Universidad
del Cauca

Aplicación de Técnicas de Minería de Textos para el Análisis de Noticias Criminales

Camilo Ernesto Sarmiento Torres

Universidad del Cauca

Facultad de Ciencias Naturales, Exactas y de la Educación

Ingeniería Física

Popayán, Colombia

2017

Aplicación de Técnicas de Minería de Textos para el Análisis de Noticias Criminales

Camilo Ernesto Sarmiento Torres

Trabajo de investigación presentado como requisito para optar al título de:

Ingeniero Físico

Director:

Ph.D. Ingeniería Biomédica. Rubiel Vargas Cañas

Codirector:

MSc. Ingeniería de Sistemas con énfasis en Computación. Néstor Díaz Mariño

Universidad del Cauca

Facultad de Ciencias Naturales, Exactas y de la Educación

Ingeniería Física

Popayán, Colombia

2017

Resumen

El siguiente trabajo presenta la aplicación y evaluación de técnicas de minería de texto para la clasificación de noticias criminales, donde se construyó un dataset de noticias criminales de diferentes medios de prensa de la ciudad de Popayán, para identificar y caracterizar diferentes tipos de delitos; además, se construyó un diccionario de palabras en español que permite clasificar las noticias según el tipo de delito, basado en los algoritmos de Clustering K-Means, K-Medoids, CLARA. Conjuntamente se implementaron los algoritmos de clasificación supervisada Naive Bayes, K-nn, Support Vector Machine y Redes Neuronales, para los que se evaluó el desempeño de la clasificación realizada.

Palabras clave: (Minería de Datos, Minería de Texto, Agrupamiento, Clasificadores).

Abstract

The next paper presents the application and evaluation of text mining techniques for the classification of criminal news, it was build a criminal news dataset was constructed from different media outlets in the city of Popayán to identify and characterize different types of crimes; In addition, a dictionary of words in Spanish was constructed that allows to classify the news according to the type of crime, based on the algorithms of Clustering K-Means, K-Medoids, CLARA. The Naive Bayes, K-nn, Support Vector Machine and Neural Networks algorithms applied to monitor the performance of the classification.

Keywords: (Data Mining, Text Mining, Clustering, Classifiers)

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

1.1. PLANTEAMIENTO DEL PROBLEMA

El consolidado oficial de las cifras de violencia y seguridad del ministerio de defensa en Colombia para el año 2015 es el siguiente: homicidas 12.782, homicidios (no uniformados) 12.130, secuestro 213, secuestro extorsivo 123, secuestro simple 90, delitos sexuales 21.597, hurto de vehículos (automotores y motocicletas) 35.005, hurto común (residencias, comercio y personas) 144.931, hurto a entidades financieras 120, piratería terrestre 295, extorsión 5.480, violencia intrafamiliar 75.480, delitos contra recursos naturales y el medio ambiente 3.168, actos de terrorismo 443, actos de terrorismo contra infraestructura 129, acciones de grupos armados al margen de la ley 121, miembros de grupos al margen de la ley neutralizados 3.529, miembros de grupos al margen de la ley capturados 2.325, miembros de crimen organizado capturados 3.073, capturas por minería ilegal 2.265 [1]. Los índices de criminalidad para Colombia son considerablemente altos, donde se observa la aparición de diferentes temáticas criminales, para las cuales se producen constantemente grandes volúmenes de información, que no es explotada y aprovechada de la forma adecuada, dado que requiere un proceso manual, intensivo en tiempo y recursos.

Por su parte, las técnicas de minería de texto permiten analizar grandes cantidades de datos, estructurados y no estructurados, con muy buen desempeño, siendo una herramienta clave para determinar la presencia de distintas temáticas criminales tratadas en una noticia policial. Estas técnicas han sido estudiadas en otros países [2], sin embargo, en el contexto de países de habla hispana estas técnicas no se encuentran muy desarrolladas [3] debido a que: las herramientas software para procesamiento de texto no cuentan con librerías que puedan ser usadas en español [3]; y, de acuerdo a la revisión bibliográfica, no existen diccionarios de palabras adecuados al idioma español y específicamente en la temática criminal.

Así, en [4] se realizo el análisis de los niveles de cobertura policial y su relación con las estadísticas de casos de delitos reales, utilizando clasificadores bayesianos y el algoritmo de k-means. En ese estudio, se manejó un diccionario propio el cual consistía de 20 palabras, y evaluaron la técnica de k-means utilizando índice de Daives Boulin. Sin embargo, solo

se desarrolló un prototipo de apoyo a la visualización de datos donde se dejaron de lado diferentes técnicas como k-medoids, CLARA, discriminantes lineales entre otros. En el caso de la evaluación, no se tuvieron en cuenta índices como: Silhouettes, Marriot y Gamma, entre otros. Adicionalmente, en este trabajo se determinó el número de categorías según los reportados en la literatura, no obstante, existen técnicas que pueden calcular de manera automatizada el número óptimo de estas.

Por lo anterior, este trabajo pretende describir y aplicar técnicas de minería de texto en español para realizar una clasificación de diferentes temáticas criminales. En este contexto se propone la siguiente pregunta de investigación: ¿Qué se requiere para aprovechar y clasificar de forma adecuada la gran cantidad de información policial generada en los artículos de prensa locales?

1.2. OBJETIVOS

1.2.1. Objetivo General

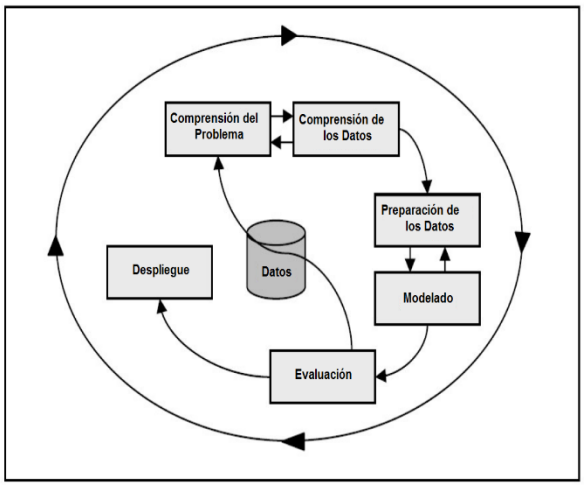
- Implementar técnicas de minería de texto para el procesamiento y clasificación de noticias criminales según el tipo de delito.

1.2.2. Objetivos Específicos

- Construir un dataset de noticias criminales de ámbito local de los últimos tres años.
- Elaborar un diccionario de palabras adecuado en minería de texto a partir del dataset y las técnicas de relevancia.
- Aplicar técnicas de minería de texto para la clasificación de noticias criminales.
- Evaluar los resultados obtenidos de la clasificación de noticias criminales.

1.3. METODOLOGÍA

La metodología adoptada en este proyecto fue una metodología tipo CRISP-DM (Cross Industry Standard Process of Data Mining) [5], que es la guía más ampliamente utilizada en proyectos de minería de datos y consta de seis etapas: Comprensión del problema, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación y Despliegue. Estas etapas se siguen de manera exhaustiva.



El ciclo consta de seis etapas (Figura No. 1.1.), en donde la sucesión de fases no es rígida y se presenta un constante movimiento de ida y vuelta, dado que es necesario para un buen desarrollo del proyecto. El resultado de una fase determina que fase o tarea se tiene que realizar sucesivamente. Las flechas indican las más importantes y frecuentes dependencias entre fases.

Figura No. 1.1. Metodología CRISP-DM [5].

1.3.1. Comprensión del problema

Probablemente sea la fase más importante y envuelve las tareas de comprensión de los objetivos y los requisitos del proyecto con el fin de convertirlos en objetivos técnicos y en un plan del trabajo [5]. Para obtener el mejor provecho de Data Mining, es necesario entender de la forma más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. Es importante convertir el conocimiento adquirido de la temática, en un problema de Data Mining y en un plan preliminar para conseguir los objetivos propuestos.

1.3.2. Comprensión de los datos

Incluye la recolección de datos, con el objetivo de establecer un primer contacto con el problema, se describen y exploran, familiarizándose con

ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir algunas primeras hipótesis. Esta fase junto a las dos siguientes son las que ocupan mayor esfuerzo y tiempo de un proyecto de Minería de Datos.

1.3.3. Preparación de los datos

Luego de efectuar la recolección de los datos se procede a adaptarlos para las técnicas de Data Mining, tales como técnicas de visualización de datos, búsqueda de relaciones entre variables, entre otras. Entre las tareas que incluye esta la selección de datos, limpieza, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Esta fase se encuentra muy relacionada con la fase de modelado, dado que la técnica elegida para el modelado va a determinar diferentes formas de procesar los datos. Por este motivo las fases de preparación y modelado interactúan de forma permanente.

1.3.4. Modelado

En esta fase se seleccionan las técnicas de modelado más apropiadas para el proyecto, teniendo en cuenta que se eligen en función de los siguientes criterios:

- Ser apropiada para el problema.
- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

Previo al modelado, se debe elegir un método de evaluación de los modelos para establecer el grado de bondad de ellos donde se elabora y evalúa el modelo.

1.3.5. Evaluación

Se evalúa el modelo teniendo en cuenta los criterios de cumplimiento y los criterios de éxito del problema. Se debe tener en cuenta que la fiabilidad recae sobre los datos a los que se le realizo el análisis. Se revisa el proceso teniendo en cuenta los resultados obtenidos para poder repetir algún paso anterior en el que se haya cometido un error y considerar que se pueden emplear múltiples herramientas para interpretar los resultados obtenidos. Si se ha determinado positivos los resultados podría pasarse a la siguiente fase, pero en caso contrario volver a una fase anterior, pudiendo incluso decidir empezar de cero con un nuevo proyecto.

1.3.6. Despliegue

Luego que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del problema, se recomiendan acciones basadas en la observación del modelo y sus resultados, como aplicar el modelo en diferentes conjuntos de datos. También se difunde y documenta los resultados obtenidos en un informe final.

2. Marco Teórico

2.1. NOTICIAS CRIMINALES

La noticia de índole criminal es el relato, construcción y elaboración que se refiere a un suceso con características de delito, que consiste en acciones que perjudican a alguien o algo, las cuales van en contra de lo establecido por la ley [6]. Entre los hechos que se relata se encuentran los datos más relevantes del suceso policial: víctimas, personas directamente implicadas, daños y/o la propiedad comprometida, una descripción de la causa y el relato cronológico, las acciones legales en tanto a la investigación y el proceso judicial [6]. Estas noticias son publicadas día a día en los diferentes medios de comunicación, prensa, televisión, internet, entre otros, produciendo grandes cantidades de información, donde el manejo de tales cantidades de información se convierte en una tarea ardua, por lo que es fundamental contar herramientas cada vez más especializadas, con las que se emplea el concepto de minería de texto.

2.2. TEXT MINING

Se define Text Mining o Minería de Texto al proceso de extraer información útil o desconocida a partir de grandes cantidades de textos o documentos [7]. Donde estas técnicas, derivan y son inspiradas de la minería de datos, por esta razón se encuentra gran similitud entre ambas [7].

Gran parte del enfoque en minería de texto se debe invertir en el preprocesamiento, limpieza y normalización de los documentos o textos, donde se extrae e identifica las características más representativas [7]. Las etapas del text mining son:

2.2.1. Representación de Documentos

Para extraer texto, en primera medida se debe procesar de forma adecuada los artículos, para poder aplicar las técnicas de minería de datos, para lo cual es fundamental determinar qué tipo de representación toma un documento. La representación de un documento, está basado en las palabras, aplicando el Modelo de Espacio Vectorial (“Vector Space Model”) o bolsa de palabras (“Bag of Words”).

El modelo de espacio vectorial representa documentos en lenguaje natural de una manera formal mediante el uso de vectores en un espacio

multidimensional, en este aspecto la representación más utilizada es la bolsa de palabras que se define como: Una colección de documentos compuesta por n documentos y m términos, representados por una matriz documento-termino de $n \times m$. Donde los vectores renglón representan los n documentos; y el valor asignado a cada componente refleja una medida de importancia o frecuencia ponderada que produce el termino t_i en la representación del documento j [8].

$$\boldsymbol{d_j = (w_{1j}, w_{2j} \ \ldots \ w_{mj})} \qquad \textbf{(2.1)}$$

Donde:
d_j: Documento.
w_{ij}: Contribución del termino t_i.

2.2.2. Diccionario de Palabras

Una parte fundamental para la aplicación de las técnicas de minería de texto enfocadas en la clasificación de información es el estudio y diseño de los diccionarios de palabras, donde estos son utilizados, en el filtrado, recuperación, indexado y cálculo de relevancia de información [9]. Por esto la construcción del diccionario es fundamental para llevar a cabo exitosamente un proceso de minería de texto. El diccionario de palabras representa documentos en lenguaje natural de una manera formal mediante el uso de vectores en un espacio multidimensional [9]. Se define un diccionario $D=(C,X)$ como un conjunto de palabras que describen un conjunto semántico C en un documento X [9]. Dado un concepto semántico C y una colección de documentos X, utilizando un vocabulario V, devuelve un ranking de palabras $w \in V$, de modo que las palabras que se refieren al concepto C aparecen más que las palabras que no se refieren a C [9].

