# Proyecto

Integrantes: María Paula Herrero, Rosana Vargas

Una organización de Derechos Humanos (DDHH) requiere identificar noticias relacionadas a violaciones de DDHH que sean publicadas en los sitios web de los periódicos nacionales de Venezuela, con el fin de usar dicha información como base para escribir sus informes sobre la situación de los DDHH en Venezuela (trimestrales, anuales).

La idea es automatizar este proceso y entregar vía correo electrónico a los integrantes de la organización el grupo de noticias seleccionados en formato pdf.

**Objetivo**: Identificar si una noticia de un periódico publicado en línea (web) pertenece a una categoría de derechos humanos en particular

**Fuente de datos**: URL de noticias seleccionadas por un experto (Raúl Cubas) durante tres meses y tabla de clasificación de dichas noticas.

**Tarea de minería de datos**: clasificación

**Categorización de noticias de DDHH**

**Motivación principal** = posibilidad de realizar una clasificación automática de noticias según la nomenclatura de clasificación de DDHH que se usa en la Organización, en función de los temas abordados en el texto.

Este es un trabajo que se hace a diarios e implica revisar muchos sitios web, es

Importante para llevar a cabo una asignación adecuada de clasificaciones.

**Objetivo**: Realizar, de manera automática, una categorización de noticias según la clasificación de DDHH.

**Fuente de datos**: URL de noticias seleccionadas por un experto (Raúl Cubas) durante tres meses y tabla de clasificación de dichas noticas.

**Tarea de minería de datos**: clasificación

(1) Recolección de datos:

El corpus está constituido por noticias categorizadas según clasificación de DDHH para un período de 4 meses en el año 2018 publicado en los periódicos El Nacional y El Universal.

|  |  |  |
| --- | --- | --- |
| **CATEGORIA** | **DESCRIPCION** | **TOTAL** |
| 1.1 | Derecho a la vida | 15 |
| 1.2 | Derecho a la libertad personal | 39 |
| 1.3 | Derecho a la Integridad personal | 6 |
| 1.7 | Derecho a la libertad de expresión e información | 3 |
| 1.10 | Derecho a la justicia | 8 |
| 2.1 | Derecho a la Salud | 25 |
| 2.2 | Derecho a la Educación | 9 |
| 2.3 | Derecho al trabajo | 44 |
| 2.6 | Derecho a la seguridad social | 5 |
| 2.8 | Derecho a la Vivienda | 22 |
| 2.9 | Derecho de los Pueblos Indios | 3 |
| 2.10 | Derecho a la alimentación | 10 |
| 3.1 | Derecho a la tierra | 3 |
| 3.2 | Derecho al ambiente | 12 |
| 4.3 | Derecho a manifestar pacíficamente | 1 |
| 5 | Actividades de las ONG nacionales e internacionales de ddhh en Venezuela | 9 |
| 12 | Participación | 6 |
| 15 | Provea en prensa | 1 |
| 18 | Instancias internacionales de ddhh | 22 |
| CONTEXTO | CONTEXTO | 22 |
|  | Total: | 265 |

Un dato posee los siguientes atributos:

|  |  |
| --- | --- |
| **Atributo** | **Significado** |
| id | Identificador único |
| idPeriodico | Identificador único de la noticia en el periódico |
| periodico | Periódico (EN: El Nacional, EU: El Universidad) |
| url | Dirección URL donde se consigue la noticia en la Web |
| fecha | Fecha en formato dd/mm/AAAA |
| fechaPublicacion | Fecha de publicación del periódico |
| autor | Nombre del autor de existir, si no aparece NO\_TIENE |
| seccion | Sección del periódico a la que pertenece la noticia |
| palabrasClaves | Palabras claves con la que el periódico clasifica la noticia (EU no usa palabras claves) |
| titular | Titular de la noticia |
| subtitular | Sumario o subtitular de la notica (párrafo destacado) |
| imagen | URL de la imagen asociada a la noticia |
| contenido | Todos los párrafos del texto de la noticia |
| derecho | Identificador del DDHH principal asociado a esta noticia. Ver lista de identificadores |
| otrosDerechos | Otros identificadores de DDHH asociados a la misma noticia, separados por comas de existir, si no está vacío |
| subDerechos | Identificadores de las subcategorías del DDHH principal asociado de existir, si no está vacío |
| EP | SI: indica que la noticia ocurre En Protestas (EP)  No: indica que la noticia no ocurre en protestas |

