: Conjunto generado de clasificación de texto

Texto	Clase
x w w y x z w x x w w w x	Positivo
x z x y w z z z w w w z z	Negativo
w w y z y w z x z z x x z y z z z y z	Positivo
y z x x y z y y y z z z y z z w z x z	Negativo

variable aleatoria representa la longitud de la secuencia. El tercer paso es sacar la muestra de las distribuciones categóricas definidas previamente. Las muestras son guardadas en la lista `D` junto con la clase a la que pertenece 0 y 1.

```
pos = multinomial(1, [0.20, 0.20, 0.35, 0.25])
neg = multinomial(1, [0.35, 0.20, 0.25, 0.20])
length = norm(loc=15, scale=3)
D = []
m = {k: chr(122 - k) for k in range(4)}
id2w = lambda x: " ".join([m[_] for _ in x.argmax(axis=1)])
for l in length.rvs(size=1000):
    D.append((id2w(pos.rvs(round(l))), 1))
    D.append((id2w(neg.rvs(round(l))), 0))
```

La Tabla 4.1 muestra los primeros cuatro ejemplos generados con el procedimiento anterior. La primera columna muestra la secuencia y asociada a cada secuencia se muestra la clase que corresponde a la secuencia.

El primer paso es encontrar la verosimilitud dado el conjunto de datos `D`. El siguiente código calcula la verosimilitud de la clase positiva.

```
D_pos = []
[D_pos.extend(data.split()) for data, k in D if k == 1]
words, l_pos = np.unique(D_pos, return_counts=True)
w2id = {v: k for k, v in enumerate(words)}
l_pos = l_pos / l_pos.sum()
l_pos
```

```
array([0.2487688 , 0.35664848, 0.195195  , 0.19938773])
```

Un procedimiento equivalente se puede realizar para obtener la verosimilitud de la clase negativa.

```
D_neg = []
[D_neg.extend(data.split()) for data, k in D if k == 0]
_, l_neg = np.unique(D_neg, return_counts=True)
l_neg = l_neg / l_neg.sum()
l_neg
```

```
array([0.20145082, 0.24923466, 0.19732464, 0.35198988])
```

La probabilidad a priori se puede calcular con la siguientes instrucciones.

```
_, priors = np.unique([k for _, k in D], return_counts=True)
N = priors.sum()
prior_pos = priors[1] / N
prior_neg = priors[0] / N
```

Una ves que se han identificador los parámetros, estos pueden ser utilizados para predecir la clase dada una secuencia. El primer paso es calcular la verosimilitud, e.g., $\mathbb{P}(\text{w w x z}|\ \mathcal{y})$. Se observa que la secuencia tiene se tiene que transformar en términos, esto se puede realizar con el método `split`. Después, los términos se convierten al identificador que corresponde al parámetro del token con el mapa `w2id`. Una vez que se identifica el índice se conoce el valor del parámetro, se calcula el producto (como o la suma si se hace todo en términos del logaritmo) y se regresa el valor de la verosimilitud.

```
def likelihood(params, txt):
    params = np.log(params)
    _ = [params[w2id[x]] for x in txt.split()]
    tot = sum(_)
    return np.exp(tot)
```

La verosimilitud se combina con la probabilidad a priori, con esta información se calcula la evidencia y para obtener la probabilidad a posteriori tanto para la clase positiva (`post_pos`) como para la negativa (`post_neg`). La clase corresponde a la etiqueta que presenta la máxima probabilidad, última línea (`hy`).

```
post_pos = [likelihood(l_pos, x) * prior_pos for x, _ in D]
post_neg = [likelihood(l_neg, x) * prior_neg for x, _ in D]
evidence = np.vstack([post_pos, post_neg]).sum(axis=0)
post_pos /= evidence
post_neg /= evidence
hy = np.where(post_pos >= post_neg, 1, 0)
```

4.3.1 Clasificador de Texto

En la sección anterior se trabajo desde la creación de un conjunto de datos sintético que fue generado mediante dos distribuciones Categóricas, donde a cada distribución se le asignó una clase, e.g., positiva o negativa. Esto permitió observar todas las partes de modelado, en la realidad se desconoce el procedimiento que genera los textos y el proceso de aprendizaje empieza con un conjunto de datos, en este ejemplo se utilizará un conjunto de datos de polaridad que tiene cuatro clases, negativo (N), neutral (N), ausencia de polaridad (NEU), y positivo (P).

> Conjunto de Datos
>
> Es pertinente mencionar que el conjunto de datos fue etiquetado usando un clasificador de texto y ninguna valoración humana fue realizada para verificar que las etiquetas sean correctas.
> Este conjunto se usa dentro de EvoMSA (Graff et al. (2020)) como conjunto de prueba para realizar pruebas unitarias.

El conjunto de datos se obtiene con la siguiente instrucción.

As can be observed, $\mathcal{D}$ is equivalent to the one used in the Categorical Distribution example. The difference is that sequence of letters is changed with a sentence. Nonetheless, a feasible approach is to obtain the tokens using the `split` method. Another approach is to retrieve the tokens using a Tokenizer, as covered in the Text Normalization Section.

The following code uses the `TextModel` class to tokenize the text using words as the tokenizer; the tokenized text is stored in the variable `D`.

```
tm = TextModel(token_list=[-1])
tok = tm.tokenize
D = [(tok(x), y) for x, y in D]
```

Before estimating the likelihood parameters, it is needed to encode the tokens using an index; by doing it, it is possible to store the parameters in an array and compute everything `numpy` operations. The following code encodes each token with a unique index; the mapping is in the dictionary `w2id`.

```
words = set()
[words.update(x) for x, y in D]
w2id = {v: k for k, v in enumerate(words)}
```

Previously, the classes have been represented using natural numbers. The positive class has been associated with the number 1, whereas the negative class with 0. However, in this dataset,

the classes are strings. It was decided to encode them as numbers to facilitate subsequent operations. The encoding process can be performed simultaneously with the estimation of the prior of each class. Please note that the priors are stored using the logarithm in the variable `priors`.

```
uniq_labels, priors = np.unique([k for _, k in D], return_counts=True)
priors = np.log(priors / priors.sum())
uniq_labels = {str(v): k for k, v in enumerate(uniq_labels)}
```

It is time to estimate the likelihood parameters for each of the classes. It is assumed that the data comes from a Categorical distribution and that each token is independent. The likelihood parameters can be stored in a matrix (variable `l_tokens`) with K rows, each row contains the parameters of the class, and the number of columns corresponds to the vocabulary's size. The first step is to calculate the frequency of each token per class which can be done with the following code.

```
l_tokens = np.zeros((len(uniq_labels), len(w2id)))
for x, y in D:
    w = l_tokens[uniq_labels[y]]
    cnt = Counter(x)
    for i, v in cnt.items():
        w[w2id[i]] += v
```

The next step is to normalize the frequency. However, before normalizing it, it is being used a Laplace smoothing with a value 0.1. Therefore, the constant 0.1 is added to all the matrix elements. The next step is to normalize (second line), and finally, the parameters are stored using the logarithm.

```
l_tokens += 0.1
l_tokens = l_tokens / np.atleast_2d(l_tokens.sum(axis=1)).T
l_tokens = np.log(l_tokens)
```

4.3.1.1 Prediction

Once all the parameters have been estimated, it is time to use the model to classify any text. The following function computes the posterior distribution. The first step is to tokenize the text (second line) and compute the frequency of each token in the text. The frequency stored in the dictionary `cnt` is converted into the vector `x` using the mapping function `w2id`. The final step is to compute the product of the likelihood and the prior. The product is computed in log-space; thus, this is done using the likelihood and the prior sum. The last step is to compute the evidence and normalize the result; the evidence is computed with the function `logsumexp`.

```
def posterior(txt):
    x = np.zeros(len(w2id))
    cnt = Counter(tm.tokenize(txt))
    for i, v in cnt.items():
        try:
            x[w2id[i]] += v
        except KeyError:
            continue
    _ = (x * l_tokens).sum(axis=1) + priors
    l = np.exp(_ - logsumexp(_))
    return l
```

The posterior function can predict all the text in $\mathcal{D}$; the predictions are used to compute the model's accuracy. In order to compute the accuracy, the classes in $\mathcal{D}$ need to be transformed using the nomenclature of the likelihood matrix and priors vector; this is done with the `uniq_labels` dictionary (second line).

```
hy = np.array([posterior(x).argmax() for x, _ in D])
y = np.array([uniq_labels[y] for _, y in D])
(y == hy).mean()
0.974
```