2.2.3. Proceso de Tokenizacion

Teniendo el total de los documentos de forma estandarizada, se aplica un proceso mediante el cual el texto se divide en unidades, denominados tokens, que corresponden a las palabras en el idioma en el que se realiza la extracción. Esto se ejecuta antes de que se aplique cualquier algoritmo de procesamiento, el cual depende claramente del lenguaje de programación en el que se está trabajando y las pautas en específico de este [10].

Para determinar las diferentes palabras, se identifican los delimitadores de cada uno de los tokens, que usualmente corresponden a signos de puntuación, espacios en blanco y caracteres que no son alfabéticos [10].

2.2.4. Eliminación de Stopwords

Con el enfoque de bolsa de palabras “Bag of Words”, es posible tener decenas de miles de diferentes palabras que ocurren en un conjunto de documentos. Gran parte de ellas no se consideran relevantes en la caracterización de un documento. Estas palabras se denominan Stopwords (palabras vacías) [11].

Las palabras que aparecen en la mayoría de los documentos, así como las que ocurren muy poco, poseen poca información útil a la hora de diferenciar los diferentes tipos de documentos [11]. Por lo cual es imprescindible reducir el tamaño del espacio de características, en donde se eliminan una lista de palabras (stopwords), tales como artículos, pronombres, conjunciones, entre otros.

2.2.5. Proceso de Stemming

Luego de que cada documento ha sido separado en tokens y se haya eliminado las stopwords, es conveniente reducir cada una de las palabras a sus raíces, excluyendo de esta manera los sufijos [12].

Esta implementación se basa en que las palabras presentan muchas variables de forma, para lo cual la se busca una única palabra que englobe y entregue la mejor descripción [12], ejemplo:

Tabla No. 2.1 Ejemplo de Steeming

PALABRA	RAIZ
Sal	Sal
Salero	
Salado	
Salar	

2.2.6. Selección de Atributos

La calidad de cualquiera de los métodos de agrupación: clustering y clasificación son altamente sensible al ruido de los atributos que se utilizan para el proceso de agrupación [13]. Por ejemplo, la palabra “el” es una característica común en todos los documentos, pero no representa un valor útil para mejorar la calidad de la agrupación, por lo tanto, es fundamental seleccionar las características de manera efectiva de modo que el ruido de las palabras se elimina antes de la agrupación, para lo cual se aplican una

serie de métodos que reducen las dimensiones del espacio de atributos, para que los documentos sean más susceptibles a agruparse [13].

Para esta selección de atributos existen los métodos supervisados y no supervisados dependiendo de las técnicas de agrupación.

2.2.6.1. Selección No Supervisada de Atributos

Dado que los diferentes tipos de categorías son desconocidos, se aplican los métodos de selección no supervisada de atributos, como frecuencia de termino (*tf*) y la contribución de termino (Term Contribution).

La frecuencia de termino (*tf*), es el número de veces que aparece un término en un determinado documento [14]:

$$\boldsymbol{tf(n) = \sum_{n} D1} \tag{2.2}$$

La frecuencia de aparición de termino n en un documento $D1$ es la suma de las ocurrencias de dicho termino.

La Contribucion de Termino (TC) [14], es la medida del peso o ponderación de cada termino en cada documento y se calcula multiplicando la frecuencia de aparición de cada termino (*tf*) y su frecuencia inversa de documento (*idf*).

Para lo cual se define la frecuencia inversa de documento (*idf*), así:

$$\boldsymbol{idf(n) = \log_{10} \frac{N}{DF(n)}} \tag{2.3}$$

Donde:

N: es el número total de documentos de la colección.

DF: (Document Frequency) es el número de documentos en los que aparece el termino n a lo largo de toda la colección.

En algunos casos es conveniente normalizar la frecuencia de termino, dividiendo por la frecuencia máxima en el documento para que la longitud de los documentos no afecte a la relevancia del término.

Por lo tanto, la contribución de termino:

$$\boldsymbol{TC(n,d) = tf(n,d) \times idf(n)} \tag{2.4}$$

Donde d representa a un documento determinado.

Entonces TC asigna al termino n un peso en el documento D que es:

- Más alto, cuando el termino n aparece muchas veces dentro de un número pequeño de documentos (dando la máxima capacidad de discriminación para esos documentos), este factor es el que mejor identifica a un documento dado.

- Más bajo, cuando el termino n aparece pocas veces en un documento o se presenta en muchos documentos (dando una menor capacidad de discriminación).

- El más bajo cuando el termino n aparece en todos los documentos.

- Cero cuando el termino no aparece en los documentos.

2.2.7. Medidas de Similitud

Luego de ser representados los documentos como vectores, podemos medir su similitud. Ya que al estar representados los documentos en un espacio multidimensional, los documentos serán similares entre más elementos compartan, es decir entre más palabras tengan en común [15].

Si la similitud la expresamos en una escala entre 0 a 1, dos documentos A y B, son tanto más similares cuando más cerca sea su valor de similitud a 1 [15].

Se utiliza el cálculo de distancia como medida de similitud entre documentos entre las cuales encontramos, distancia euclidiana, manhattan, minkowski, coseno, entre otras.

- Distancia euclidiana: Se define la distancia euclidiana (*DE*), como una función utilizada para medir la distancia entre dos puntos en el espacio multidimensional. Mide la distancia en línea recta entre dos puntos, para lo cual la distancia mínima entre dos puntos es cero, entre más se aleje del cero la distancia entre dos puntos, más diferentes son los vectores y por lo tanto los documentos que representan [16]. Se representa mediante la siguiente ecuación:

$$\boldsymbol{DE\left(\overrightarrow{v1},\overrightarrow{v2}\right) = \sqrt{\sum_{j=1}^{t}(x_i - y_i)^2}} \quad \textbf{(2.5)}$$

Donde:

$\overrightarrow{v1},\overrightarrow{v2}$ son los vectores de términos.

x_i, y_i son los ponderados de términos.

- Distancia Manhattan: Se define la distancia Manhattan como el cálculo de la suma de las diferencias absolutas entre sus coordenadas y se representa mediante la siguiente ecuación [17]:

$$\boldsymbol{DM\left(\overrightarrow{v1},\overrightarrow{v2}\right) = \sum_{i=1}^{n}|x_i - y_i|} \quad \textbf{(2.6)}$$

- Distancia Minkowski: Se define la distancia de Minkowski como la medida generalizada de una amplia gama de distancias [17], así:

$$\boldsymbol{DMin\left(\overrightarrow{v1},\overrightarrow{v2}\right) = \left(\sum_{i=1}^{n}(x_i - y_i)^p\right)^{1/p}} \quad \textbf{(2.7)}$$

Dónde: $p>0$

- Medida del Coseno: Una de las limitantes de la distancia euclidiana es que la variabilidad del largo de los documentos afecta a la métrica, por esta razón es más frecuente utilizar la medida del coseno.

  Lo más frecuente es utilizar el coseno del ángulo entre los vectores como medida de similitud (coseno de Salton), si los documentos son iguales el ángulo vale 0 y por lo tanto el coseno es 1, mientras que si los dos vectores son ortogonales el coseno vale 0. Se define [16]:

$$\boldsymbol{SCos\left(\overrightarrow{v1},\overrightarrow{v2}\right) = \frac{\overrightarrow{v1}\cdot\overrightarrow{v2}}{\left|\overrightarrow{v1}\right| \times \left|\overrightarrow{v2}\right|} = \frac{\sum_{i=1}^{t} x_i \times y_i}{\sqrt{\sum_{i=1}^{t} {x_i}^2 \times \sum_{i=1}^{t} {y_i}^2}}} \quad \textbf{(2.8)}$$

Dado que el ángulo del coseno es una medida de similitud, se pasa a una medida de distancia restando su valor por 1:

$$\boldsymbol{DCos = 1 - SCos} \quad \textbf{(2.9)}$$

2.2.8. Clustering de Documentos

Clustering se refiere a agrupar objetos que son similares entre sí y disimiles a los objetos pertenecientes a otros grupos [11]. En muchos campos se producen beneficios que se obtienen al agrupar objetos similares. Ejemplo:

- En una aplicación de recuperación de documentos, es posible que se desee encontrar documentos que contengan información similar.

- En una aplicación de análisis de delitos podemos buscar clusters de altos volúmenes de crímenes tales como robos e intentar determinar posibles relaciones con asesinos.

Existen muchos algoritmos para la agrupación, pero este trabajo se centra en los métodos para los cuales la medida de similitud entre objetos está basada en una medida de la distancia entre ellos.

2.2.8.1. K-Means

K-Means es un algoritmo de agrupamiento, en donde cada objeto del dataset se asigna a un único grupo o partición denominada cluster (existen otros métodos que permiten que un objeto este en múltiples clusters) [11]. El termino cluster hace referencia a grupo de objetos que tienen características similares entre ellos. Cada uno de los grupos es representado por la media o media ponderada de sus puntos, denominado centroide, centro de gravedad, centro geométrico o puntos medios del cluster [11].

Para este método de agrupación se empieza por decidir cuantos clusters se desea formar a partir de los datos. Se denomina a este valor k, que se representa con un número entero. Una vez iniciado el proceso se escogen k puntos iniciales aleatorios (que en este caso representan un documento) en todo el conjunto de datos $x = \{x_1, x_2, \dots, x_n\}$ y se calcula la distancia que hay entre los centroides (puntos iniciales) $u = \{u_1, u_2, \dots, u_k\}$ y el resto de los puntos, luego se resignan (los documentos) a los puntos que estén más próximos. Posteriormente se vuelven a calcular los centroides de cada uno

de los clusters $c = \{c_1, c_2, \ldots, c_k\}$ basado en los actuales miembros, por último, cada documento es reasignado al cluster que tenga el centroide más cercano o sea más similar. Se repite el proceso hasta que los centroides no cambien o se alcance un máximo de iteraciones [11].

El método de agrupación implementando en el algoritmo de K-Means, para formar dos grupos (clusters), se describe en la Figura No. 2.1, en donde los puntos rojos representan los centroides:

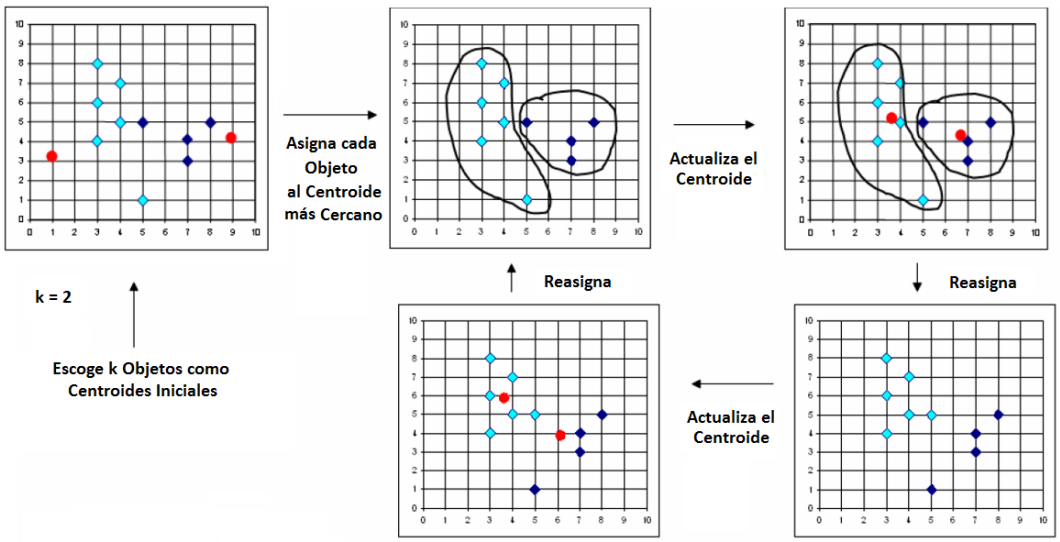


Figura No. 2.1 Representación algoritmo K-Means [18].

En general, el objetivo del algoritmo de K-Means es reducir al mínimo la distancia entre los puntos de cada grupo y su centroide, que se representa mediante la siguiente ecuación [12]:

$$\boldsymbol{SC = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - u_i\|^2} \qquad \textbf{(2.10)}$$

Donde el centroide u_i del cluster c_i es:

$$\boldsymbol{u_i = \frac{1}{n_i} \sum_{x \in c_i}^{n} d(x_i, u_i)} \qquad \textbf{(2.11)}$$

Determinar el mejor valor de k juega un papel crítico en el rendimiento del modelo. Para seleccionar el mejor valor de k, no hay una regla que se aplica todas las situaciones [12].