De estas noticias se consideran el titular, el subtitular, el contenido y el derecho

También se considera para una segunda clasificación EP, dado que es necesario categorizar por derecho y EP.

No se consideran los atributos: id, idPeriodico, url, fecha, fechaPublicacion, imagen; por que funcionan como constantes

No se consideran los atributos otrosDerechos y subDerechos, porque solo vamos a clasificar según Derecho principal

No se consideran los atributos Periódico, autor, sección, las palabras claves; porque no aporta información para realizar la clasificación pero sirven para procesar la notica.

(2) Preparación de datos:

* Se revisaron los documentos PDF, se extrajo los datos necesarios y se creó el archivo CSV a usar en el proceso de Minería.
  + Usando Python se creó un archivo csv con los siguientes datos de cada archivo PDF entregado:
    - No.
    - Nombre archivo
    - Fecha
    - Titular
    - Periodico
    - Fecha Periodico
    - URL
    - Derecho PPal
    - Otros Derechos
    - SubDerechos
    - EP

Este script de Python, toma cada uno de los PDFs del directorio “toda” y del nombre de cada archivo obtiene el derecho principal (que es el primero que aparece), los otros derechos y el subderecho. También extrae la fecha y el periódico. Si el primer derecho que aparece no tiene dos niveles, se extrae y se pone como subderecho el original. Además, si aparecen la sigla EP en el nombre se coloca el atributo EP en SI, en caso contrario se coloca NO. Por último, del cuerpo del archivo pdf se obtiene el URL y el titular.

* + Usando Python, a partir del archivo CVS anterior, se buscó el contenido de la noticia a través del URL y se crea un segundo archivo CSV con la siguiente información:
    - id
    - idPeriodico
    - periodico
    - url
    - fecha
    - fechaPublicacion
    - autor
    - sección
    - palabrasClaves
    - titular
    - subtitular
    - imagen
    - contenido
    - derecho
    - otrosDerechos
    - subDerechos
    - EP

Usando el campo URL, el script se conecta al sitio web del periódico y se trae el contenido de la noticia. De ahí obtiene los datos del idPeriódico, el periódico como tal (ya no usamos el del nombre del pdf dado que puede estar incorrecto), la fecha de publicación, el autor, la sección donde el periódico lo clasifica, las palabras claves que el periódico le coloca (si no tiene se coloca NO\_EXISTE), el titular, subtitular, el url de la imagen, el contenido. Los últimos campos son los mismos que el csv anterior (derecho, otrosDerechos, subDrechos, EP)