4.3.1.2 Training

Solving supervised learning problems requires two phases; one is the training phase, and the other is the prediction. The posterior function handles the later phase, and it is missing to organize the code described in a training function. The following code describes the training function; it requires the dataset's parameters and an instance of `TextModel`.

```
def training(D, tm):
    tok = tm.tokenize
    D =[(tok(x), y) for x, y in D]
    words = set()
    [words.update(x) for x, y in D]
    w2id = {v: k for k, v in enumerate(words)}
    uniq_labels, priors = np.unique([k for _, k in D], return_counts=True)
    priors = np.log(priors / priors.sum())
    uniq_labels = {str(v): k for k, v in enumerate(uniq_labels)}
    l_tokens = np.zeros((len(uniq_labels), len(w2id)))
    for x, y in D:
        w = l_tokens[uniq_labels[y]]
```

```
        cnt = Counter(x)
        for i, v in cnt.items():
            w[w2id[i]] += v
    l_tokens += 0.1
    l_tokens = l_tokens / np.atleast_2d(l_tokens.sum(axis=1)).T
    l_tokens = np.log(l_tokens)
    return w2id, uniq_labels, l_tokens, priors
```

4.4 Modelado Vectorial

xxx

5 Representación de Texto

El **objetivo** de la unidad es

Paquetes usados

```
from EvoMSA import BoW,\
                   DenseBoW
from microtc.utils import tweet_iterator
from wordcloud import WordCloud
import numpy as np
import pandas as pd
from matplotlib import pylab as plt
import seaborn as sns
```

5.1 Bolsa de Palabras Dispersa

La idea de una bolsa de palabras discretas es que después de haber normalizado y segmentado el texto (Capítulo 2), cada token t sea asociado a un vector único $\mathbf{v_t} \in \mathbb{R}^d$ donde la i-ésima componente, i.e., $\mathbf{v}_{\mathbf{t}_i}$, es diferente de cero y $\forall_{j \neq i} \mathbf{v}_{\mathbf{t}_j} = 0$. Es decir la i-ésima componente está asociada al token t, se podría pensar que si el vocabulario está ordenado de alguna manera, entonces el token t está en la posición i. Por otro lado el valor que contiene la componente se usa para representar alguna característica del token.

El conjunto de vectores $\mathbf{v}$ corresponde al vocabulario, teniendo d diferentes token en el mismo y por definición $\forall_{i \neq j} \mathbf{v_i} \cdot \mathbf{v_j} = 0$, donde $\mathbf{v_i} \in \mathbb{R}^d$, $\mathbf{v_j} \in \mathbb{R}^d$, y $(\cdot)$ es el producto punto. Cabe mencionar que cualquier token fuera del vocabulario es descartado.

Usando esta notación, un texto x está representado por una secuencia de términos, i.e., $(t_1, t_2, ...)$; la secuencia puede tener repeticiones es decir, $t_j = t_k$. Utilizando la característica de que cada token está asociado a un vector $\mathbf{v}$, se transforma la secuencia de términos a una secuencia de vectores (manteniendo las repeticiones), i.e., $(\mathbf{v}_{\mathbf{t_1}}, \mathbf{v}_{\mathbf{t_2}}, ...)$. Finalmente, el texto x se representa como:

$$\mathbf{x} = \frac{\sum_t \mathbf{v_t}}{\|\sum_t \mathbf{v_t}\|}, \tag{5.1}$$

donde la suma se hace para todos los elementos de la secuencia, $\mathbf{x} \in \mathbb{R}^d$, y $\|\mathbf{w}\|$ es la norma Euclideana del vector $\mathbf{w}$.

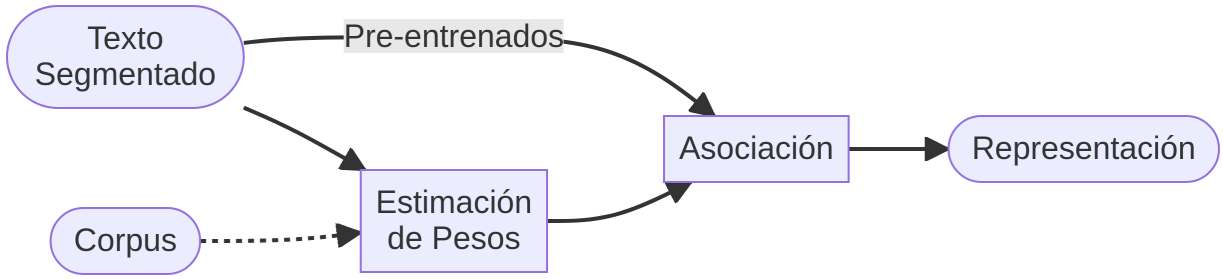


Figura 5.1: Diagrama Bolsa de Palabras Dispersa

Antes de iniciar la descripción detallada del proceso de representación utilizando una bolsa de palabras dispersas, es conveniente ilustrar este proceso mediante la Figura 5.1. El **texto segmentado** es el resultado del proceso ilustrado en Figura 2.1. El texto segmentado puede seguir dos caminos, en la parte superior se encuentra el caso cuando los pesos han sido identificados previamente y en la parte inferior es el procedimiento cuando los pesos se estiman mediante un corpus específico que normalmente es un conjunto de entrenamiento.

5.1.1 Pesado de Términos

Como se había mencionado el valor que tiene la componente i-ésima del vector $\mathbf{v_{t_i}}$ corresponde a una característica del término asociado, este procedimiento se le conoce como el **esquema de pesado**. Por ejemplo, si el valor es 1 (i.e., $\mathbf{v_{t_i}} = 1$) entonces el valor está indicando solo la presencia del término, este es el caso más simple. Considerando la Ecuación 5.1 se observa que el resultado, $\mathbf{x}$, cuenta las repeticiones de cada término, por esta característica a este esquema se le conoce como **frecuencia de términos** (*term frequency (TF)*).

Una medida que complementa la información que tiene la frecuencia de términos es el inverso de la frecuencia del término (*Inverse Document Frequency (IDF)*) en la colección, esta medida propuesta por Sparck Jones (1972) se usa en un método de pesado descrito por Salton y Yang (1973) el cual es conocido como **TFIDF**. Este método de pesado propone el considerar el producto de la frecuencia del término y el inverso de la frecuencia del término (*Inverse Document Frequency (IDF)*) en la colección como el peso del término.

5.1.2 Ejemplos

En los siguientes ejemplos se usa una bolsa de palabras con un pesado TFIDF pre-entrenada, los datos de esta bolsa de palabras se encuentra en el atributo `BoW.bow`. El tamaño del vocabulario es 131072, que está compuesto por palabras, gramas de palabras y caracteres. En el siguiente ejemplo se muestran los primeros tres gramas con sus respectivos valores TFIDF de la frase *Buen día*. Se puede observar que el `tm` regresa una lista de pares, donde la primera parte es el identificador del término, e.g., 11219 y el segundo es el valor TFIDF, e.g., 0.3984. La lista tiene un tamaño de 27 elementos, el resto de los 131072 componentes son cero dado que no se encuentran en el texto representado.

```
bow = BoW(lang='es')
tm = bow.bow
vec = tm['Buen día']
vec[:3]
```

```
[(11219, 0.3984336285263178),
 (11018, 0.3245843730253675),
 (24409, 0.2377856890280623)]
```

El uso del identificador del término se puede reemplazar por el término para poder visualizar mejor la representación del texto en el espacio vectorial. El diccionario que se encuentra en `BoW.names` hace la relación identificador a término. Se puede ver que el primer elemento del vector es el bigrama *buen~dia*, seguido por *buen* y el tercer término es *dia*. Los siguientes términos que no se muestran corresponden a gramas de caracteres. El valor TFIDF no indica la importancia del término, mientras mayor sea el valor, se considera más importante de acuerdo al TFIDF. En este ejemplo el bigrama tiene más importancia que las palabras y la palabra *buen* es más significativa que *dia*.

```
[(bow.names[k], v)
 for k, v in vec[:3]]
```

```
[('buen~dia', 0.3984336285263178),
 ('buen', 0.3245843730253675),
 ('dia', 0.2377856890280623)]
```

Con el objetivo de ilustrar una heurística que ha dado buenos resultados en el siguiente ejemplo se presentan las primeras cuatro componentes del texto *Buen día colegas*. Se puede observar como los valores de IDF de los términos comunes cambiaron, por ejemplo para el caso de *buen~dia* cambio de 0.3984 a 0.2486. Este es el resultado de que los valores están normalizados tal como se muestra en la Ecuación 5.1. Por otro lado, se observa que ahora el término más significativo es la palabra *colegas*.

```
txt = 'Buen día colegas'
[(tm.id2token[k], v)
 for k, v in tm[txt][:4]]
```

```
[('buen~dia', 0.24862785236357487),
 ('buen', 0.20254494048246244),
 ('dia', 0.1483814139998851),
 ('colegas', 0.3538047214393573)]
```

Una manera de visualizar la representación es creando una nube de palabras de los términos, donde el tamaño del termino corresponde al valor TFIDF. En la Figura 5.2 muestra la nube de palabras generada con los términos y sus respectivos valores IDF del texto *Es un placer estar platicando con ustedes.*