2.2.8.2. K-Medoids (PAM – Partitional Around Medoids)

K-Medoids es un algoritmo de agrupamiento que presenta una variación a K-Means, que trabaja de igual manera dividiendo el conjunto de datos en particiones o grupos, e intentando minimizar la distancia entre puntos y su centroide, pero bajo la salvedad que se escogen datapoints como centros, es decir los centros del cluster deben pertenecer al dataset, y el objetivo es determinar el mejor datapoint que represente al centro del cluster, conocido como medoide. trabaja con una métrica arbitraria de distancia entre datapoints [12]. De forma técnica un medoide es un objeto del cluster cuyo promedio de disimilitud a todos los objetos en un cluster es mínimo [19]. La función objetivo se representa mediante la ecuación:

$$\boldsymbol{min} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{x \in c_i}} \boldsymbol{d(x, u_i)} \qquad \textbf{(2.12)}$$

Donde x representa un objeto, u_i su medoide y c_i el cluster.

Inicialmente el algoritmo determina los k medoides y luego cada objeto que no es un centro se agrupa a su medoide más cercano. Posteriormente cambia los medoides con otros objetos candidatos hasta alcanzar un minimo de distancia entre objetos y su medoide, lo que continua hasta que no se produzcan más cambios de medoides [19].

2.2.8.3. CLARA

Es un algoritmo de agrupamiento basado en el enfoque de K-Medoids, diseñado para agrupar grandes cantidades de datos y su nombre viene de las siglas Clustering LARge Applications [20]. El agrupamiento se realiza en dos pasos, se extrae una muestra del conjunto de datos y se agrupa en K subconjuntos utilizando el método de K-Medoids, donde k determina el número de subconjuntos, entonces cada objeto que no pertenece a la muestra se asigna al subconjunto más cercana [20]. El proceso se repite hasta alcanzar que la distancia media entre los datos y el medoide sea la mínima. Esto produce un agrupamiento de todo el conjunto de datos.

2.2.8.4. Evaluación de Clustering

El clustering es un proceso no supervisado que presenta una amplia sensibilidad a los parámetros de entrada, pero resulta difícil definir cuando un resultado de un agrupamiento es aceptable, por esta razón existen técnicas e índices para la validación de un agrupamiento realizado.

Existen dos tipos de validación, externa e interna, en donde la principal diferencia es si se usa o no información externa para la validación, es decir, información que no es producto de la técnica de agrupación. Por esta razón en este trabajo el enfoque se centra en la validación interna.

Debido a que los algoritmos de clustering pretenden agrupar objetos similares, las métricas de validación interna están basadas en dos criterios: cohesión y separación.

Cohesión [21]: El miembro de cada cluster debe ser lo más cercano posible a los miembros de su mismo cluster.

Separación [21]: Los clusters deben estar ampliamente separados entre ellos.

Existen muchos índices o coeficientes de validación interna pero el más ampliamente utilizado es el coeficiente de Silhouettes.

Se define el coeficiente de Silhouettes mediante la siguiente expresión [21]:

$$\boldsymbol{s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}} \tag{2.13}$$

Donde:

$a(i)$ es la distancia media entre el objeto y todos los otros objetos de la misma clase.

$b(i)$ es la distancia media entre el objeto y todos los otros objetos del cluster más próximo.

El valor de $s(i)$ se obtiene a partir de la siguiente regla.

$$\boldsymbol{s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & si\ a(i) < b(i) \\ 0, & si\ a(i) = b(i) \\ \frac{b(i)}{a(i)}, & si\ a(i) > b(i) \end{cases}} \tag{2.14}$$

De esta definición se obtiene que el índice de Silhouette está en un rango $-1 \leq s(i) \leq 1$. Donde un valor alto cercano a uno indica un buen agrupamiento, un valor cercano a 0 indica que se encuentra en los extremos del cluster y un valor negativo indica que el objeto se encuentra mal agrupado.

Este índice se representa gráficamente calculando los índices para cada uno de los objetos del dataset, donde se pueda observar toda la agrupación a través de una sola gráfica.

Donde el coeficiente de Silhouette para todo el agrupamiento es [21]:

$$\boldsymbol{s = \frac{1}{n}\sum_{i=1}^{n} s(i)} \qquad \textbf{(2.15)}$$

2.2.8.5. Determinar Numero Optimo de K

Dado que los métodos de agrupación requieren que se especifique el número de grupos que se genere y que son altamente sensibles al número de estos, es fundamental elegir el número óptimo de estos. Desafortunadamente no existe una respuesta definitiva, porque la agrupación optima es de alguna manera subjetiva y depende del método utilizado para medir las similitudes y los parámetros utilizados en la agrupación [22]. Para lo cual se describen dos métodos ampliamente utilizados: método de Elbow y Silhouettes.

- **Método de Elbow**

El método de Elbow también conocido como método del codo utiliza los valores de inercia obtenidos tras aplicar las técnicas de clustering a diferentes números de clusters desde 1 hasta n, siendo la inercia la suma de distancias cuadradas de cada objeto del cluster a su centroide, se representa mediante la expresión [22]:

$$\boldsymbol{Inercia = \sum_{i=0}^{n} \|x_i - u\|^2} \qquad \textbf{(2.16)}$$

Posterior a obtener los valores de inercia, se representan en una gráfica bidimensional la inercia versus el número de clusters. El valor de inercia siempre decrece a medida que aumenta el número de

clusters. En donde un cambio brusco de inercia nos dirá el número optimo, similar al punto que representaría al codo de un brazo. En algunos casos resulta difícil apreciar el codo e inclusive se pueden observar dos o más codos [22].

- **Método de Silhouettes**

El enfoque de medida de Silhouette fue descrito en la sección 2.2.8.4. Que mide la calidad de agrupación, es decir determina que tan bien agrupado este cada objeto dentro de un grupo, en donde una anchura promedio alta de silhouette indica una buena agrupación. Se calcula la silueta promedio para diferentes valores de k y el numero optimo se obtiene a partir del valor más alto.

2.2.9. Clasificación de Documentos

La clasificación de documentos se basa en determinar una función de asignación de una categoría $F: D \times C \rightarrow \{0,1\}$, donde D es el conjunto de todos los documentos posibles y C es el conjunto de categorías predefinidas. El valor de la función $F(D, C)$ es 1 si pertenece a la categoría C y 0 en caso contrario. La tarea es construir una función de aproximación denominada clasificador que produzca resultados lo más cercano posibles a la categoría verdadera [7]. La clasificación de documentos es un proceso supervisado debido a que se debe tener conocimiento previo de las clases a la cual pertenece un documento, para lo cual existen clasificadores como Naive Bayes, k-nn, SVM, redes neuronales, entre otros.

2.2.9.1. Clasificador Naive Bayes

Es una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados, basado en el teorema de Bayes, que en términos ordinarios este tipo de clasificador considera que cada una de las características contribuye de forma independiente a la probabilidad de pertenecer a una clase independientemente de la presencia o ausencia de las otras características, permitiendo simplificar los cálculos implicados [13].

El teorema de Bayes permite identificar si un documento d_j pertenece a una clase C_k calculando la probabilidad [13], así:

$$\boldsymbol{P(C_k|d_j) = \frac{P(d_j|C_k)P(C_k)}{P(d_j)}} \quad \textbf{(2.17)}$$

Donde la probabilidad $P(d_j)$ es constante para todas las categorías y para calcular la probabilidad $P(d_j|C_k)$ se debe representar al documento como un vector de características $d_j = (w_{1j}, w_{2j} \ \ldots \ w_{mj})$ y tener en cuenta cada característica contribuye de forma independiente teniendo en cuenta la clase, así:

$$\boldsymbol{P(d_j|C_k) = \prod P(w_j|C_k)} \quad \textbf{(2.18)}$$

Donde se calcula la probabilidad de pertenecía a una clase determinada y se clasifica con el valor más alto.

Es denominada una técnica supervisada por que necesita tener ejemplos para entrenar el modelo y poder realizar la clasificación, se divide en dos etapas, crear el modelo y prueba, así:

Crear el modelo: A partir de una muestra en la cual ya es conocida su clase.

- Calcular las probabilidades a priori de cada clase
- Se distribuye cada clase por separado con sus características.
- Se aplica la corrección de Laplace, para que los valores de cero no presenten problema.
- Se normaliza para tener un rango entre [0,1].

Prueba: Clasificar a partir de nuevos ejemplos.

- Se calcula la probabilidad de pertenencia de cada uno de los nuevos documentos.
- Se asignan a una determinada clase.

2.2.9.2. Clasificador K-nn

El clasificador K-nn (K nearest neighbors) o K vecinos más cercanos, es un método supervisado que permite estimar la función de densidad de probabilidad o la probabilidad directa de que un elemento x pertenezca a la clase Ck a partir de la información proporcionada por un conjunto de

entrenamiento [23]. Este mide la proximidad entre objetos con alguna medida de similitud y es conocido como un método Lazy Leaner, dado que no construye un modelo a partir de los datos de entrenamiento y solo se ejecuta cuando llega la instancia de prueba [23], donde la idea fundamental es que el objeto de prueba se clasificará en la clase más frecuente de sus k vecinos más próximos [23].

El algoritmo consta de los siguientes pasos [23]:

- Se calcula la similitud del objeto de prueba con los objetos de entrenamiento.
- Se determina los k documentos más similares al objeto de prueba.
- Se asigna el objeto de prueba a la clase que ocurre más frecuente a sus k vecinos más próximos.

La elección de k depende fundamentalmente del conjunto de datos, valores grandes de k reducen el ruido, pero crean problemas en los límites entre clases parecidas, ya que incluirá puntos de otras clases [23].

2.2.9.3. Redes Neuronales

Una red neuronal artificial (ANN) es un esquema computacional distribuido e inspirada en la estructura del sistema nervioso de los seres humanos. La arquitectura de una red neuronal es formada conectando múltiples procesadores elementales, siendo éste un sistema adaptivo que posee un algoritmo para ajustar sus pesos (parámetros libres) para alcanzar los requerimientos de desempeño del problema basado en muestras representativas [26]. De forma análoga a los otros algoritmos de clasificación se divide en etapas de entrenamiento y prueba.

La red neuronal feedfoward (FANN), es una de las más estudiadas dado que tiene diversos acampos de aplicación y se define así [26]:

Constituida por tres capas: Capa de entrada, Capa oculta y Capa de salida.

Dado un conjunto de observaciones, la tarea de aprendizaje neuronal es construir un estimador $G_\lambda(x, w)$ de la función desconocida $h(x)$ del cual se conoce solo un conjunto de datos:

$$G_{\lambda}(x, w) = \gamma_2 \left(\sum_{j=1}^{\lambda} w_j{}^{[2]} \gamma_1 \left(\sum_{I=1} w_{Ij}{}^{[1]} x_1 + w_{M+1,j}{}^{[1]} \right) + w_{\lambda+1}{}^{[2]} \right) \quad \textbf{(2.19)}$$

Donde $w = (w_1, \dots, w_d)^T$ es un vector paramétrico a ser estimado y equivale a las ponderaciones de las conexiones entre las neuronas de la red, γ_1 es una función no lineal acotada y diferenciable con forma de función sigmoide o radia basal, γ_2 es una función que puede ser lineal o no lineal y λ es el parámetro de control que indica el número de neuronas escondidas.

La función γ_1 típicamente es la función sigmoidal dada por:

$$\gamma_1 (z) = \frac{1}{1+e^{-z}} \quad \textbf{(2.20)}$$

Si la función γ_2 se elige no lineal, debe ser estrictamente monótona, acotada y diferenciable. La función sigmoidal satisface estos requisitos.

Por lo tanto, el ajuste de la red se produce como resultado de la estimación de los parámetros basado en una muestra de tamaño n. La estimación es obtenida minimizando la función de costo:

$$w_n{}^{LS} = argmin\{L_n(w) : w \in W \subseteq R^d\} \quad \textbf{(2.21)}$$

Donde $L_n(w)$ viene dada por la función del promedio de errores al cuadrado entre el dato estimado y el dato real:

$$L_n = \frac{1}{2n} \sum_{t=1}^{n} (y_t - G(x^t, w))^2 \quad \textbf{(2.22)}$$

2.2.9.4. Support Vector Machine (SVM)

Las Maquinas de Soporte Vectorial son un conjunto de algoritmos de aprendizaje supervisado que permite resolver problemas de clasificación. Dado un conjunto de muestras de entrenamiento, podemos etiquetar las clases y entrenar unas SVM para construir un modelo que prediga la clase de una muestra [7].

Este modelo representa los puntos de una muestra en el espacio, separando las clases en espacios lo más amplios posibles mediante un hiperplano, él cual es calculado mediante los vectores de soporte.

Los vectores de soporte, son trazados a través de los puntos en los extremos de separación de cada clase, por lo tanto, cuando las muestras de prueba se ponen en correspondencia con el modelo, en función del espacio en el que pertenezcan pueden ser clasificados a una determinada clase [7]. El problema se centra en encontrar la separación optima, el hiperplano que tenga la máxima distancia (margen) con los puntos que están más cerca del mismo.