* + Una vez generado el csv con todos los datos, se abre este archivo con un editor de texto, se cambian las dobles comillas ("") por (\"), se abre en rapidMiner, indicando que los campos se separan con punto y coma (;) y que no tome comentario (#).
* Separación de objetos o tokens (tokenization)
  + Los tokens identifican partes en las cuales se puede dividir un texto.
  + Pueden ser de diferente granularidad, dependiendo de la aplicación (párrafos, frases, n-gramas, palabras, ...)
* Eliminación de palabras no informativas (stop words) y signos de puntuación:
  + Construcción un **diccionario de palabras frecuentes en el español** que no aportan información para la tarea de clasificación de textos (artículos, adjetivos, pronombres, entre otros)
  + Estas palabras fueron eliminadas de los documentos aplicando un proceso de comparación con el diccionario
  + Se realizaron pruebas eliminando palabras que no nos parecían tener sentido en la clasificación (como Nicolás, maduro, Colombia, etc.), pero no vario el rendimiento, por lo tanto se dejaron.
* Lematización (stemming): aplicación del algoritmo **Porter Stemmer** para el español
  + Formas morfológicas de una misma palabra tienen el mismo significado.
  + El efecto de la lematización es una reducción de la dimensionalidad (al reducir un conjunto de palabras a una sola representación)
* Generación de vectores: se obtuvo un total de **5.063** términos (lexicón)
* Reducción de la dimensionalidad:
  + Se aplicaron diferentes técnicas de selección de atributos con el fin de identificar los términos más informativos para la tarea de categorización. Se estudiaron los resultados de Ganancia de Información, Chi-Cuadrado, Razón de Ganancia y Correlación, decidiendo que la mejor selección de palabras la hacía Ganancia de Información según la opinión del experto, seguido por el de Correlación.
  + Para definir el número de características finales se elaboró un sistema de votación eligiendo aquellos atributos que tenían un peso igual o mayor a 0,095 según el criterio de Ganancia de Información.
  + La cantidad de atributos obtenidos luego de evaluar el sistema de votación fueron 226.
  + Sin embargo, cuando se probó este grupo de datos con el algoritmo Bayes ingenuo, luego de varias pruebas se concluyó que el mejor peso era el 0,03
* Balancear las clases
  + Se eliminaron las categorías con muy poca muestra dejando solamente la siguiente data:

|  |  |  |
| --- | --- | --- |
| **CATEGORIA** | **DESCRIPCION** | **TOTAL** |
| 1.1 | Derecho a la vida | 15 |
| 1.2 | Derecho a la libertad personal | 39 |
| 2.1 | Derecho a la Salud | 25 |
| 2.3 | Derecho al trabajo | 44 |
| 2.8 | Derecho a la Vivienda | 22 |
| 2.10 | Derecho a la alimentación | 10 |
| 3.2 | Derecho al ambiente | 12 |
| 18 | Instancias internacionales de ddhh | 22 |
| CONTEXTO | CONTEXTO | 22 |
|  | Total: | 211 |

* + Luego se probó quitando las categorías con menos muestra (1.1, 2.10 y 3.2), pero esto no mejoró el modelo mucho.
  + También se probó quitando las categorías 18 y CONTEXTO, esto sí represento una mejora para el modelo. Evaluando con el experto llegamos a la conclusión que ambas son categorías más generales, mientras que las demás son solo derechos, por eso la diferencia. Sin embargo, se decidió no eliminarlas para realizar una mejor categorización cuando se aplica a la información extraída directamente de los sitios web.
  + Por último, para mejorar el balance de muestra se agregaron más casos a las clases con menos datos, esto se realizó porque no se obtuvieron buenos resultados con el algoritmo de sobre muestreó Smote, donde la generación del modelo mejoro, logrando una Exactitud = 80.12% y G-media = 74,23% pero a la hora de evaluar los datos de prueba la matriz de confusión y la medidas resultantes de las clases sobre muestreadas no representaron una mejore en la clasificación al obtenido sin el uso del Smote. A continuación se muestran los resultados:

**Pruebas Sin Smote (Exactitud = 51,06%, G-media = 48,63%):**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **true 2.1** | **true 18** | **true 1.2** | **true 2.3** | **true CONTEXTO** | **true 2.10** | **true 2.8** | **true 3.2** | **true 1.1** | **class precision** |
| **pred. 2.1** | 6 | 0 | 0 | 3 | 4 | 5 | 2 | 3 | 0 | **26,09%** |
| **pred. 18** | 1 | 4 | 6 | 0 | 8 | 1 | 0 | 2 | 0 | **18,18%** |
| **pred. 1.2** | 0 | 0 | 14 | 3 | 7 | 0 | 0 | 0 | 0 | **58,33%** |
| **pred. 2.3** | 0 | 0 | 1 | 13 | 1 | 1 | 0 | 0 | 0 | **81,25%** |
| **pred. CONTEXTO** | 1 | 1 | 2 | 0 | 6 | 0 | 0 | 0 | 1 | **54,55%** |
| **pred. 2.10** | 0 | 0 | 0 | 2 | 2 | 7 | 0 | 0 | 0 | **63,64%** |
| **pred. 2.8** | 0 | 0 | 0 | 1 | 1 | 3 | 15 | 0 | 0 | **75,00%** |
| **pred. 3.2** | 0 | 0 | 0 | 0 | 1 | 4 | 0 | 1 | 1 | **14,29%** |
| **pred. 1.1** | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 6 | **85,71%** |
| **class recall** | **75,00%** | **80,00%** | **58,33%** | **59,09%** | **20,00%** | **33,33%** | **88,24%** | **16,67%** | **75,00%** |  |

**Pruebas con Smote (Exactitud = 52,48%, G-media = 48,60%):**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **true 2.1** | **true 18** | **true 1.2** | **true 2.3** | **true CONTEXTO** | **true 2.10** | **true 2.8** | **true 3.2** | **true 1.1** | **class precision** |
| **pred. 2.1** | 6 | 0 | 0 | 3 | 3 | 6 | 2 | 3 | 0 | **26,09%** |
| **pred. 18** | 1 | 3 | 5 | 0 | 4 | 1 | 0 | 2 | 0 | **18,75%** |
| **pred. 1.2** | 0 | 0 | 15 | 3 | 6 | 0 | 0 | 0 | 0 | **62,50%** |
| **pred. 2.3** | 0 | 0 | 0 | 13 | 1 | 0 | 0 | 0 | 0 | **92,86%** |
| **pred. CONTEXTO** | 1 | 2 | 2 | 0 | 13 | 1 | 0 | 0 | 3 | **59,09%** |
| **pred. 2.10** | 0 | 0 | 0 | 2 | 3 | 6 | 0 | 0 | 0 | **54,55%** |
| **pred. 2.8** | 0 | 0 | 0 | 1 | 0 | 4 | 12 | 0 | 0 | **70,59%** |
| **pred. 3.2** | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 1 | 0 | **14,29%** |
| **pred. 1.1** | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 5 | **71,43%** |
| **class recall** | **75,00%** | **60,00%** | **62,50%** | **59,09%** | **43,33%** | **28,57%** | **70,59%** | **16,67%** | **62,50%** |  |

Quedando la muestra final en:

|  |  |  |
| --- | --- | --- |
| **CATEGORIA** | **DESCRIPCION** | **TOTAL** |
| 1.1 | Derecho a la vida | 19 |
| 1.2 | Derecho a la libertad personal | 39 |
| 2.1 | Derecho a la Salud | 25 |
| 2.3 | Derecho al trabajo | 44 |
| 2.8 | Derecho a la Vivienda | 22 |
| 2.10 | Derecho a la alimentación | 22 |
| 3.2 | Derecho al ambiente | 14 |
| 18 | Instancias internacionales de ddhh | 22 |
| CONTEXTO | CONTEXTO | 23 |
|  | Total: | 230 |

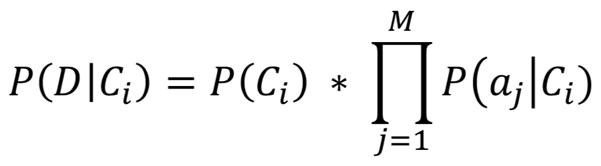
* + Esta nueva muestra, con el peso en 0,03 según el criterio de Ganancia de Información, nos dieron un total de 230 registros y 1.306 atributos para trabajar el modelo.

**Resultado de la fase de preparación de datos**:

Vista minable = Matriz de datos (tabla atributo-valor) con **230** registros (filas) y **1.306** atributos, incluyendo la clase.