Figura 5.2: Nube de términos

El texto se representa en un espacio vectorial, entonces es posible comparar la similitud entre dos textos en esta representación, por ejemplo, en el siguiente ejemplo se compara la similitud coseno entre los textos *Es un placer estar platicando con ustedes.* y *La lluvia genera un caos en la ciudad.* El valor obtenido es cercano a cero indicando que estos textos no son similares.

```
txt1 = 'Es un placer estar platicando con ustedes.'
txt2 = 'La lluvia genera un caos en la ciudad.'
vec1 = tm[txt1]
vec2 = tm[txt2]
f = {k: v for k, v in vec1}
np.sum([f[k] * v for k, v in vec2 if k in f])
```

```
0.01645519294478695
```

Complementando el ejemplo anterior, en esta ocasión se comparan dos textos que comparten el concepto *plática*, estos son *Es un placer estar platicando con ustedes.* y *Estoy dando una platica en Morelia.* se puede observar que estos textos son más similares que los ejemplos anteriores.

```
txt1 = 'Es un placer estar platicando con ustedes.'
txt2 = 'Estoy dando una platica en Morelia.'
vec1 = tm[txt1]
vec2 = tm[txt2]
f = {k: v for k, v in vec1}
np.sum([f[k] * v for k, v in vec2 if k in f])
```

```
0.2035427118119315
```

Habiendo realizado la similitud entre algunos textos lleva a preguntarse cómo será la distribución de similitud entre varios textos, para poder contestar esta pregunta, se utilizarán los datos de Delitos, los cuales se guardan en la variable `D` tal y como se en las siguientes instrucciones.

```
fname = 'delitos/delitos_ingeotec_Es_train.json'
D = list(tweet_iterator(fname))
```

El primer paso es representar todos los textos en el espacio vectorial de la bolsa de palabras, lo cual se logra con el método `BoW.transform` (primera linea), el segundo paso es calcular la similitud entre todos los textos, como se muestra en la segunda linea.

```
X = tm.transform(D)
sim = np.dot(X, X.T)
```

La distribución de similitud se muestra en la Figura 5.3 se puede observar que las similitudes se encuentran concentradas cerca del cero, esto indica que la mayoría de los textos están distantes, esto es el resultado de la bolsa de palabras discreta que se enfoca en modelar las palabras y no el significado de las mismas.

5.2 Bolsa de Palabras Densa

La Figura 5.4 muestra el procedimiento que se sigue para representar un texto en una bolsa de palabras dispersa. En primer lugar la bolsa de palabras densa considera que los vectores asociados a los términos se encuentra pre-entrenados y en general no es factible entrenarlos en el momento, esto por el tiempo que lleva estimar estos vectores.

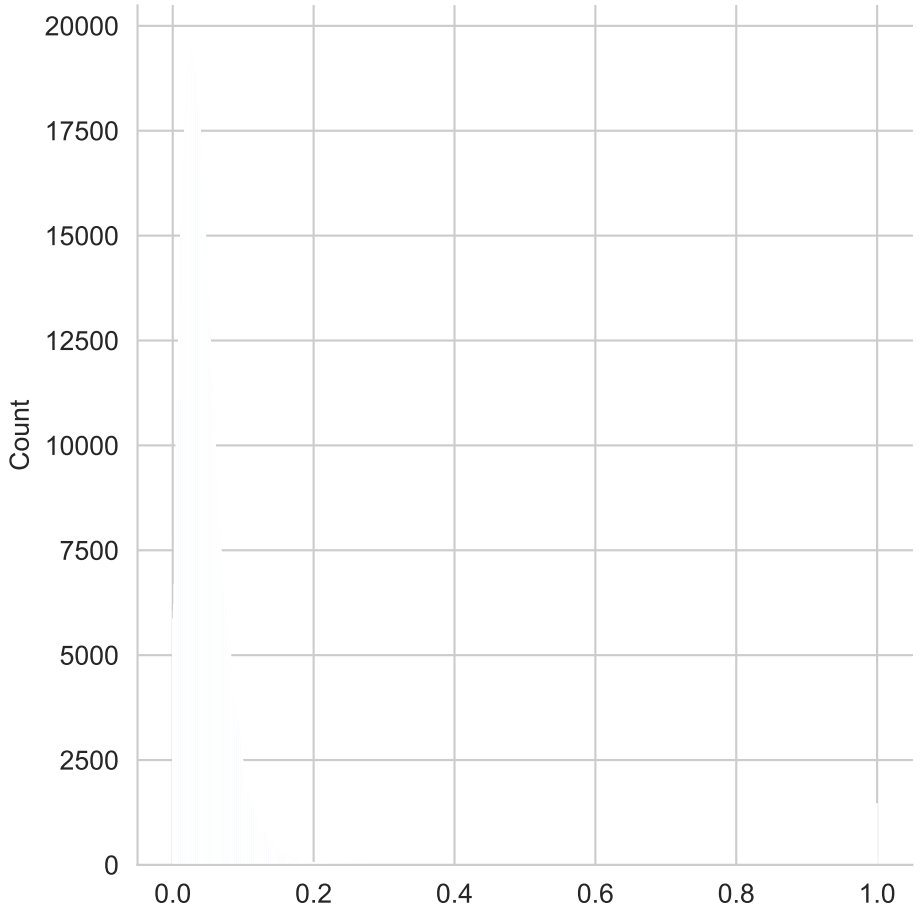


Figura 5.3: Histograma de la similitud

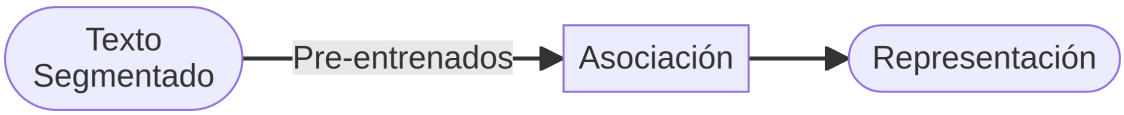


Figura 5.4: Diagrama Bolsa de Palabras Densa

El texto se representa como el vector $\mathbf{u}$ que se calcula usando la Ecuación 5.2 donde se observa que es la suma de los vectores asociados a cada término más un coeficiente $\mathbf{w_0}$. En particular el coeficiente $\mathbf{w_0} \in \mathbb{R}^M$ no se encuentra en todas las representaciones densas, pero en la representación que se usará contiene este vector, M es la dimensión de la representación densa.

$$\mathbf{u} = \sum_t \mathbf{u_t} + \mathbf{w_0}. \tag{5.2}$$

En vector $\mathbf{u_t}$ está asociado al término t, en particular este vector en la representación densa que se describirá está definido en términos de una bolsa de palabras dispersa (Ecuación 5.1) como se puede observar en la Ecuación 5.3

$$\mathbf{u_t} = \frac{\mathbf{W v_t}}{\|\sum_t \mathbf{v_t}\|}, \tag{5.3}$$

donde $\mathbf{W} \in \mathbb{R}^{M \times d}$ es la matriz que hace la proyección de la representación dispersa a la representación densa, se puede observar esa operación está normalizada con la norma Euclideana de la representación dispersa.

Combinando las Ecuación 5.2 y Ecuación 5.3 queda la

$$\begin{aligned} \mathbf{u_t} &= \sum_t \frac{\mathbf{W v_t}}{\|\sum_t \mathbf{v_t}\|} + \mathbf{w_0} \\ &= \mathbf{W} \frac{\sum_t \mathbf{v_t}}{\|\sum_t \mathbf{v_t}\|} + \mathbf{w_0}, \end{aligned}$$

donde se puede observar la representación dispersa (Ecuación 5.1), i.e., $\frac{\sum_t \mathbf{v_t}}{\|\sum_t \mathbf{v_t}\|}$ lo cual resulta en la Ecuación 5.4

$$\mathbf{u} = \mathbf{W x} + \mathbf{w_0}, \tag{5.4}$$

que representa un texto en el vector $\mathbf{u} \in \mathbb{R}^M$.

Para algunas representaciones densas, las componentes de la matriz de transformación $\mathcal{W}$ están asociadas a conceptos, en el caso que se analiza estas están asociadas a palabras claves o emojis.

5.2.1 Ejemplos

Continuando con los ejemplos presentados para la bolsa dispersa (Sección 5.1.2) en esta sección se hace el análisis con la representación de palabras densa. El primer paso es inicializar la clase que contiene las representaciones densas, esto se hace con la siguiente instrucción.

```
dense = DenseBoW(lang='es',
                 voc_size_exponent=15,
                 emoji=False, keyword=True,
                 distance_hyperplane=True,
                 dataset=False)
```

Para representar un texto en el espacio vectorial denso se utiliza el método `transform`, por ejemplo la siguiente instrucción representa el texto *Es un placer estar platicando con ustedes.* Solo se visualizan los valores de las primeras tres componentes.

```
txt1 = 'Es un placer estar platicando con ustedes.'
dense.transform([txt1])[0, :3]
```

```
array([-0.0042934 , -0.00429635, -0.00515905])
```

Lo primero que se observa es que los valores son negativos, a diferencia del caso disperso donde todos los valores son positivos. En este tipo de representación cada componente está asociada a una palabra las cuales se pueden conocer en el atributo `names`. El siguiente código muestra las tres primeras palabras asociadas al ejemplo anterior.

```
dense.names[:3]
```

```
['semanas', 'cuatro', 'piensa']
```

Siguiente la idea de utilizar una nube de palabras para visualizar el vector que representa el texto modelado, La Figura 5.5 muestra las nubes de palabras generada con las características y sus respectivos valores del texto *Es un placer estar platicando con ustedes.* Durante la generación de la nube de palabras se decidió representar genera una nube de palabras con las palabras con coeficiente negativo más significativo y aquellas con los coeficientes positivos más significativos. Se puede observar que las palabras positivas contienen componentes que están relacionados al enunciado, pero al mismo tiempo leyendo los términos positivos es complicado construir el texto representado. Adicionalmente las términos negativos que se observan en la nube de palabras en su mayoría son hashtags que tiene muy poca relación al texto representado.