La solución del hiperplano se expresa de la siguiente manera:

$$\boldsymbol{W^T X_i + b = 0} \tag{2.23}$$

Donde:
W es el vector ortogonal al plano.
b es el coeficiente de intercepción.

En una clasificación binaria, la función de clasificación depende del signo así:

$$\boldsymbol{f(x) = signo(W^T X_i + b)} \tag{2.24}$$

La elección del signo es arbitraria y nos permitirá clasificar, clase uno $W^T X_i + b \leq -1$, clase dos $W^T X_i + b \leq +1$.

Y el margen se calcula a partir de la distancia entre los dos planos de los vectores de soporte:

$$\boldsymbol{d^- = d^+ = \frac{|WX + b|}{\|W\|} = \frac{1}{\|W\|}} \tag{2.25}$$

$$\boldsymbol{Margen = d^- + d^+ = \frac{2}{\|W\|}} \tag{2.26}$$

Dado que maximizar la margen es proporcional a minimizar el problema inverso:

$$\boldsymbol{\varphi(W) = \frac{1}{2}\|W\|^2} \tag{2.27}$$

Para resolver este problema usamos los multiplicadores de Lagrange y se obtiene la función a optimizar sujeta a la condición α.

$$\boldsymbol{Maximizar\theta(\alpha) = \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i} \quad \textbf{(2.28)}$$

Sujeto:

$$\boldsymbol{\sum_i \alpha_i y_i = 0\,, \quad \alpha_i \geq 0} \quad \textbf{(2.29)}$$

Inicialmente fue desarrollado para resolver problemas de clasificación binaria y lineal, en donde existe una única frontera de separación representada por una línea en R^2 y por un hiperplano en R^n, pero en la clasificación de documentos se constituye una clasificación multiclase y que no es linealmente separable.

Para conjuntos que no son linealmente separables se transforma el espacio de características en otro que es linealmente separable denominado truco del kernel (kernel trick).

Como los vectores aparecen en productos escalares se aplica la transformación de la función del kernel asi:

$$\boldsymbol{K(x,z) = \varphi(x)^T \varphi(z)} \quad \textbf{(2.30)}$$

Y se puede integrar en la función a optimizar y para lo cual existen muchas funciones de kernel entre las cuales están, polinomial, funciones de base radial, sigmoide, multi-cuadratico inverso, kernel de intersección entre otras.

Por otra parte, en la prueba del modelo, existen variables que pueden encontrarse dentro del margen del hiperplano, por esta razón se agrega un margen de tolerancia ε que permite que estén en un margen inferior a uno, que también es regulada con él parámetro C.

Para realizar la clasificación multiclase se realiza la combinación de varios clasificadores binarios en la configuración uno vs. uno [24]:

- Entrenamiento: Entrenar un clasificador binario para cada combinación posible de dos clases i y j. Donde existen $n(n-1)/2$ clasificadores distintos.

- Prueba: Aplicar todos los clasificadores y acumular un voto a la clase ganadora en cada caso. Resultando ganador la clase con el mayor número de votos.

2.2.9.5. Evaluación de la Clasificación

Para medir el desempeño de los algoritmos de aprendizaje supervisado se emplea la matriz de confusión, que permite visualizar a través de una tabla las predicciones hechas por cada uno de los algoritmos. Cada columna representa el número de predicciones de cada clase y cada fila representa a las instancias en la clase real [25].

Tabla No. 2.2 Matriz de Confusión

		Predicción	
		Clase 1	**Clase 2**
Real	**Clase 1**	*TP*	*FN*
	Clase 2	*FP*	*TN*

La tabla informa falsos positivos (*FP*), falsos negativos (*FN*), verdaderos positivos (*TP*) y verdaderos negativos (*TN*).

- *TP*: Numero de objetos de la clase 1 clasificados correctamente como clase 1.
- *FP*: Numero de objetos pertenecientes a la clase 2 clasificados erróneamente en la clase 1.
- *FN*: Numero de objetos pertenecientes a la clase 1 clasificados erróneamente como clase 2.
- *TN*: Numero de objetos pertenecientes a la clase 2 clasificados correctamente como clase 2.

Dada la matriz de confusión se definen diferentes medidas de evaluación:

Accuracy [25] o exactitud que determina el número de objetos que son correctamente clasificados, mediante la siguiente expresión:

$$\boldsymbol{Exactitud = \frac{\sum TP + \sum TN}{n}} \qquad \textbf{(2.31)}$$

Error rate [25] o tasa de error determina el número de objetos que son erróneamente clasificados, mediante la expresión:

$$\boldsymbol{Tasa\ de\ Error = \frac{\sum FP + \sum FN}{n}} \qquad \textbf{(2.32)}$$

Donde n es el número total de objetos.

Recall [25] determina número de objetos de la clase 1 que son correctamente clasificados como clase 1, mediante la expresión:

$$\boldsymbol{Recall = \frac{TP}{TP + FN}} \quad \textbf{(2.33)}$$

Precision [25] determina el número de objetos de la clase 1 que realmente pertenecen a la clase 1, mediante la expresión:

$$\boldsymbol{Precision = \frac{TP}{TP + FP}} \quad \textbf{(2.24)}$$

F-Score [25] es la media armónica de precisión y sensibilidad, mediante la expresión:

$$\boldsymbol{FScore = \frac{2(Precision \times Recall)}{(Presicion + Recall)}} \quad \textbf{(2.26)}$$

2.3. ANTECEDENTES DE INVESTIGACION

Las técnicas de Minería de datos como las técnicas de minería de texto han dado un impulso notable en el campo de la criminalística, permitiendo extraer conocimiento de conjuntos extensos de información y proporcionando herramientas que ayudan a identificar las características de la delincuencia, en este contexto, el trabajo realizado por Keyvanpour, Javideh y Ebrahimi [27], determino propiedades importantes de informes narrativos de policía en texto plano, donde se utilizó el enfoque SOM de redes neuronales para agrupar información en el proceso de coincidencia de delito y demostró la factibilidad de las técnicas, rescatando la importancia de su implementación en el análisis delictivo.

Un segundo trabajo desarrollado por Sergei Ananyan [28], analizo reportes históricos de un departamento de policía de Virginia, donde se identificaron patrones históricos, indagando la relación entre tipo de delito y la ubicación del incidente, relaciones entre tipo de delito, ubicación y el tipo de arma, implementando las técnicas de minería de texto, donde se planteó un proceso automatizado y se desarrolló un software para facilitar el análisis de los crímenes tanto en la prevención como en la identificación de delitos.

En labor análoga fue presentado por Kianmehr y Alhajj [29], la implementación de las Maquinas de Soporte Vectorial (SVM) para la predicción y clasificación de datos espaciales en el análisis de crímenes, permitiendo predecir lugares probables de ocurrencia de crímenes, contribuyendo en estrategias de seguridad policial.

En [30] se destaca la importancia de cada una de las técnicas de minería de datos y el impacto que tienen en el ámbito de estudio de la criminalística, dado que permiten la extracción de patrones, desde datos como texto, imágenes o audio, para identificar automáticamente a sospechosos y personas, presentando beneficios en el procesamiento eficaz y veloz del flujo de la información, en la clasificación y agrupación de delitos a un determinado grupo delictivo, para el seguimiento del comportamiento y atribuir delitos a pandillas, en la detección de delitos financieros y el seguimiento de delitos web. De forma consecuente se presenta una lista de ocho categorías de crímenes basados en la clasificación de diferentes expertos en la temática criminal: Accidentes de tránsito, delitos sexuales, robos, fraudes, incendios intencionales, delitos de drogas, crímenes violentos y cibercrimen.

Finalmente, el estudio más próximo al contexto colombiano fue desarrollado en Chile por Torres [4], en donde se realizo el análisis de los niveles de cobertura policial y su relación con las estadísticas de casos de delitos reales utilizando técnicas de minería de texto, identificando temáticas

criminales de una base de datos y desarrollando un prototipo de herramienta de visualización geográfica para la cobertura de noticias en relación al número de casos reales.

3. Metodología

La metodología que adopta este proyecto es tipo CRISP-DM (Cross Industry Standard Process of Data Mining) [5], donde se siguen las fases, pero no se entra en profundidad donde no aplique, las Figuras No. 3.1 - 3.2 - 3.3 - 3.4-3.5 son adaptadas de [5]:

3.1. COMPRENSIÓN DEL PROBLEMA

El problema central indagado en este trabajo es la clasificación de noticias criminales dependiendo del tipo de delito.

- **Determinar objetivos del problema:**
  - **Background:** Se realizó una revisión bibliográfica en donde se identificaron los avances realizados en minería de datos respecto a la temática criminal, en el contexto mundial y latinoamericano (Sección 2.3). Donde se identificó la problemática en el contexto nacional y local (Sección 1.1).
  - **Objetivos del problema:** Se identificó la problemática, para la cual se plantearon los objetivos del problema, dando como resultado los objetivos del trabajo de investigación.
  - **Criterios de Éxito:** Como criterio de éxito de la resolución del problema se plantea realizar una clasificación de crímenes según el tipo de delito.
- **Determinar objetivos de Minería de Datos:**
  - **Metas de Minería de Datos:** Se plantearon como metas de minería de datos, determinar qué características son más significativas para diferencias y clasificar las noticias criminales según el tipo de delito, para lo cual se obtiene y elabora un diccionario de palabras, donde se identificaron diferentes categorías a partir de las temáticas criminales para finalmente realizar una clasificación de noticias.
  - **Criterios de Éxito de Minería de Datos:** Como criterio de éxito se plantean índices altos de cohesión, separación y exactitud, entre las diferentes clasificaciones.

La Figura No. 3.1 Describe el proceso realizado.

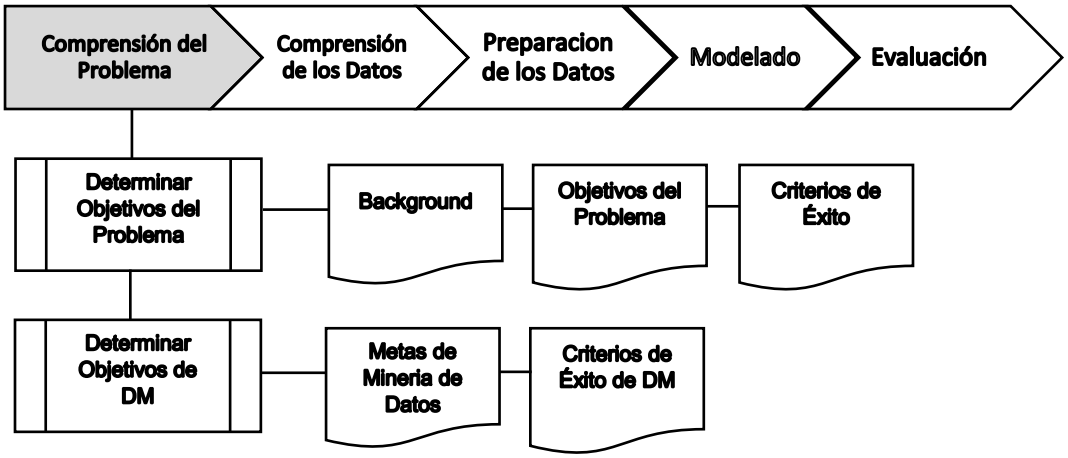


Figura No. 3.1 Comprensión del Problema.

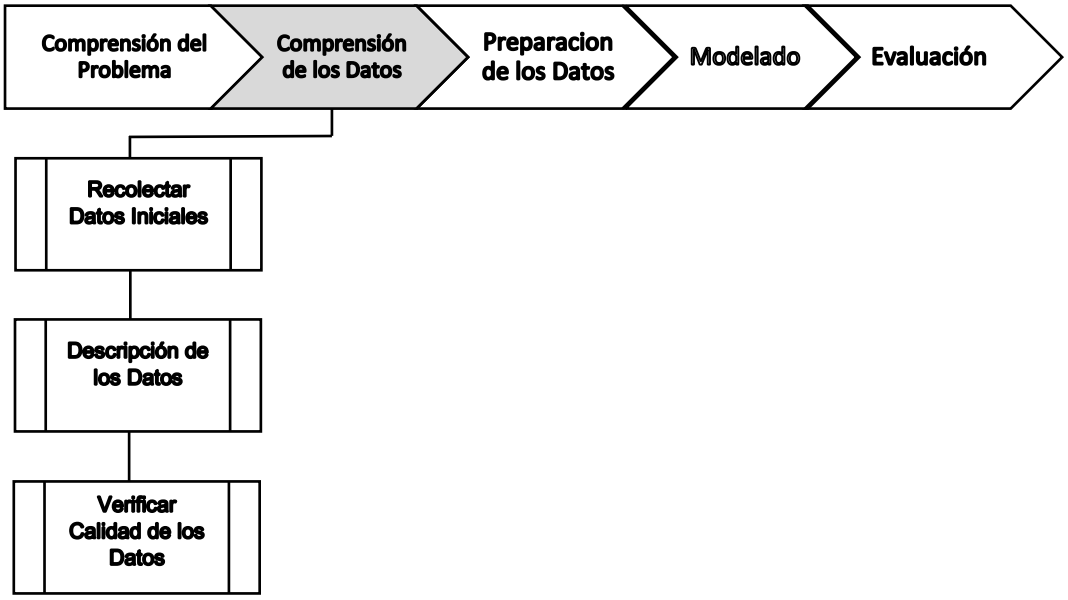
3.2. COMPRENSIÓN DE LOS DATOS

- **Recolectar datos Iniciales:** Se seleccionaron tres medios de noticias locales distintos que distribuyen contenido vía web, El nuevo Liberal, Periódico Virtual y Notivision. En un intervalo de tiempo de tres años (2015-2017).