(3) Minería de datos:

* Algoritmo: Bayes Ingenuo con corrección laplace:



* Medida de rendimiento: exactitud predictiva, precisión y recall (sensibilidad), kappa.
* Técnica de evaluación: validación cruzada de 10 particiones
* Matriz de Confusión:

Exactitud = 80.43%

Kappa = 0,775

G-medias = 81,20%

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **true 2.10** | **true 2.8** | **true 2.3** | **true CONTEXTO** | **true 2.1** | **true 1.1** | **true 1.2** | **true 3.2** | **true**  **18** | **class precision** |
| pred. 2.10 | 15 | 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | **75,00%** |
| pred. 2.8 | 0 | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **100,00%** |
| pred. 2.3 | 2 | 1 | 39 | 0 | 3 | 0 | 0 | 0 | 0 | **86,67%** |
| pred. CONTEXTO | 2 | 0 | 0 | 16 | 1 | 0 | 0 | 1 | 4 | **66,67%** |
| pred. 2.1 | 3 | 0 | 3 | 1 | 20 | 0 | 1 | 2 | 0 | **66,67%** |
| pred. 1.1 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | **100,00%** |
| pred. 1.2 | 0 | 0 | 1 | 2 | 0 | 0 | 36 | 0 | 2 | **87,80%** |
| pred. 3.2 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 11 | 0 | **78,57%** |
| pred. 18 | 0 | 0 | 0 | 4 | 0 | 2 | 2 | 0 | 16 | **66,67%** |
| **class recall** | **68,18%** | **68,18%** | **88,64%** | **69,57%** | **80,00%** | **89,47%** | **92,31%** | **78,57%** | **72,73%** |  |

* Este es el modelo que se escogió por funcionar mejor, sin embargo, se probaron otros algoritmos: árboles de decisión, basados en reglas y K-NN. De los tres, K-NN fue el que tuvo mejor rendimiento luego de Bayes ingenúo. Los resultados se muestran en la siguiente tabla:.

|  |  |  |
| --- | --- | --- |
|  | **Exactitud** | **G-media** |
| **Bayes Ingenuo** | **80,43%** | **81,2%** |
| K-NN | 73,48% | 70,76% |
| Árbol | 54,77% | 49,1% |
| Reglas | 67,28% | 63,97% |

(4) Evaluación e interpretación:

Exactitud = 58.33%

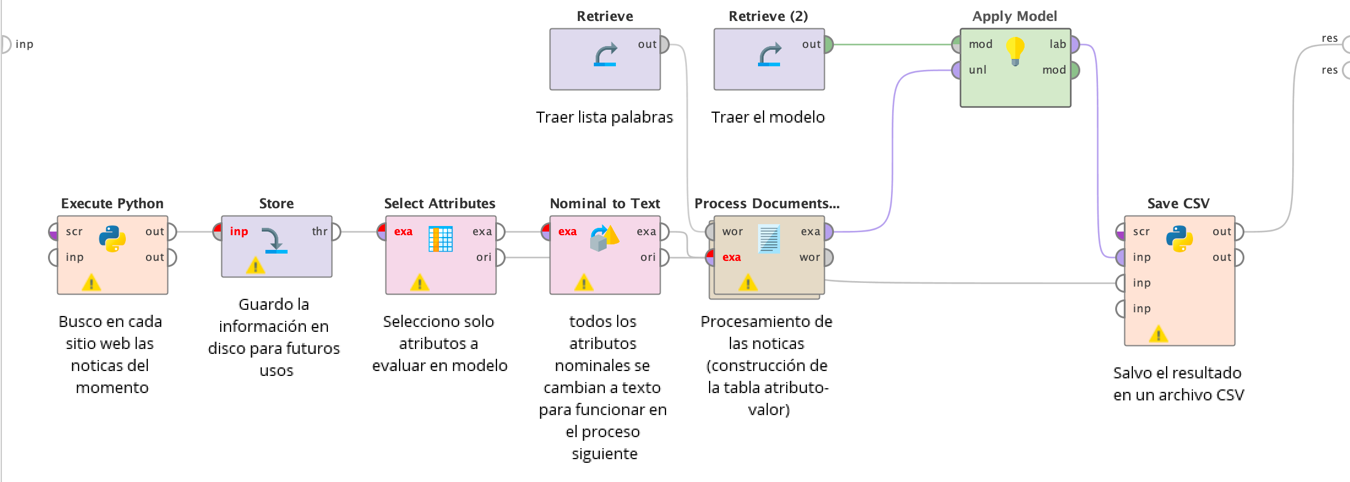
Kappa: 0,51

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CATEGORIA** | **DESCRIPCION** | **TOTAL** | **Precision** | **Recall** |
| 1.1 | Derecho a la vida | 3 | 66,67% | 28,57% |
| 1.2 | Derecho a la libertad personal | 24 | 58,33% | 70% |
| 2.1 | Derecho a la Salud | 8 | 37,50% | 42,86% |
| 2.3 | Derecho al trabajo | 22 | 59,09% | 86,67% |
| 2.8 | Derecho a la Vivienda | 17 | 41,18% | 100% |
| 2.10 | Derecho a la alimentación | 9 | 55,56% | 26,32% |
| 3.2 | Derecho al ambiente | 3 | 66,67% | 33,33% |
| 18 | Instancias internacionales de ddhh | 5 | 60% | 42,86% |
| CONTEXTO | CONTEXTO | 29 | 72,41% | 65,62% |
|  | **Total**: | 120 | **G-media**: | 59,17% |
|  |  |  |  |  |

* Los resultados con las pruebas fueron mejores sin las categorías 18 y CONTEXTO, sin embargo, al momento de probar el modelo con datos extraídos directamente de los sitios web se notó que era necesario usar ambas categorías.

(5) Difusión y uso:

* Se realizaron tres scripts en Python para la aplicación de este modelo en la vida real: categorizarDDHHRetrieveData.py, categorizarDDHHCatCsv.py, categorizarDDHHPdf.py
* categorizarDDHHRetrieveData.py se conecta a los sitios web de El Universal (http://www.eluniversal.com/buscador) y de El nacional (http://www.el-nacional.com/ultimo-minuto) y extrae las noticias más recientes dejando el resultado en dataFrame de panda.
* categorizarDDHHCatCsv.py obtiene los resultados de pasar el modelo por esta información y guarda la información en un archivo csv.
* categorizarDDHHPdf.py toma la información del csv y la guarda en archivos pdf para hacer la entrega al usuario final.
* Los dos primeros scripts son usados en RapidMiner. El proceso que se debe correr para poder aplicar el uso es el siguiente:



* Los resultados obtenidos de aplicar este modelo con datos de la vida real no fueron del todo satisfactorios. Los resultados reflejan bastante bien lo que nos dio en el paso de evaluación e interpretación. El modelo está clasificando bien el 60% de los casos, el otro 40% no queda correcto, sobre todo si las noticias no son de DDHH. Se probaron distintas técnicas para mejorar esto, pero no encontramos como hacerlo más asertivo. Vimos en los datos de entrada que el modelo siempre categoriza, pero hay noticias que no son de DDHH y el modelo debería descartarlas. Intentamos ver si luego de aplicar el modelo algo nos indicaba el porcentaje con el que categoriza una noticia, para no tomarla encueta si el mismo es muy bajo, pero no encontramos cómo hacerlo.
* Creemos debemos mejorar el conjunto de datos de entrenamiento para lograr un mejor resultado, crear una nueva clase para descartar las noticias que no están relacionada con derechos humanos y que se debería seguir trabajando para mejorar la generación de un modelo mas asertivo.