Figura 5.5: Nube de características para el texto *Es un placer estar platicando con ustedes.*

Esta representación también permite comparación de similitud entre textos, en el siguiente ejemplo se calcula la similitud entre el texto *Es un placer estar platicando con ustedes.* y los textos *La lluvia genera un caos en la ciudad.* y *Estoy dando una platica en Morelia.* tal y como se hizo para la representación dispersa. Se puede observar que existe una mayor similitud entre los textos que contienen el concepto **plática**, lo cual es equivalente a lo que se observó en el ejemplo con bolsa de palabras discretas, pero los valores son significativamente mayores que en ese caso.

```
txt1 = 'Es un placer estar platicando con ustedes.'
txt2 = 'La lluvia genera un caos en la ciudad.'
txt3 = 'Estoy dando una platica en Morelia.'
X = dense.transform([txt1, txt2, txt3])
np.dot(X[0], X[1]), np.dot(X[0], X[2])
```

```
(0.7728943423183761, 0.8721107462230386)
```

Los valores de similitud entre los enunciados anteriores, se puede visualizar en una nube de palabras, utilizando solo las características positivas. La Figura 5.6 muestra las nubes de palabras generadas, en ellas es complicado comprender la razón por la cual la frases que tiene el concepto *plática* están más cercanas, es probable que la cola de la distribución, es decir, las palabras menos significativas son las que acercan las dos oraciones.

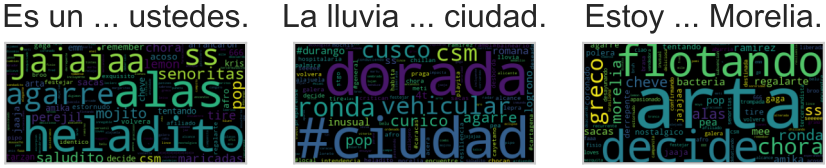


Figura 5.6: Nube de características positivas.

Al igual que en el caso disperso se puede calcular la distribución de similitud. Las siguientes instrucciones calcula la similitud coseno entre todos los ejemplos del conjunto de entrenamiento ($\mathcal{T}$).

```
X = dense.transform(D)
sim = np.dot(X, X.T)
```

La Figura 5.7 muestra el histograma de las similitudes calculada mediante la bolsa densa. Aquí se puede observar que la gran mayoría de los ejemplos tiene una similitud mayor y tiene una desviación estándar mayor que la vista en la Figura 5.3.

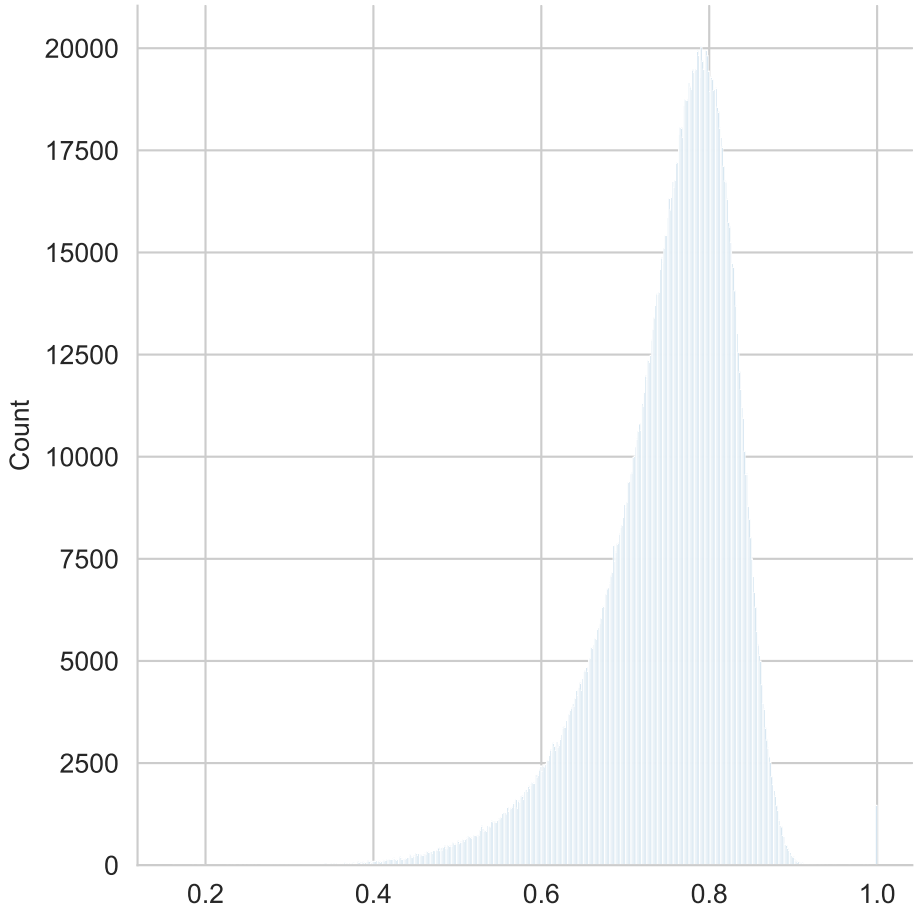


Figura 5.7: Histograma de la similitud usando bolsa de palabras densas

6 Clasificación de Texto

El **objetivo** de la unidad es

Paquetes usados

```
from EvoMSA import BoW,\
                   DenseBoW,\
                   StackGeneralization
from microtc.utils import tweet_iterator
from IngeoML import CI, SelectFromModelCV
from sklearn.metrics import f1_score,\
                            recall_score,\
                            precision_score
from wordcloud import WordCloud
import numpy as np
import pandas as pd
from matplotlib import pylab as plt
import seaborn as sns
```

6.1 Introducción

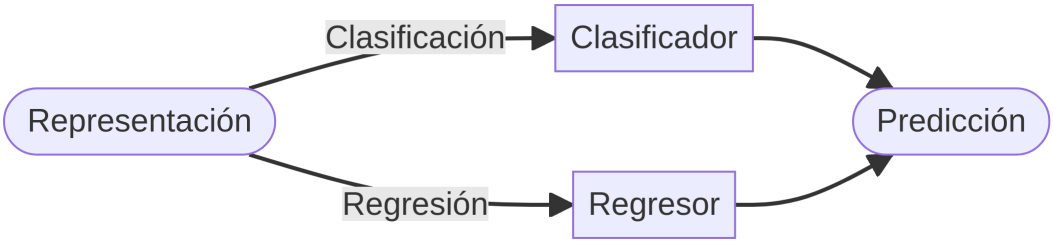


Figura 6.1: Diagrama de predicción

El conjunto de datos se puede conseguir en la página de Delitos aunque en esta dirección es necesario poblar los textos dado que solamente se encuentra el identificador del Tweet.

Para leer los datos del conjunto de entrenamiento y prueba se utilizan las siguientes instrucciones. En la variable `D` se tiene los datos que se utilizarán para entrenar el clasificador basado en la bolsa de palabras y en `Dtest` los datos del conjunto de prueba, que son usados para medir el rendimiento del clasificador.

```
fname = 'delitos/delitos_ingeotec_Es_train.json'
fname_test = 'delitos/delitos_ingeotec_Es_test.json'
D = list(tweet_iterator(fname))
Dtest = list(tweet_iterator(fname_test))
```

En la siguiente instrucción se observa el primer elemento del conjunto de entrenamiento. Se puede observar que en el campo `text` se encuentra el texto, el campo `klass` representa la etiqueta o clase, donde 0 representa la clase negativa y 1 la clase positiva, es decir, la presencia de un delito. El campo `id` es el identificador del Tweet y `annotations` son las clases dadas por los etiquetadores a ese ejemplo.

```
D[81]
```

```
{'annotations': [0, 0, 0],
 'id': 1107040319986696195,
 'klass': 0,
 'text': 'To loco'}
```

6.2 Bolsa de Palabras Dispersa

Se inicia con la creación de un clasificador basado en una bolsa de palabras dispersa, el clasificador es una máquina de soporte vectorial lineal (`LinearSVC`). La siguiente instrucción usa la clase `BoW` para crear este clasificador de texto. El primer paso es seleccionar el lenguaje, en este caso español (es) y después se entrena usando el método `fit`.

```
bow = BoW(lang='es').fit(D)
```