- **Descripción de los datos:** Cada una de las noticias fue almacenada en un documento independiente, según la fuente en formato .txt en codificación UTF-8, con su respectivo título o titular, fecha y descripción o contenido.

- **Verificar la calidad de los datos:** Se identificaron noticias repetidas y se eliminaron. Se excluyeron noticias con campos no relacionados con la temática criminal.

La Figura No. 3.2 describe el proceso realizado.



Figuras No. 3.2 Comprensión de los Datos.

3.3. PREPARACIÓN DE LOS DATOS

- **Dataset:** Se distribuyeron en tres grupos que se atribuyen a la fuente de noticias, en donde se inspecciona el tipo de delito para las fuentes El periódico Virtual y Notivision para los algoritmos de clasificación supervisada.

- **Estructuración de Datos:** Cada uno de los artículos fue representado en un vector de tokens, en donde todas las palabras se pasan de mayúsculas a minúsculas y se eliminan las tildes, signos de puntuación y números.

- **Limpieza de Datos:** Se elaboró una lista de stopwords con palabras de dos o menos caracteres, también con más de veinte caracteres, palabras propias de los medios web como (http, www, xml, etc), símbolos, signos de puntuación, preposiciones, artículos, adverbios, pronombres, conjunciones, nombres propios de personas, regiones, departamentos, ciudades, municipios, barrios, localidades, apellidos, meses, días, números y conjugaciones de los verbos: ser, estar, haber, hacer, ir y tener.

- **Integración de Datos:** Se implementó el método de stemming teniendo en cuenta las siguientes reglas, aplicadas a las palabras con los valores más altos de frecuencia (*tf*) y contribución de termino (TC):

- Plural a singular.
- Femenino a masculino.
- Palabras con sufijos: ción, sión, miento a verbo infinitivo.
- Eliminar el sufijo: mente.
- Todas las conjugaciones a verbos infinitivos.

➢ **Formateo de los Datos:** Se aplicó la selección de atributos de forma independiente para cada una de las fuentes de noticias, donde inicialmente se eliminaron términos de esparcimiento proporcional a palabras que están frecuente mente en los documentos, pero no representan un valor significativo, de esta forma atenuar el efecto de dispersión que generan el gran volumen de datos. Finalmente se empleó la selección de atributos, donde se tabularon las características más relevantes de cada documento, con los ponderados de frecuencia de termino *tf* y contribución de termino *TC*.

En última medida se realizó un barrido de palabras respecto a su frecuencia de aparición en el conjunto de documentos, en donde se identificaron palabras carentes de sentido para la clasificación e identificación de cada documento y las palabras de índole judicial que no presentaban valores significativos, donde cada una fue realimentando la lista de stopwords.

La Figura No. 3.3 describe el proceso realizado.

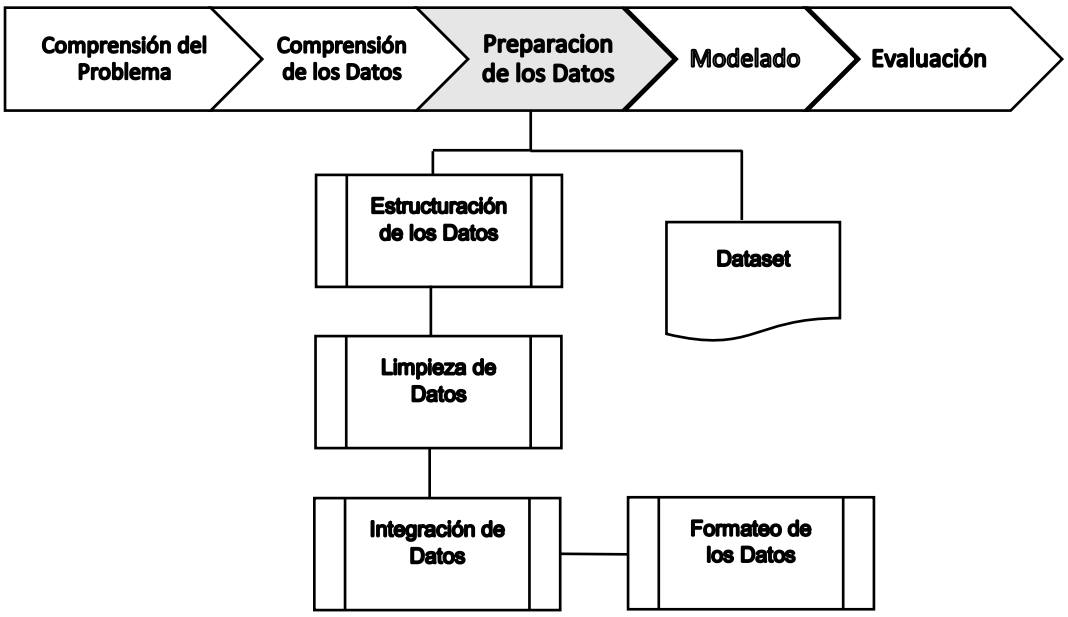


Figura No. 3.3 Preparación de los Datos.

3.4. MODELADO

- **Selección de técnicas de modelado:** Se seleccionaron tres algoritmos no supervisados de clasificación: K-Means, K-Medoids y CLARA. Y cuatro algoritmos supervisados de clasificación Naive Bayes, K-nn, SVM y Redes Neuronales (Capitulo 2).

- **Generación Plan de Prueba:** Se emplearon inicialmente los algoritmos no supervisados y posteriormente los supervisados, en donde se implementa el diccionario obtenido a partir de los métodos no supervisados en los algoritmos supervisados.

- **Construcción del Modelo**: Los algoritmos no supervisados se emplearon inicialmente a las fuentes Periodico Virtual y Notivision, debido a que se tenía conocimiento previo del número de categorías o tipos de delitos, para lo que no fue necesario estimar el número de k, en donde cada una de las técnicas fue aplicada en relación a la selección de atributos y las diferentes combinaciones de métricas. Para la fuente El nuevo Liberal, se estimó el numero óptimo de k teniendo en cuenta los métodos de Elbow y Silhouettes.

  Posterior a tener los resultados de los algoritmos no supervisados, se procedió a identificar las categorías detectadas según el tipo de delito, para lo cual se seleccionaron las palabras con los valores más altos de *tf* y TC, visualizados mediante un TagCloud y posteriormente computando estas palabras significativas para obtener el diccionario final respecto a cada una de los tipos de crímenes.

  En última medida se implementaron los algoritmos supervisados para la clasificación de los documentos, se utilizaron la fuente Periodico Virtual y Notivision, pues que se tenía información previa respecto al tipo de delitos. Dividiendo en fase de entrenamiento con el 70% de los documentos y una fase de testeo con el 30% de los documentos restante, aplicando los algoritmos de Naive Bayes, K-nn, SVM y Redes Neuronales.

  La Figura No. 3.4 describe el proceso realizado.

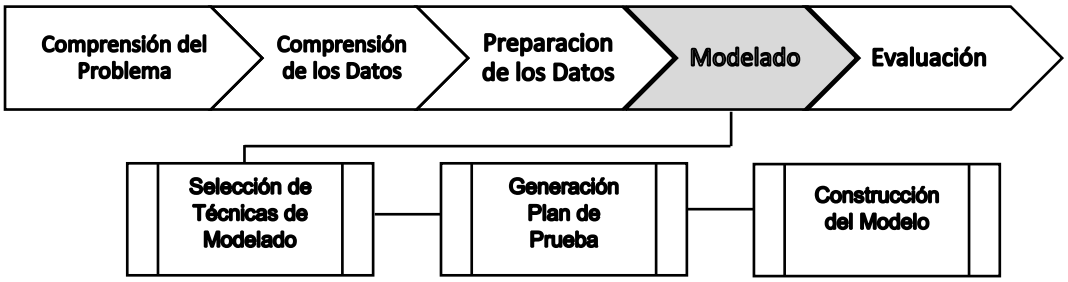


Figura No. 3.4 Modelado.

3.5. EVALUACIÓN

- **Evaluación de los resultados:** Se evaluaron los resultados obtenidos aplicando inicialmente los coeficientes de Silhouettes, que permitieron validar cada uno de los algoritmos utilizados en la clasificación no supervisada y posteriormente se calculó la matriz de confusión para visualizar la correcta agrupación de cada uno de los documentos respecto al tipo de delito.

  En los algoritmos supervisados de clasificación, se evaluaron los resultados a través de la matriz de confusión, calculando la exactitud y la tasa de error respecto a la clasificación.

- **Valoración de los resultados:** Se seleccionaron los modelos de algoritmos no supervisados con los valores más altos de coeficientes de Silhouettes y los algoritmos supervisados con los valores más altos de exactitud.

- **Modelos Aprobados:** Se elaboraron las conclusiones del trabajo.

  La Figura No. 3.5 describe el proceso realizado.

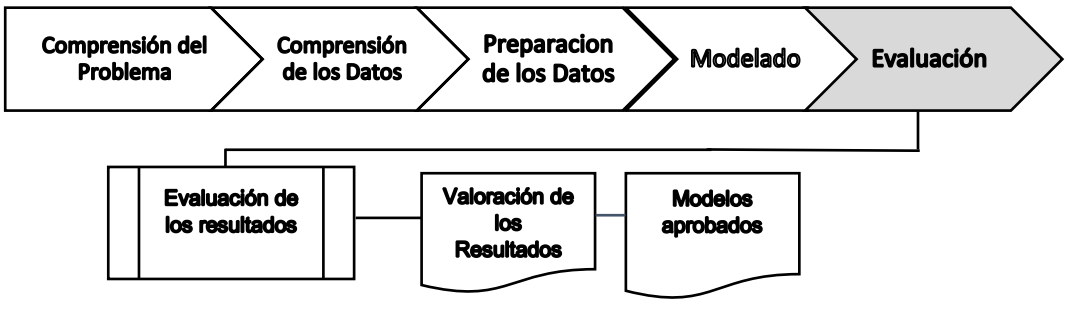


Figura No. 3.5 Evaluación

4. Resultados

4.1. CONSTRUCCIÓN DEL DATASET

Para la elaboración del dataset, se recolectaron noticias de índole criminal de tres medios de noticias de la ciudad de Popayán, El Nuevo Liberal, Periódico Virtual y Notivision, que difunden contenido a través de páginas web, en un rango de tiempo de tres años, las cuales se almacenaron por notica en un documento, en codificación UTF-8 en formato .txt.

Se obtuvo el siguiente dataset conformado por 913 noticias criminales:

Tabla No.4.1 Distribución de Noticias Seleccionadas

Fuente	Numero de Noticias	Intervalo de Tiempo
El Nuevo Liberal	312	2015-2017
Periódico Virtual	422	2013-2017
Notivision	179	2015-2017

Que en relación al número total de términos se distribuye de la siguiente manera:

Tabla No. 4.2 Distribución número de términos

Fuente	No de términos
El Nuevo Liberal	9.562
Periódico Virtual	8.223
Notivision	4.461

Por lo tanto, el dataset está conformado por 22.246 términos, que debido a la alta dimensionalidad se aplicaron las técnicas de eliminación de stopwords y se obtuvo la siguiente distribución.