Habiendo entrenado el clasificador de texto es momento de utilizarlo para predecir, las siguientes dos instrucciones muestra el uso de la instancia `bow` para predecir clase del texto *me golpearon y robaron la bicicleta en la noche.* Se puede observar que la clase es 1, lo cual indica que el texto menciona la ejecución de un delito.

```
txt = 'me golpearon y robaron la bicicleta en la noche'
bow.predict([txt])
```

```
array([1])
```

El método `predict` recibe una lista de textos a predecir, en la siguiente instrucción se predicen todas las clases del conjunto de prueba (`Dtest`), la predicciones se guardar en la variable `hy_bow`.

```
hy_bow = bow.predict(Dtest)
```

Habiendo realizado la predicciones en el conjunto de prueba ($\mathcal{D}$), es momento de utilizar estas para medir el rendimiento, en esta ocasión se mide el valor f_1 para cada clase. El primer valor (0.9461) corresponde a la medida f_1 en la clase negativa y el segundo (0.7460) corresponde al valor en la clase positiva.

```
y = np.r_[[x['klass'] for x in Dtest]]
f1_score(y, hy_bow, average=None)
```

```
array([0.94612795, 0.74603175])
```

Con el objetivo de conocer la variabilidad del rendimiento del clasificador en este conjunto de datos, la siguientes instrucciones calcula el intervalo de confianza; para realizarlo se utiliza la clase `CI` la cual recibe la estadística a calcular, en este caso medida f_1. El siguiente paso es llamar a la clase con las entradas para calcular el intervalo, estas corresponden a las mediciones y predicciones del conjunto de prueba.

```
ci = CI(statistic=lambda y, hy: f1_score(y, hy,
                                         average=None))
ci_izq, ci_der = ci(y, hy_bow)
```

El intervalo izquierdo es $[0.9263, 0.6503]$ y el derecho tiene los valores $[0.9640, 0.8326]$. Para complementar la información del intervalo de confianza, la Figura 6.2 muestra el histograma y la densidad estimada para calcular el intervalo de confianza. Se ve que la varianza en la clase negativa es menor además de que tiene un rendimiento mejor que en la clase positiva.

Una manera de poder analizar el comportamiento del clasificador de texto implementado es visualizar en una nube de palabras las características que tienen el mayor peso en la decisión. Esto se realiza en las siguientes instrucciones siendo el primer paso obtener los coeficientes de la máquina de soporte vectorial lineal, los cuales se guardan en la variable `ws`. El segundo componente es el valor de IDF que tiene cada uno de los términos, esto se encuentran en el atributo `BoW.weights` tal y como se muestra en la segunda instrucción del siguiente código.

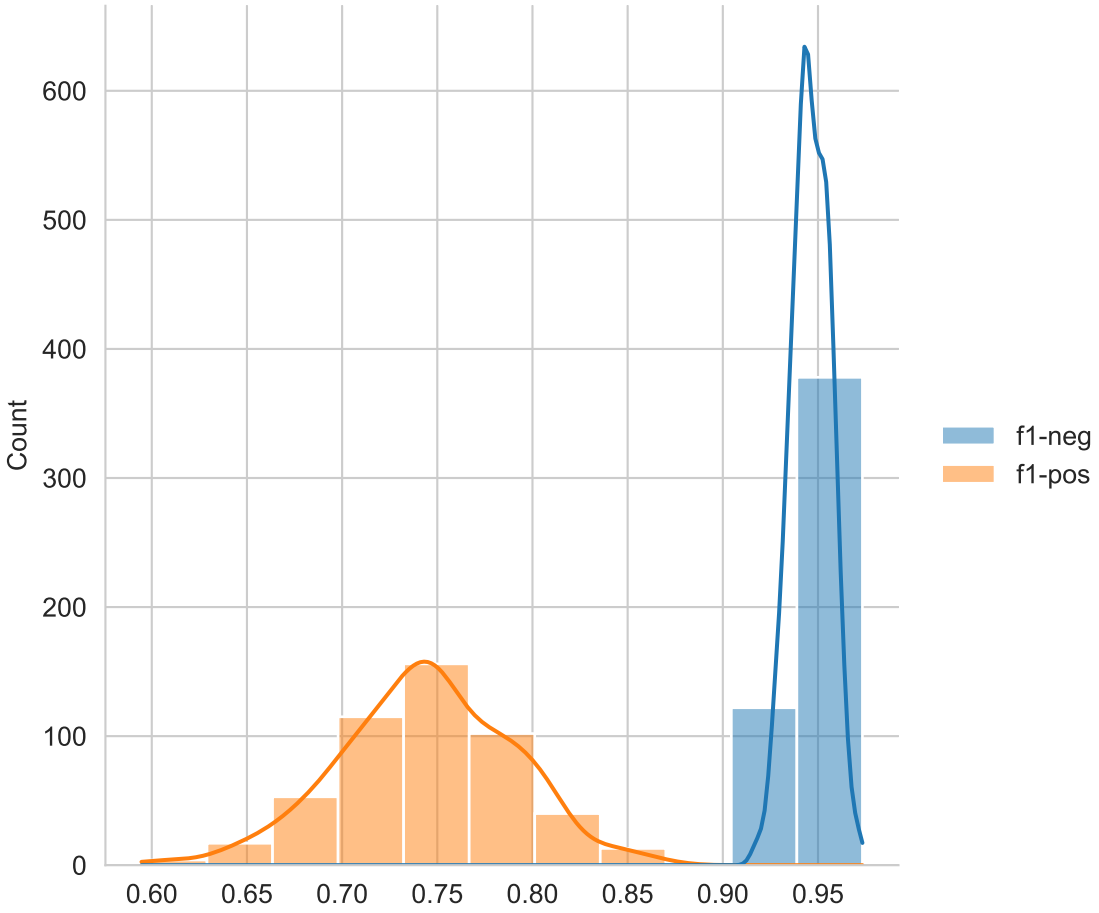


Figura 6.2: Histograma de f1 por clase

```
ws = bow.estimator_instance.coef_[0]
idfs = bow.weights
```

Teniendo los valores del clasificador y del IDF, solamente es necesario obtener su producto y separar los términos positivos de los negativos, tal y como se muestra en el siguiente código. La Figura 6.3 muestra las nubes de palabras generadas, cabe mencionar que aquellos términos que tienen como prefijo `q:` corresponden a q-gramas de caracteres y los términos que tienen el caracter `~` corresponden a bigramas de palabras.

```
tokens_pos = {name: w * idf
              for name, idf, w in zip(bow.names,
                                      idfs, ws)
              if w > 0}
tokens_neg = {name: w * idf * -1
              for name, idf, w in zip(bow.names,
                                      idfs, ws)
              if w < 0}
```

En la figura se puede observar que las características más importantes corresponden a la presencia de *asesinan*, *asesinan a*, *tiroteo* entre otras y por el lado de palabras relacionadas a la clase negativa se observan *secuestrado*, *asesina*, *un asesino*, entre otras. En Sección 6.4 se muestran ejemplos que ayudan a comprender el hecho de que la palabra *asesinan* sea considerada como positiva y por otro lado *asesina* se en la clase negativa.

Figura 6.3: Nubes de términos positivos y negativos

6.3 Bolsa de Palabras Densas

Es ahora el turno de hacer el análisis del clasificador de texto que se basado en bolsa de palabras densas. La siguiente instrucción muestra el uso de este clasificador donde se usa una bolsa de palabras con un vocabulario de 2^{15} (parámetro `voc_size_exponent=15`) y además se seleccionan como representaciones aquellos entrenados en los conjuntos de emojis (`emoji=True`)

y las palabras claves (`keyword=True`). En esta caso, los parámetros del clasificador no son estimados, es decir, no se llama al método `fit`. Esto es porque en este ejemplo se van a seleccionar aquellas representaciones que mejor representan al problema de Delitos utilizando una máquina de soporte vectorial lineal.

```
dense = DenseBoW(lang='es',
                 voc_size_exponent=15,
                 emoji=True, keyword=True,
                 dataset=False)
```

Para seleccionar las características que mejor representan al problema de delitos se utiliza la clase `SelectFromModelCV` la cual usa los coeficientes de la máquina de soporte vectorial para seleccionar las características más representativas, estas corresponden aquellas que tienen los coeficientes más grandes tomando su valor absoluto. La selección se realiza llamando al método `DenseBoW.select` con los parámetros que se observan en las siguientes instrucciones. En particular `SelectFromModelCV` es un método supervisado entonces se utilizarán las clase del conjunto de entrenamiento, y para poder medir el rendimiento de cada conjunto de características seleccionadas se usa una validación cruzada. La última instrucción estima los valores del clasificador con las características seleccionadas.

```
macro_f1 = lambda y, hy: f1_score(y, hy, average='macro')
kwargs = dense.estimator_kwargs
estimator = dense.estimator_class(**kwargs)
kwargs = dict(estimator=estimator,
              scoring=macro_f1)
dense.select(D=D,
             feature_selection=SelectFromModelCV,
             feature_selection_kwargs=kwargs)
dense.fit(D)
```

Como se mencionó la clase `SelectFromModelCV` selecciona aquellas características que mejor rendimiento dan, la clase mantiene los valores estimados en cada selección, las siguientes instrucciones ejemplifican como obtener los valores de rendimiento en las selecciones. La variable `perf` es una diccionario donde la llave es el número de características y el valor es el rendimiento correspondiente. La Figura 6.4 muestra es rendimiento se puede observar la dinámica donde con un poco menos de 1000 características se tiene un valor de rendimiento cercano a 0.9.

```
select = dense.feature_selection
perf = select.cv_results_
```

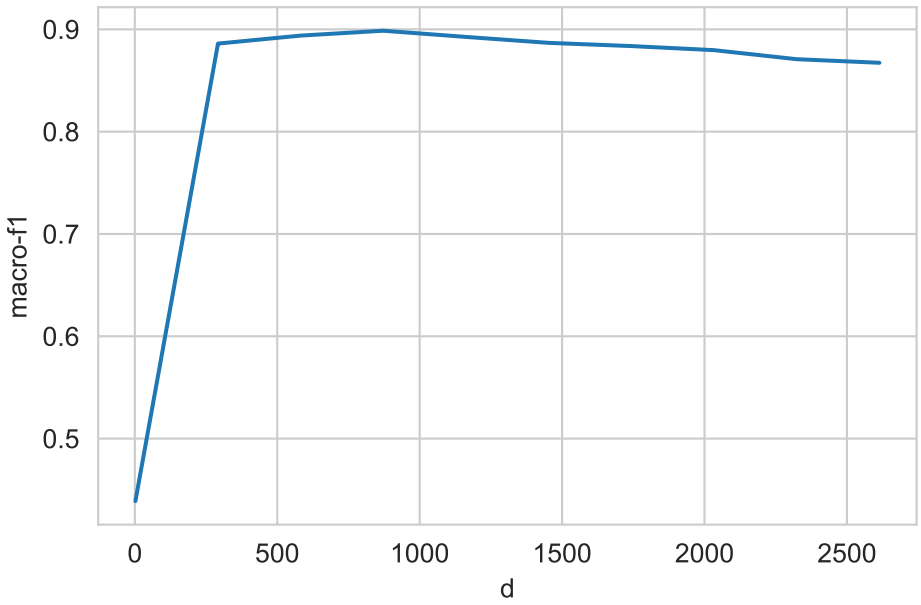


Figura 6.4: Rendimiento Variando el Número de Características

Después de haber seleccionado el número de características, se utiliza un código equivalente al usado en `BoW` para predecir las clases del conjunto de prueba ($\mathcal{G}$), tal y como se muestra en la siguiente instrucción.

```
hy_dense = dense.predict(Dtest)
```

El rendimiento en f_1 del clasificador basado en una bolsa se muestra con el siguiente código. Este valor puntual se complementa con la Figura 6.5 donde se muestra la distribución de esta medición y se compara con la obtenida con el clasificador de bolsa de palabras dispersa (i.e., `bow`).

```
f1_score(y, hy_dense, average=None)
```

```
array([0.94158076, 0.75362319])
```

En un clasificador basado en palabras densas también se puede comprender su comportamiento mostrando aquellas características que tiene un mayor peso al momento de decidir la clase. En las siguientes instrucciones se agrupan las características positivas y las negativas, utilizando el valor estimado por la máquina de soporte vectorial lineal (`w`). Considerando que cada característica está asociada a una palabra o emoji, entonces se pueden visualizar mediante una nube de palabras.

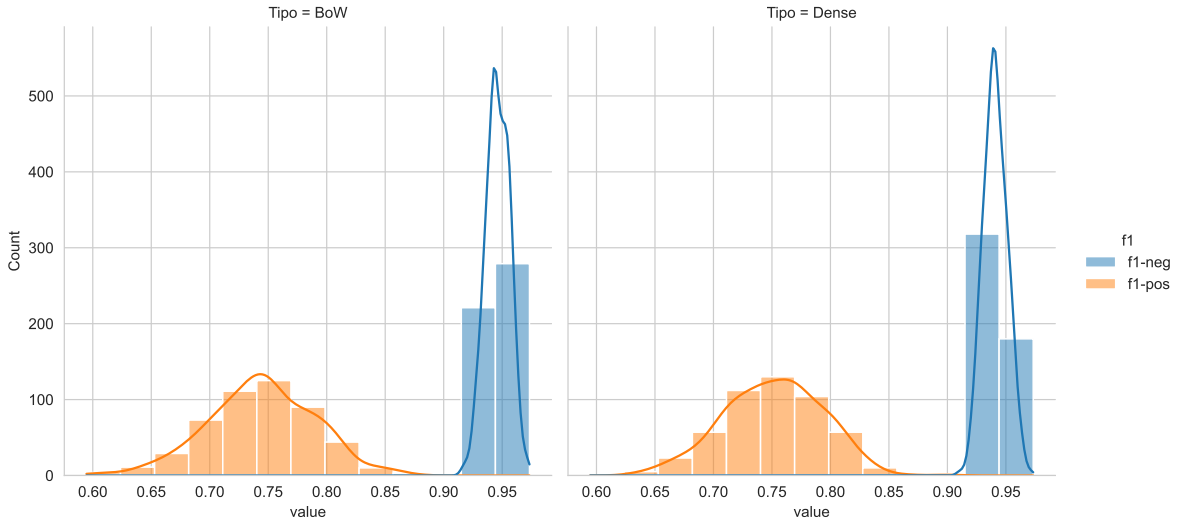


Figura 6.5: Histogramas de f1 por clase

```
w = dense.estimator_instance.coef_[0]
names = np.array(dense.names)
carac_pos = {k: v for k, v in zip(names, w) if v > 0}
carac_neg = {k: v * -1 for k, v in zip(names, w) if v < 0}
```

La Figura 6.6 muestra las nubes de palabras de las características positivas y negativas, se puede observar que una característica significativa de la clase positiva corresponde al modelo **robo**, **muere**, entre otros y de la clase negativa se observa **comentando**, **ocurrir*, entre otras.

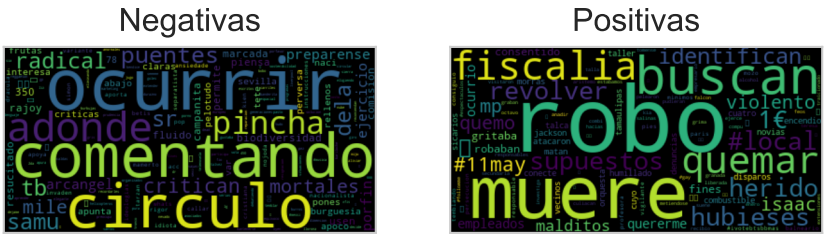


Figura 6.6: Nube de características positivas y negativas

6.4 Análisis Mediante Ejemplos

Hasta el momento se ha presentado un análisis global de los clasificadores dispersos (`bow`) y densos (`dense`), en esta sección se especializa el análisis al nivel de ejemplos. Lo primero que se realiza es ver el valor de la función de decisión, el signo de este valor indica la clase, un valor positivo indica clase positiva y el signo negativo corresponde a la clase negativa. El valor absoluto de la función indica de manera proporcional la distancia que existe al hiperplano que define las clases. Dado que se está utilizando este valor para contrastar el comportamiento de los algoritmos entonces la distancia entre el ejemplo al hiperplano está dado por la función de decisión dividida entre la norma de los coeficientes. La primera línea calcula la norma de los coeficientes estimados tanto para el clasificador disperso (`bow_norm`) y el denso (`dense_norm`)

```
bow_norm = np.linalg.norm(bow.estimator_instance.coef_[0])
dense_norm = np.linalg.norm(dense.estimator_instance.coef_[0])
```

Con las normas se procederá a calcular la función de decisión para el ejemplo *Asesinan a persona en Jalisco*, este ejemplo es positivo dado que menciona la ocurrencia de un delito. En las siguientes instrucciones se calcula la distancia al hiperplano la cual se puede observar que es positiva indicando que el texto es positivo.

```
array([[0.03104444]])
```

Complementando la distancia del clasificador disperso se presenta la distancia del clasificador denso en el siguiente código. También se puede observar que su valor es positivo, pero este se encuentre más cercano al hiperplano de decisión, lo cual indica que existe una mayor incertidumbre en su clase.

```
array([[0.00906055]])
```

Realizando el mismo procedimiento pero para texto *La asesina vivía en Jalisco.* Lo primero que se debe de notar es que el texto es negativo dado que se menciona que existe una asesina, pero el texto no indica que se haya cometido algún delito, esta fue una de las reglas que se siguió para etiquetar los textos tanto del conjunto de entrenamiento ($\mathcal{T}$) como del conjunto de prueba ($\mathcal{G}$). Pero es evidente que el texto anterior y el actual son sintácticamente muy similares, pero con una semántica diferente.