Tabla No.4.3 Distribución de términos posterior a stopwords

Fuente	No de términos
El Nuevo Liberal	8.773
Periódico Virtual	7.860
Notivision	4.012

Donde se obtuvo una reducción a 20.645 términos, pero que a su vez se realizó una limpieza adicional, recorriendo a lo largo de la distribución de frecuencias de ocurrencia de los términos, y se eliminó las palabras que no representan un significado alguno en cada una de las bases de datos y anexándolo a la lista de stopwords así:

Para la fuente El Nuevo Liberal:

Palabras que presentan una frecuencia mayor a 100 se distribuyen:

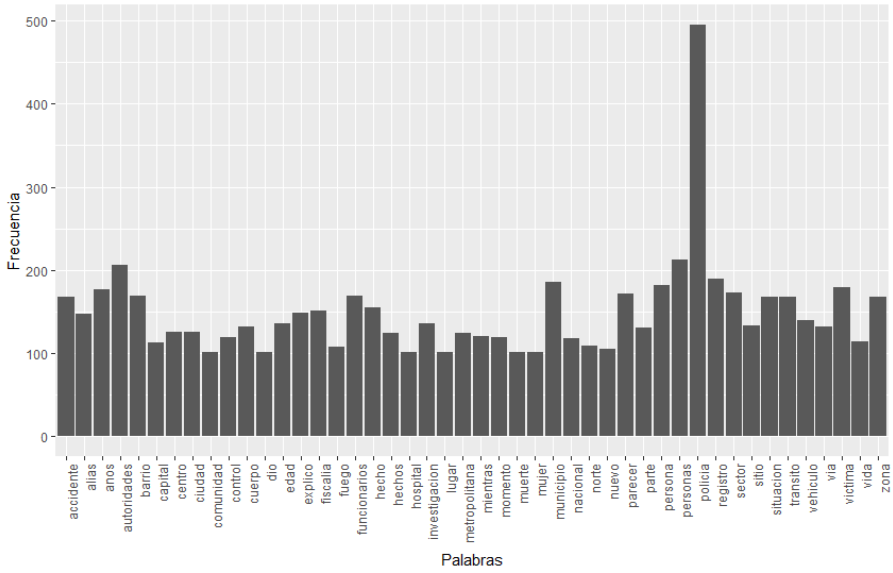


Figura No. 4.1 Frecuencias mayor a 100 de la fuente El Nuevo Liberal.

Se obtuvo la siguiente lista de palabras que no representan un valor significativo:

Tabla No. 4.4 Palabras no significativas fuente El Nuevo Liberal, frecuencia mayor a 100

anos	comunidad	hecho	mas	nacional	persona
barrio	dio	hechos	mientras	norte	zona
capital	edad	hospital	momento	nuevo	Sitio
centro	explico	investigacion	mujer	parecer	situacion
ciudad	funcionarios	lugar	municipio	parte	sector

Luego se realiza un barrido con las palabras que presentan una frecuencia menor que 100 y mayor a 50:

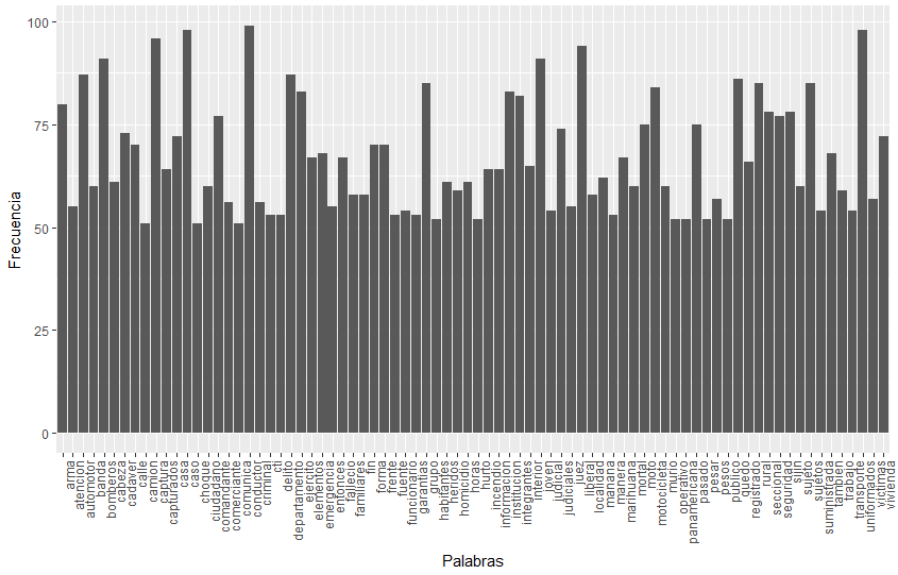


Figura No. 4.2 Frecuencia menor 100 y mayor 50 de la fuente El Nuevo Liberal

Para lo cual se obtuvo la siguiente lista de palabras que no presentan un valor significativo:

Tabla No. 4.5 Palabras no significativas con frecuencia menor a 100 y mayor a 50 de la fuente El Nuevo Liberal

atencion	elementos	frente	informacion	manana	registrado	trabajo
cabeza	entonces	fuente	institucion	manera	seccional	sujetos
ciudadano	familiares	funcionario	interior	pasado	sujeto	
comunica	fin	habitantes	joven	publico	suministrada	
departamento	forma	horas	liberal	quedo	tambien	

La distribución de palabras que presentan una frecuencia menor a 50 y mayor que 30:

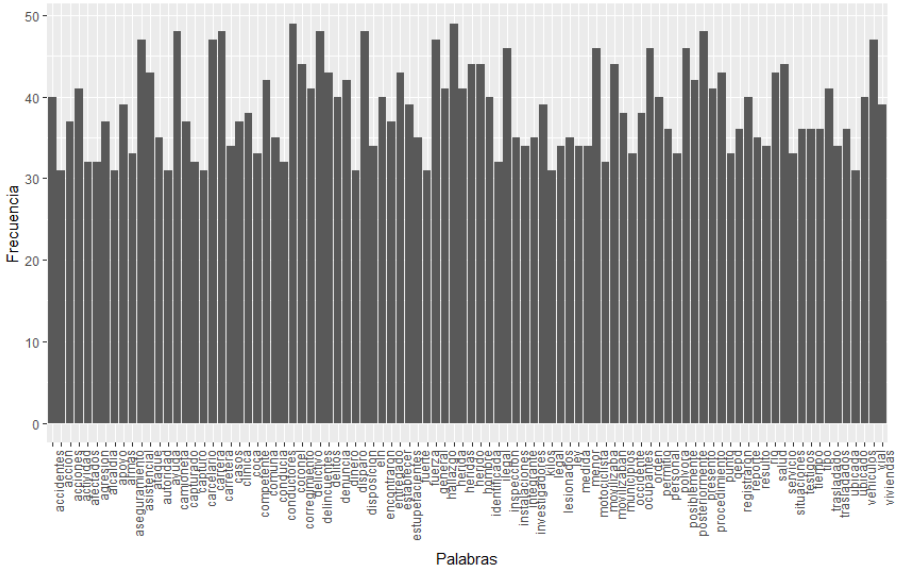


Figura No. 4.3 Frecuencia menor a 50 y mayor que 30 de la fuente El Nuevo Liberal.

De las cuales las palabras que no presentan un valor significativo:

Tabla No. 4.6 Palabras no significativas con frecuencia menor que 50 y mayor que 30 de la fuente El Nuevo Liberal.

accion	asistencial	comuna	entregado	hombre
acciones	autoridad	coronel	establecer	identificada
actividad	ayuda	corregimiento	fuerte	inspeccion
alcaldia	carrera	dinero	fuerza	instalaciones
apoyo	casos	disposicion	general	investigadores
aseguramiento	competente	encontraron	hallazgo	legal
ley	orden	presento	servicio	trasladados
medida	permitio	procedimiento	situaciones	ubicada
menor	personal	punto	testigos	ubicado
municipios	polvora	registraron	tiempo	vehiculo
occidente	posiblemente	reporte	tipo	via
ocupantes	posteriormente	resulto	trasladado	vivienda

Palabras que presenta una frecuencia menor a 30 y mayor a 20:

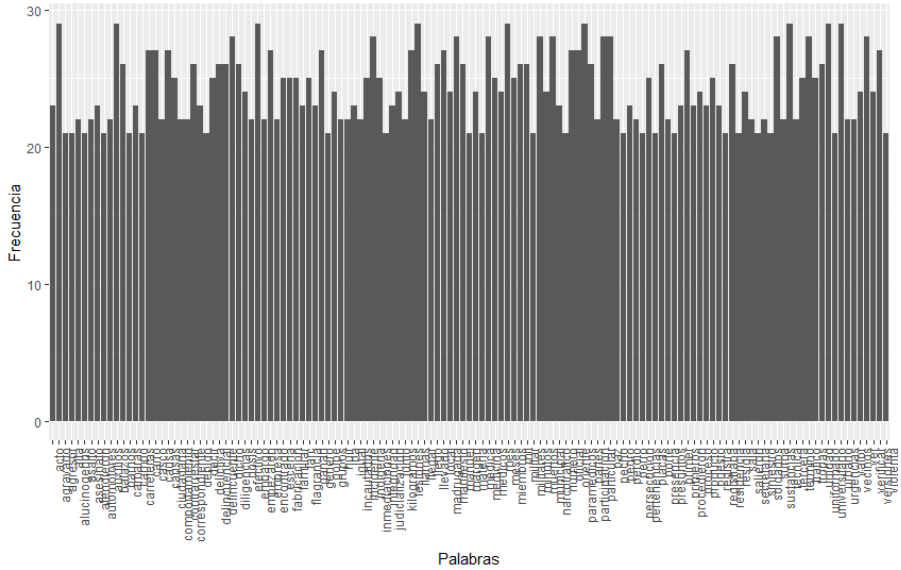


Figura No. 4.4 Frecuencia menor a 30 y mayor que 20 de la fuente El Nuevo Liberal.

Palabras que no presentan un valor significativo:

Tablo No. 4.7 Palabras no significativas con frecuencia menor a 30 y mayor a 20 de la fuente El Nuevo Liberal.

acto	causas	diligencias	ganan	llego	mes	numero
alta	ciudadana	efectivo	genero	llevado	meses	oficial
atendieron	comportamiento	embargo	iba	local	mil	oriente
brazos	contrarrestar	empezaron	igual	madrugada	militar	partes
barrios	correspondiente	empresa	indigena	manifesto	militares	participacion
camaras	debido	encontraron	inmediaciones	manuel	minutos	paz
campo	decir	encontrado	junto	margen	municipal	pecho
causa	dicho	familiar	llegar	materia	particular	pedro
perro	placas	porte	presencia	presuntos	primer	primeros
salieron	secretaria	suma	tecnica	tercera	termino	unidad
urbano	urgencias	vecinos	ver	valor	verificar	versiones
procedieron	proceso	producto	propuesta	puesto	registra	registraba
residia	sala	universitario	viviendas	rural		

Estas palabras fueron adicionadas a la lista de stopwords general y se repitió el proceso para cada una de las fuentes, reduciendo la cantidad de

términos. Las palabras significativas de las otras dos fuentes se muestran en el Anexo I.

Donde se obtuvo que el dataset se redujo a 17.586 terminos.

Luego, se identificaron palabras que poseen contenido de índole judicial pero que no son significativas a la hora de discriminar entre las diferentes temáticas criminales, que son anexadas a la lista de stopwords:

Tabla No. 4.8 Palabras de índole judicial no significativas de las tres fuentes.

captura	denuncia	metropolitana	carcel	brigada
capturados	ejercito	sijin	farc	cai
capturado	eln	uniformados	capturadas	gaula
capturo	emergencia	judicializacion	capturan	patrullas
carcelario	fiscalia	soldados	capturaron	delitos
comandante	herida	uniformados	jurisdiccion	juez
cti	heridos	capitan	juzgado	patrullero
delictivo	heridas	capturas	tropas	
delincuentas	judicial	denunciar	civil	
delito	judiciales	guerrillero	justicia	

4.2. APLICACIÓN DE ALGORITMOS DE CLUSTERING

Para realizar la clasificación de noticias según el tipo de delito se aplicaron los algoritmos de clustering así, K-Means, K-Medoids y CLARA con las dos medidas de selección de atributos, frecuencia de termino (*tf*) y contribución de termino (*TC*) y cuatro métricas de distancia, Euclidiana, Manhattan, Minkowski y Coseno para cada una de las fuentes de noticias.

Para las fuentes Periódico Virtual y Notivision, se tiene conocimiento previo del número de k, debido a que las noticias fueron seleccionadas e inspeccionando el número de categorías según los tipos de delito.

Luego de aplicar cada uno de los algoritmos de clustering se obtuvieron los coeficientes de Silhouettes promedio de toda la agrupación, para cada una de las combinaciones posibles entre algoritmos, selección de atributos y métricas de distancia, donde se obtuvo la siguiente tabla, que representa la cohesión y separación de cada uno de los grupos, para validar la clasificación:

Para la fuente Periódico Virtual:

Tabla No. 4.9 Coeficientes de Silhouettes de la fuente Periódico Virtual.

K=8	ALGORITMO	K-MEANS		K-MEDOIDS		CLARA	
	ATRIBUTO	*tf*	*TC*	*tf*	*TC*	*tf*	*TC*
DISTANCIA	Euclidiana	0.01	-0.09	0.23	0.21	0.24	0.27
	Manhattan	-0.05	-0.06	0.27	0.19	0.26	0.27
	Minkowski	-0.01	-0.1	0.23	0.21	0.24	0.27
	Coseno	0.34	0.34	**0.48**	0.41	**0.48**	0.44

De la Tabla No. 4.9 se observó que los valores más altos de índices de Silhouettes se obtienen en primera medida aplicando la métrica del coseno y los algoritmos de K-Medoids y CLARA, los cuales presentan un mejor rendimiento respecto a K-Means. Además, las medidas de selección de atributos no presentan una diferencia marcada, ya que entre las combinaciones posibles tanto *tf* como *TC* presentan valores altos uno respecto al otro.

Para visualizar la clasificación se presentan las gráficas de las agrupaciones con los valores promedio más altos de coeficiente de Silhouettes para cada uno de los algoritmos.

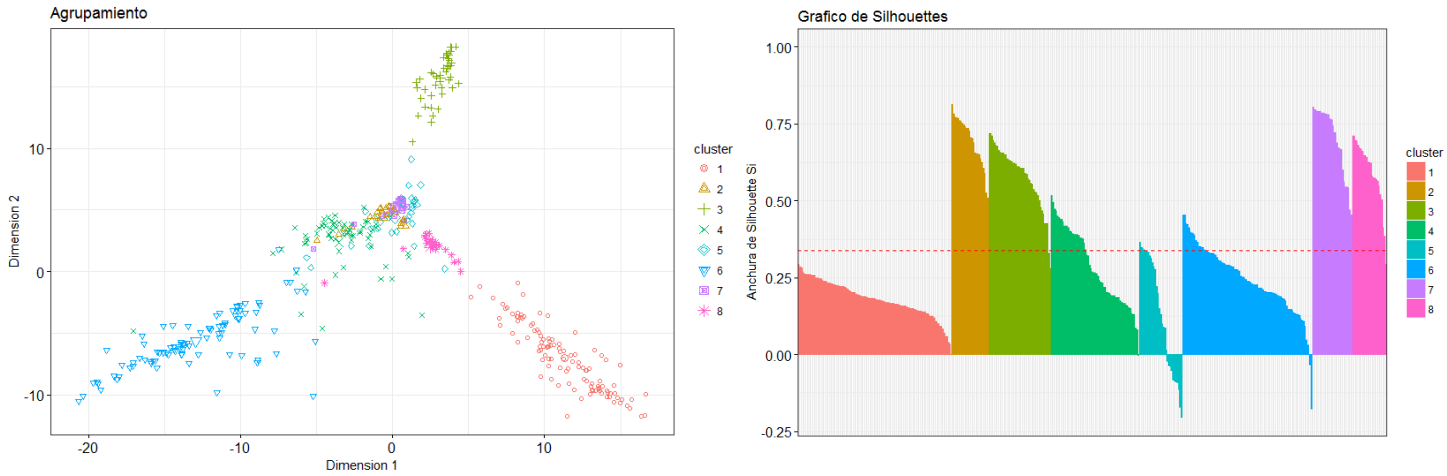


Figura No. 4.5 Agrupamiento con K-Means, *TC* y distancia del Coseno de la fuente Periódico Virtual.

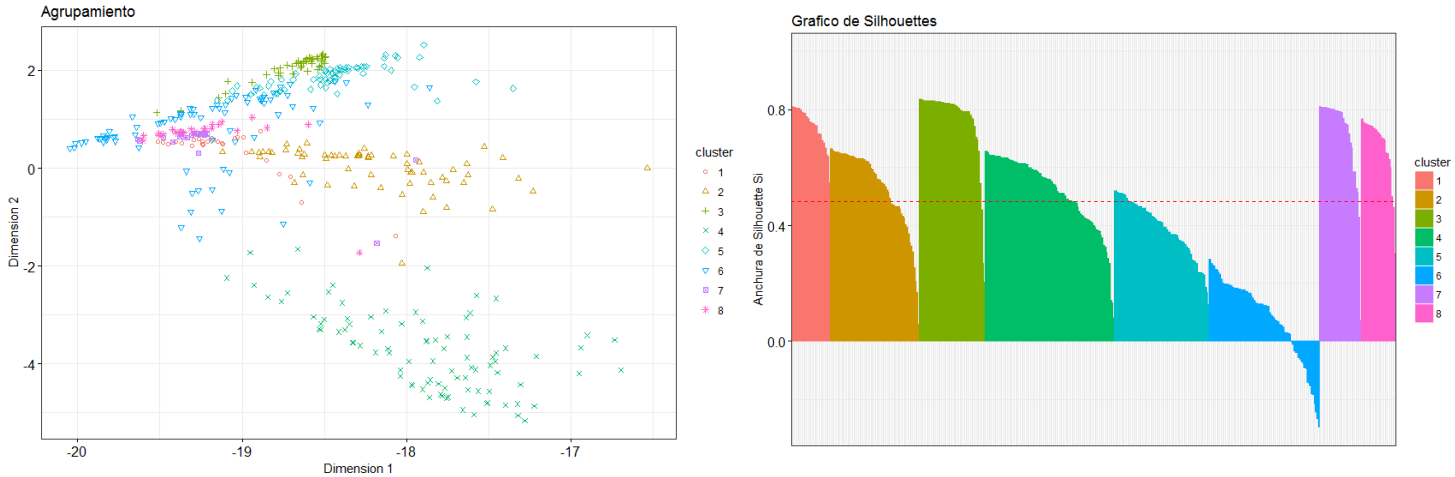


Figura No. 4.6 Agrupamiento con K-Medoids, *tf* y distancia del Coseno de la fuente Periódico Virtual.

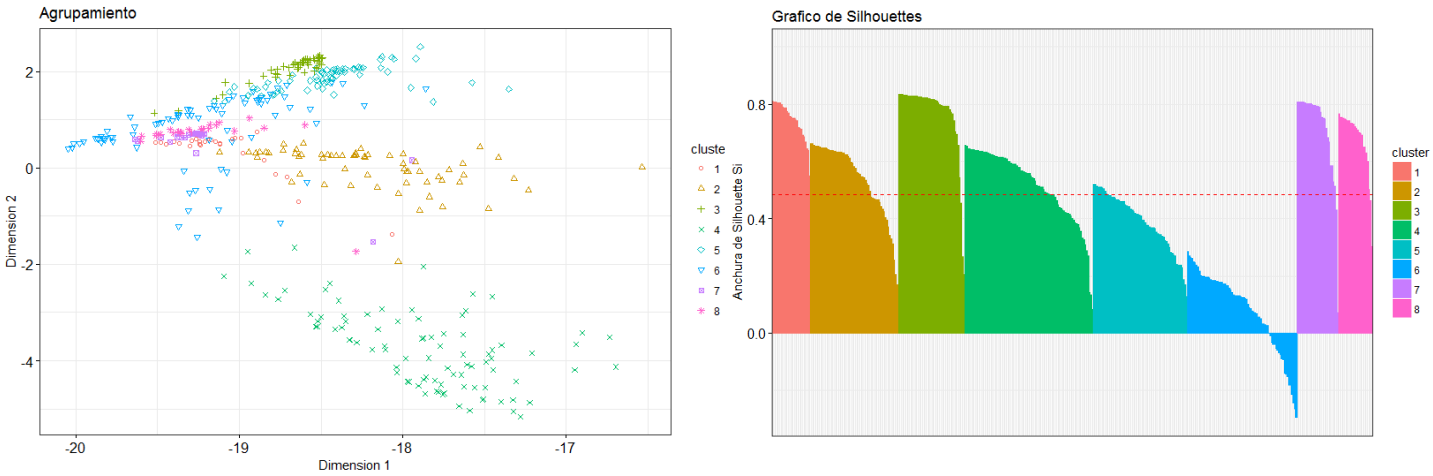


Figura No. 4.7 Agrupamiento con CLARA, *tf* y distancia del Coseno de la fuente Periódico Virtual

Las Figuras No. 4.5 - 4.6 – 4.7 a la izquierda representan el agrupamiento donde cada punto representa un documento del dataset, que se refiere a una noticia criminal, cada cluster representado por un único color equivale a una categoría o tipo de crimen. A la derecha el grafico de los índices de Silhouettes para cada uno de los documentos del dataset, separados por cada cluster, en donde se traza una línea horizontal que representa el coeficiente promedio de Silhouettes en donde el valor promedio más alto de Silhouettes se obtuvo del algoritmo CLARA.

Para la fuente Notivision:

Tabla No. 4.10 Coeficientes de Silhouettes de la fuente Notivision

K=5	ALGORITMO	K-MEANS		K-MEDOIDS		CLARA	
	ATRIBUTO	*tf*	*TC*	*tf*	*TC*	*tf*	*TC*
DISTANCIA	Euclidiana	-0.01	-0.03	0.26	0.34	0.26	0.26
	Manhattan	-0.01	-0.2	**0.36**	0.2	**0.36**	0.2
	Minkowski	-0.01	-0.03	0.26	0.34	0.26	0.26
	Coseno	0.25	0.13	0.35	0.2	0.35	0.2

La fuente Notivision se selección el valor de K=5 dado que presento un número menor de categorías.

De la Tabla No. 4.10 se observó que los valores más altos de índices de Silhouettes se obtiene aplicando la métrica del coseno en el algoritmo de K-Means y de forma análoga valores altos para la métrica manhattan en el algoritmo de K-Medoids y CLARA. En donde las medidas de selección de atributos no presentan una diferencia marcada, ya que entre las combinaciones posibles tanto *tf* como *TC* presentan valores altos uno respecto al otro.

Para la visualización del agrupamiento se presentan las gráficas en el Anexo II, con los valores promedio más altos de coeficiente de Silhouettes para cada uno de los algoritmos.

Para la fuente El Nuevo Liberal se aplicaron los métodos para estimar el número óptimo de clusters debido a que el dataset de noticias presentaba una amplia gama de noticias de ámbito criminal y teniendo en cuenta que algunas de estas noticias presentaban combinaciones entre los diferentes tipos de delitos.

En la siguiente tabla se expresa los valores óptimos de k según el método de Elbow.

Tabla No. 4.11 Numero de k según método de Elbow.

	ALGORITMO	**K-MEANS**		**K-MEDOIDS**		**CLARA**	
	ATRIBUTO	***tf***	***TC***	***tf***	***TC***	***tf***	***TC***
DISTANCIA	Euclidiana	3-4	4-8	3-5	4-5	4-6	4-5
	Manhattan	4-5	4-5	4	4	4	4-7
	Minkowski	3-4	4-8	3-5	4-5	4-6	4-5
	Coseno	2-6-8	3-5-9	6	6	6	6

En la siguiente tabla se expresan los valores óptimos de k según el método de Silhouettes.

Tabla No. 4.12 Numero de k según método de Silhouettes

	ALGORITMO	**K-MEANS**		**K-MEDOIDS**		**CLARA**	
	ATRIBUTO	***tf***	***TC***	***tf***	***TC***	***tf***	***TC***
DISTANCIA	Euclidiana	3-4-5	3-4	3-8	3-4-5	3-4-6	3-4-5
	Manhattan	3-5	3-4-5	3-4-6	3-4-5	3-4-6	3-4-5
	Minkowski	3-4-5	3-4	3-8-7	3-4-5	3-4-6	3-4-5
	Coseno	3-7-8	6-5-8	4-6	3-6-9	4-5-6	3-6-7

La graficas siguientes muestran la estimación del número óptimo de k, según los criterios de Elbow y Silhouettes.