El siguiente código predice la función de decisión del clasificador disperso, la distancia es el valor absoluto del número presentado y el signo indica el lado del hiperplano, se observa que es negativo, entonces el clasificador indica que pertenece a la clase negativa.

```
array([[-0.03643014]])
```

El mismo procedimiento se realiza para el clasificador denso como se indica a continuación, obteniendo también un valor negativo y con una magnitud similar al encontrado por el clasificador disperso.

```
array([[-0.03598119]])
```

Continuando con el análisis, se puede visualizar los coeficientes más significativos para realizar la predicción. Por ejemplo, las siguientes instrucciones muestran los 5 coeficientes más significativos para predecir el texto *Asesinan a persona en Jalisco.*

```
[('asesinan', 0.21959432035885817),
 ('asesinan~a', 0.2082828806362095),
 ('q:sina', 0.14511790633072927),
 ('q:n~a~', 0.0838802284104354),
 ('q:an~a', 0.07344378043319356)]
```

Un procedimiento equivalente se realiza para el clasificador denso, tal y como se muestra en el siguiente código.

```
[('ocurrir', 0.06683457750173856),
 ('muere', 0.053880044741064774),
 ('consiguio', -0.05168798790090579),
 ('critican', -0.04459207389659752),
 ('hubieses', -0.04426955144485894)]
```

La Figura 6.7 muestra la nube de palabras de los términos y características más significativas para la predicción del ejemplo positivo (*Asesinan a persona en Jalisco*). La nube de palabras para el ejemplo negativo (*La asesina vivía en Jalisco*) se muestra en la Figura 6.8. La nube de palabras está codificada de la siguiente manera, las palabras que corresponden a la clase positiva está en mayúsculas y las de la clase negativa en minúsculas. Por ejemplo, en Figura 6.7 se observa que la palabra *asesinan* es relevante para la clasificación del ejemplo así como la característica **ocurrir**.

En el caso del ejemplo negativo, la Figura 6.8 muestra q-gramas de caracteres asociados a la clase positiva y también es evidente la palabra *asesina*. En el caso del clasificador denso también se observan características positivas como **ocurrir** y características negativas como **critican** y **empleados**.

Complementando los ejemplos anteriores, la Figura 6.9 muestra la nube de palabras obtenidas al calcular la función de decisión del texto *Le acaban de robar la bicicleta a mi hijo.* Se observa que este texto corresponde a la clase positiva y la función de decisión normalizada del clasificador disperso es -0.0335 y del clasificador denso corresponde a -0.0798. Ambas

Figura 6.7: Nube de características ejemplo positivo

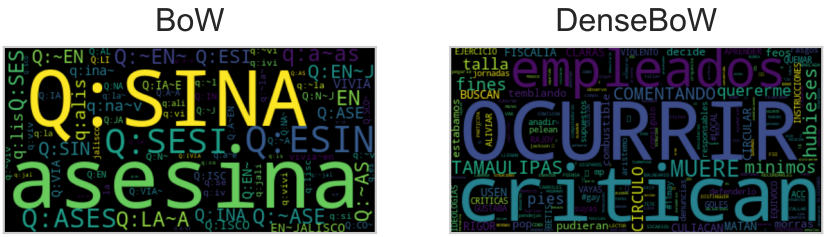


Figura 6.8: Nube de características ejemplo negativo

funciones de decisión indican que la clase es negativa, lo cual es un error. La figura muestra que los q-gramas de caracteres y las características positivas dominan las nubes, pero estas no tienen el peso suficiente para realizar una predicción correcta.

Figura 6.9: Nube de características en un ejemplo positivo con predicción negativa

6.5 Combinando Modelos

La siguiente pregunta es conocer si los modelos anteriores se pueden combinar para realizar una mejor predicción. En esta sección se utiliza la técnica de Stack Generalization (Wolpert (1992), Graff et al. (2020)) para combinar los dos modelos. La siguiente linea entrena el clasificador, el cual recibe como parámetros los clasificador a juntar.

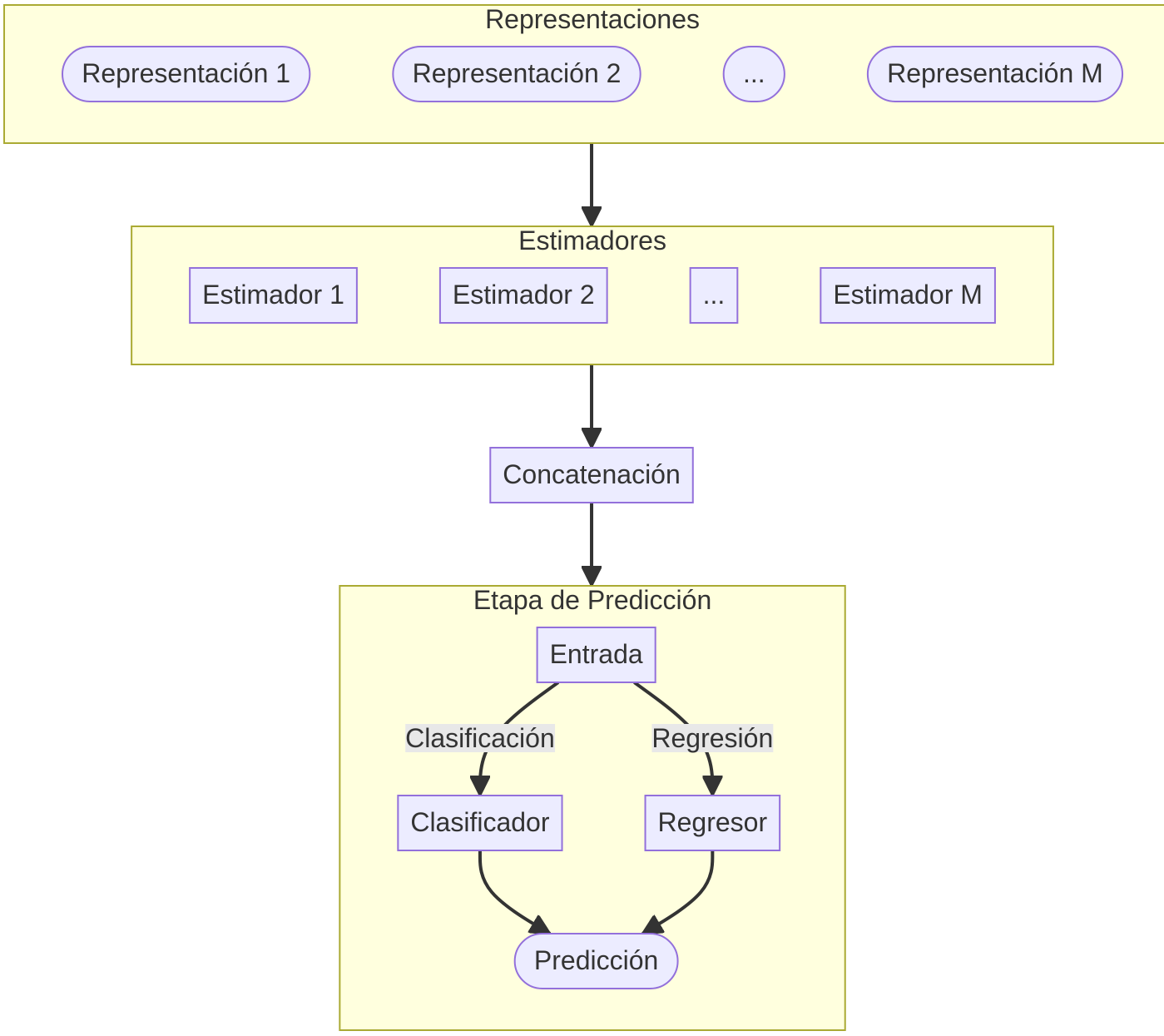


Figura 6.10: Diagrama de predicción en mezcla de modelos

Tabla 6.1: Rendimiento

	Recall neg	Recall pos	Precision neg	Precision pos
`bow`	0.9894	0.6184	0.9065	0.9400
`dense`	0.9648	0.6842	0.9195	0.8387
`stack`	0.9613	0.7500	0.9349	0.8382

Tabla 6.3: Rendimiento

	f1 neg	f1 pos	macro-f1
`bow`	0.9461	0.7460	0.8461
`dense`	0.9416	0.7536	0.8476
`stack`	0.9479	0.7917	0.8698

```
stack = StackGeneralization([bow, dense]).fit(D)
```

Siguiendo el procedimiento de los clasificadores dispersos y densos, la siguiente linea predice la clase de los ejemplos del conjunto de prueba y calcula su rendimiento en términos de la medida f_1.

```
hy_stack = stack.predict(Dtest)
f1_score(y, hy_stack, average=None)
```

```
array([0.94791667, 0.79166667])
```

Para poder comparar el rendimiento de los tres clasificadores desarrollados, la Tabla 6.1 presenta el rendimiento con las medidas recall y precision en las dos clases. Se puede observar que el `bow` tiene el mejor recall en la clase negativa y mejor precision en la clase positiva. Por otro lado el mejor recall en la clase positiva y precision en la clase negativa lo tiene `stack`.

Con respecto al rendimiento en términos de f_1, la Tabla 6.3 presenta la información con respecto a cada clase y la última columna contiene el macro-f_1. Los valores indican que en la clase positiva el mejor valor corresponde a `stack` lo cual se ve reflejado en el macro-f_1. El algoritmo de Stack Generalization nos indica que se hizo una mejora en la predicción de la clase positiva y la clase negativa se mantuvo constante al menos con respecto de la medida f_1.

7 Tareas de Clasificación de Texto

El **objetivo** de la unidad es

8 Bases de Conocimiento

El **objetivo** de la unidad es

9 Visualización

El **objetivo** de la unidad es

Paquetes usados

```
from EvoMSA import DenseBoW
from microtc.utils import tweet_iterator
from IngeoML import SelectFromModelCV
from sklearn.metrics import recall_score
from wordcloud import WordCloud
import numpy as np
import pandas as pd
import umap
import textwrap
from matplotlib import pylab as plt
import seaborn as sns
import plotly.express as px
```

9.1 Introducción

El conjunto de datos se puede conseguir en la página de Delitos aunque en esta dirección es necesario poblar los textos dado que solamente se encuentra el identificador del Tweet.

Para leer los datos del conjunto de entrenamiento y prueba se utilizan las siguientes instrucciones. En la variable `D` se tiene los datos que se utilizarán para entrenar el clasificador basado en la bolsa de palabras y en `Dtest` los datos del conjunto de prueba, que son usados para medir el rendimiento del clasificador.

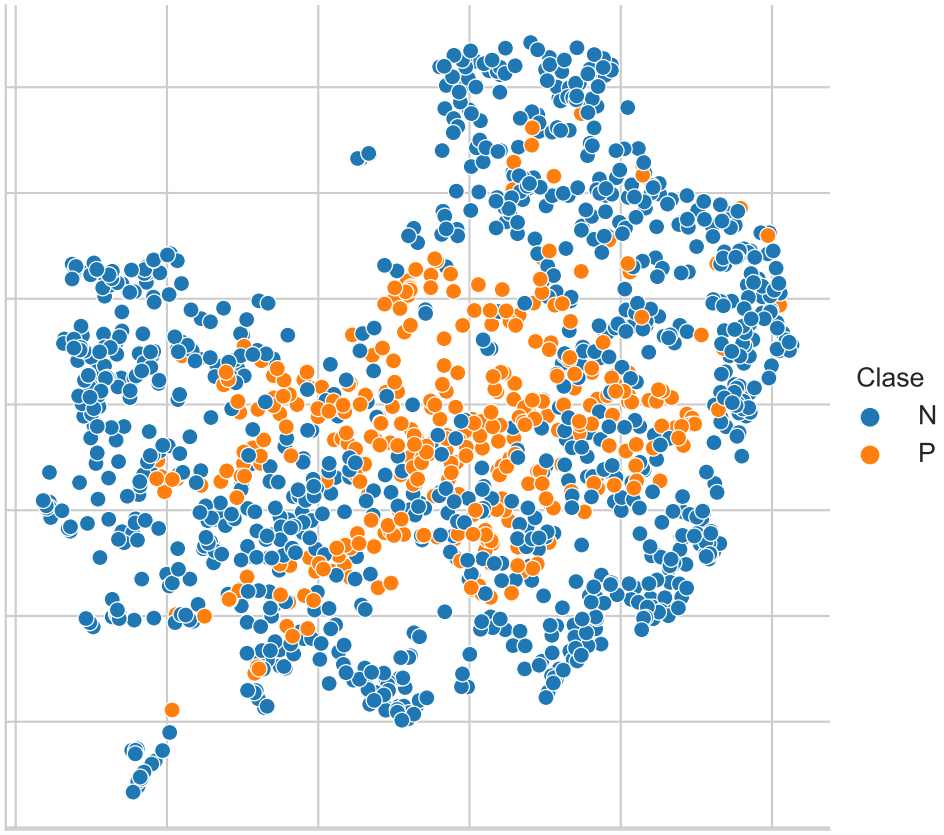
```
fname = 'delitos/delitos_ingeotec_Es_train.json'
fname_test = 'delitos/delitos_ingeotec_Es_test.json'
D = list(tweet_iterator(fname))
Dtest = list(tweet_iterator(fname_test))
```

9.2 Representación

```
dense = DenseBoW(lang='es', dataset=False,
                 emoji=True, keyword=True,
                 voc_size_exponent=15,
                 estimator_kwargs=dict(dual='auto'))
```

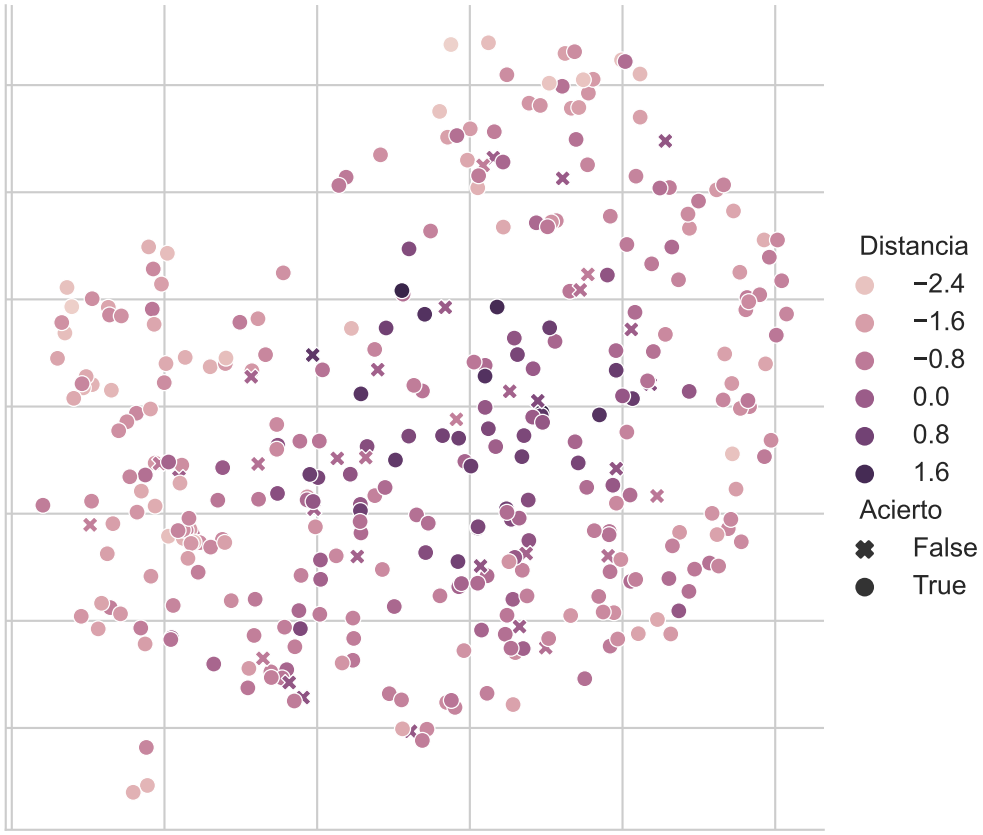
9.3 Proyección con UMAP

```
reducer = umap.UMAP(n_neighbors=5)
low_dim = reducer.fit_transform(X_dense)
```



```
dense.fit(D)
X_dense = dense.transform(Dtest)
```

```
df_dis = dense.decision_function(Dtest).flatten()
```



10 Conclusiones

El **objetivo** de la unidad es

Referencias

Graff, Mario, Sabino Miranda-Jiménez, Eric S. Tellez, y Daniela Moctezuma. 2020. «EvoMSA: A Multilingual Evolutionary Approach for Sentiment Analysis». *Computational Intelligence Magazine* 15: 76-88. https://ieeexplore.ieee.org/document/8956106.

Salton, Gerard, y Chungshu S. Yang. 1973. «On the specification of term values in automatic indexing». *Journal of Documentation* 29 (abril): 351-72. https://doi.org/10.1108/EB026562.

Sparck Jones, Karen. 1972. «A statistical interpretation of term specificity and its application in retrieval». *Journal of Documentation* 28: 11-21. https://doi.org/10.1108/EB026526.

Tellez, Eric S., Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Oscar S. Siordia, y Elio A. Villaseñor. 2017. «A case study of Spanish text transformations for twitter sentiment analysis». *Expert Systems with Applications* 81: 457-71. https://doi.org/https://doi.org/10.1016/j.eswa.2017.03.071.

Tellez, Eric S., Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Ranyart R. Suárez, y Oscar S. Siordia. 2017. «A Simple Approach to Multilingual Polarity Classification in Twitter». *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2017.05.024.

Tellez, Eric S., Daniela Moctezuma, Sabino Miranda-Jiménez, y Mario Graff. 2018. «An automated text categorization framework based on hyperparameter optimization». *Knowledge-Based Systems* 149: 110-23. https://doi.org/10.1016/j.knosys.2018.03.003.

Wolpert, David H. 1992. «Stacked generalization». *Neural Networks* 5 (2): 241-59. https://doi.org/10.1016/S0893-6080(05)80023-1.