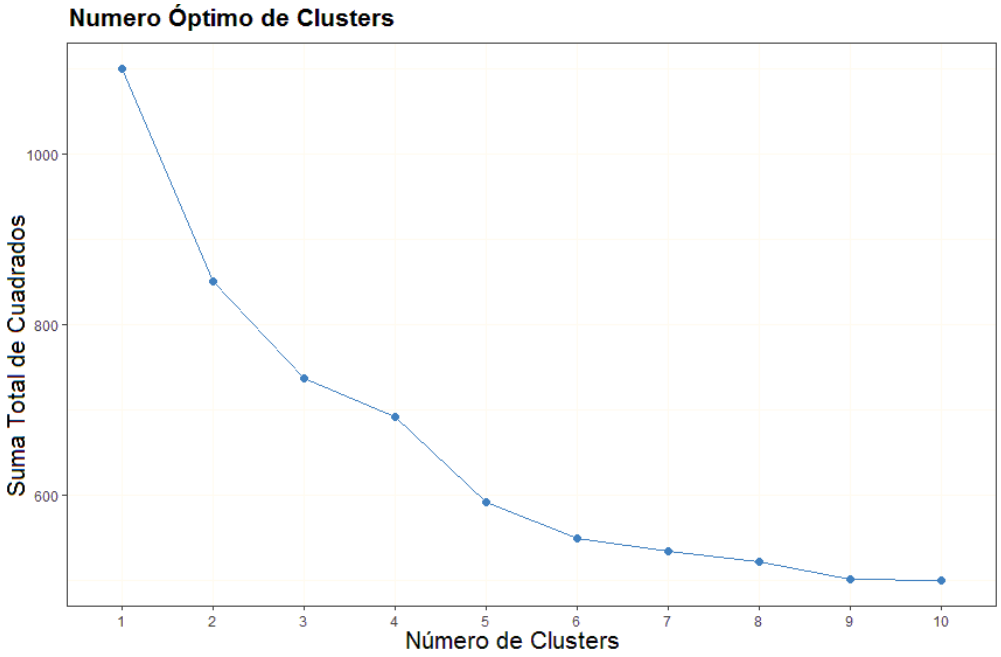


Figura No. 4.8 Numero óptimo de clusters método de Elbow

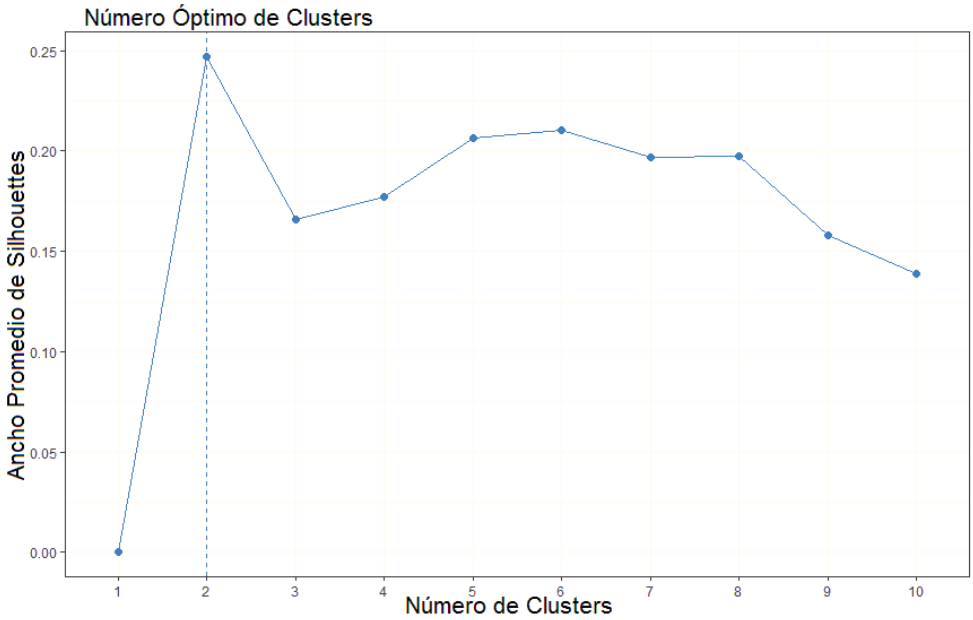


Figura No. 4.9 Numero óptimo de clusters método de Silhouettes

En la Figura No. 4.8 se representa el número óptimo de clusters con el método de Elbow para el algoritmo de K-Means con selección de atributos *tf* y métrica del coseno, se evidencia que es difícil estimar un numero óptimo de k, dado que se presentan un total de tres codos. Proceso que se repite a lo largo de todas las combinaciones posibles entre algoritmos, selección de atributos y métricas.

De forma análoga en la Figura No. 4.9 se representa el número óptimo de clusters con el método de Silhouettes para el algoritmo de K-Means con selección de atributos TC y métrica del coseno, se evidencia que el valor más alto del promedio de Silhouettes es para k=2, pero se desprecia este valor dado que representaría una clasificación binaria entre una clase positiva u otra, puesto que no permite identificar agrupaciones entre temáticas criminales. Por esta razón se consideran los siguientes coeficientes con los valores más altos, en este caso particular k=6-5-8.

Los métodos de selección de número óptimo de k presentan una alta variabilidad dependiendo del tipo de algoritmo, selección de atributos y métrica de distancia, adicional a ello, pueden producir agrupaciones no deseadas, por lo tanto, se optó por tomar como valor de k la moda entre las diferentes combinaciones de la métrica del coseno dado que es la que presenta los mejores valores, k=6.

Dado el valor de k=6, se construye la tabla de coeficientes de Silhouette para cada uno de los algoritmos:

Tabla No. 4.13 Coeficientes de Silhouettes de la fuente El Nuevo Liberal

	ALGORITMO	**K-MEANS**		**K-MEDOIDS**		**CLARA**	
	ATRIBUTO	***tf***	***TC***	***tf***	***TC***	***tf***	***TC***
DISTANCIA	Euclidiana	-0.08	-0.11	0.26	0.21	0.25	0.21
	Manhattan	-0.12	-0.21	0.26	0.2	0.29	0.2
	Minkowski	-0.06	-0.11	0.26	0.21	0.25	0.21
	Coseno	0.16	0.13	**0.33**	0.22	**0.33**	0.22

4.3. ELABORACIÓN DEL DICCIONARIO DE PALABRAS

Para identificar las diferentes temáticas criminales, se elaboró una selección de atributos teniendo en cuenta la frecuencia de termino (*tf*) y la contribución de termino (TC), se removieron los términos de esparcimiento seleccionando palabras que se encuentren en el 0.1 % de los documentos y que poseen los valores más altos de estos dos y se aplicaron las reglas de stemming a las palabras que poseían los valores más representativos, por ejemplo:

Tabla No. 4.14 Stemming de palabras más frecuentes

PALABRA	RAIZ
Violador	Violar
Violo	
Violado	
Violada	
Violadora	
Violadores	
Violadoras	
violación	
violaciones	

Teniendo en cuenta la literatura existen ocho categorías de crímenes [30], en donde se tuvieron en cuenta dos de las diferentes fuentes de noticias (dataset), dado que la fuente El Nuevo Liberal presento un contenido muy variado respecto a las categorías, con los siguientes resultados:

- La fuente Periódico Virtual, presento ocho categorías de crímenes, que en relación a la literatura no se cuenta con las categorías de fraudes y cibercrimen, pero se identificaron dos temáticas nuevas, secuestros y minería ilegal, por lo tanto, ocho corresponde al número de k.

- La fuente Notivision, presento siete categorías, en donde las temáticas de accidentes de tránsito e incendios son excluidas por presentar muy poca cantidad de documentos y no se obtuvieron fraudes y cibercrimen, por lo tanto, aplican cinco categorías que corresponden al número de k.

Se identificaron ocho tipos de crímenes constituidos por Accidentes de Tránsito, Delitos Sexuales, Homicidios, Hurtos, Incendios Intencionales,

Minería Ilegal, Secuestro y Trafico de Drogas, donde se elaboró el diccionario de palabras para cada una de las categorías comparando los valores más altos de *tf* y *TC*, representadas mediante un Tag Cloud:

4.3.1. Accidentes de tránsito

Se tienen en cuenta los valores de tf y TC respectivamente, representados por el siguiente TagCloud:

Figura No. 4.10 Tag Cloud, Frecuencia de termino vs. Contribución de Termino, Accidentes de Tránsito.

La Tabla No. 4.15 muestra las palabras más relevantes para la temática de accidentes de tránsito, la cual se caracterizó a partir de la fuente Periódico Virtual puesto que la fuente Notivision no poseía artículos respecto a la temática.

Tabla No. 4.15 Palabras más significativas de Accidentes de Tránsito

Periódico Virtual			
tf		***TC***	
accidente	motociclista	volco	fallecio
transito	cadaver	choque	pasajero
choque	falla	muerto	mecanicas
bomberos	socorristas	bomberos	eco
volco	fallecio	violenta	ocurrio
muerto	mecanicas	motociclista	cadaver
pasajero	carril	colision	movilizaba
violenta	conducia	kilometro	carril
velocidad	percance	falla	volqueta
kilometro	determinar	velocidad	levantamiento
colision	conductores	exceso	socorristas
exceso	movilizaba	coomotoristas	aparatoso

El diccionario definitivo que computa las palabras más significativas de cada una de las medidas de selección de atributos es representado mediante el siguiente TagCloud:

Figura No. 4.11 Tag Cloud del diccionario de Accidentes de Tránsito

4.3.2. Delitos Sexuales

Se tienen en cuenta los valores de tf y TC respectivamente, representados por el siguiente TagCloud:

Figura No. 4.12 Tag Cloud, Frecuencia de termino vs. Contribución de Termino, Delitos Sexuales.

Las palabras más relevantes para la temática de Delitos Sexuales, la cual se caracterizó a partir de las fuentes Periódico Virtual y Notivision se muestran en los Anexos III, Tabla No. 8.3.

El diccionario definitivo que computa las palabras más significativas de cada una de las medidas de selección de atributos es representado mediante el siguiente TagCloud:

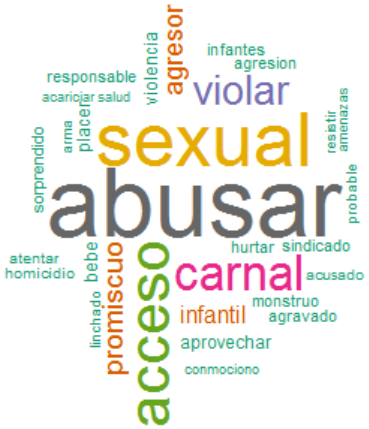


Figura No. 4.13 Tag Cloud del diccionario de Delitos Sexuales.

4.3.3. Homicidios

Se tienen en cuenta los valores de tf y TC respectivamente, representados por el siguiente TagCloud:

Figura No. 4.14 Tag Cloud, Frecuencia de termino vs. Contribución de Termino, Homicidios.

Las palabras más relevantes para la temática de Homicidios, la cual se caracterizó a partir de las fuentes Periódico Virtual y Notivision se presentan en los Anexos III, Tabla No. 8.4.

El diccionario definitivo que computa las palabras más significativas de cada una de las medidas de selección de atributos es representado mediante el siguiente TagCloud:

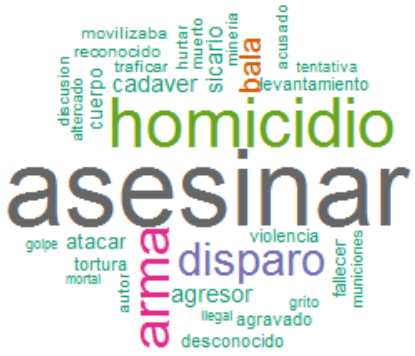


Figura No. 4.15 Tag Cloud del diccionario de Homicidios.

4.3.4. Hurtos

Se tienen en cuenta los valores de tf y TC respectivamente, representados por el siguiente TagCloud:

Figura No. 4.16 Tag Cloud, Frecuencia de termino vs. Contribución de Termino, Hurtos.

Las palabras más relevantes para la temática de Hurtos, la cual se caracterizó a partir de las fuentes Periódico Virtual y Notivision se muestran en los Anexos III, Tabla No. 8.5.

El diccionario definitivo que computa las palabras más significativas de cada una de las medidas de selección de atributos es representado mediante el siguiente TagCloud:

Figura No. 4.17 Tag Cloud del diccionario de Hurtos.

4.3.5. Incendios Intencionales

Se tienen en cuenta los valores de tf y TC respectivamente, representados por el siguiente TagCloud:

Figura No. 4.18 Tag Cloud, Frecuencia de termino vs. Contribución de Termino, Incendios Intencionales.

Las palabras más relevantes para la temática de Incendios Intencionales, la cual se caracterizó a partir de la fuente Periódico Virtual, dado que la fuente Notivision, no presentaba la tematica, se muestran en los Anexos III, Tabla No. 8.6.

El diccionario definitivo que computa las palabras más significativas de cada una de las medidas de selección de atributos es representado mediante el siguiente TagCloud:

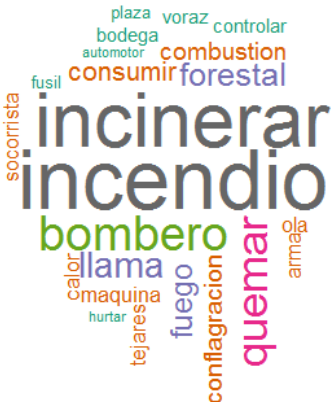


Figura No. 4.19 Tag Cloud del diccionario de Incendios Intencionales.

4.3.6. Minería ilegal

Se tienen en cuenta los valores de tf y TC respectivamente, representados por el siguiente TagCloud:

Figura No. 4.20 Tag Cloud, Frecuencia de termino vs. Contribución de Termino, Mineria Ilegal.

Las palabras más relevantes para la temática de Minería ilegal, la cual se caracterizó a partir de las fuentes Periódico Virtual y Notivision, se muestran en los Anexos III, Tabla No. 8.7.

El diccionario definitivo que computa las palabras más significativas de cada una de las medidas de selección de atributos es representado mediante el siguiente TagCloud:

Figura No. 4.21 Tag Cloud del diccionario de Minería Ilegal.

4.3.7. Secuestros

Se tienen en cuenta los valores de tf y TC respectivamente, representados por el siguiente TagCloud:

Figura No. 4.22 Tag Cloud, Frecuencia de termino vs. Contribución de Termino, Secuestros.

Las palabras más relevantes para la temática de Secuestros, la cual se caracterizó a partir de las fuentes Periódico Virtual y Notivision, se muestran en los Anexos III, Tabla 8.8.

El diccionario definitivo que computa las palabras más significativas de cada una de las medidas de selección de atributos es representado mediante el siguiente TagCloud:

Figura No. 4.23 Tag Cloud del diccionario de Secuestros.

4.3.8. Trafico de Drogas

Se tienen en cuenta los valores de tf y TC respectivamente, representados por el siguiente TagCloud:

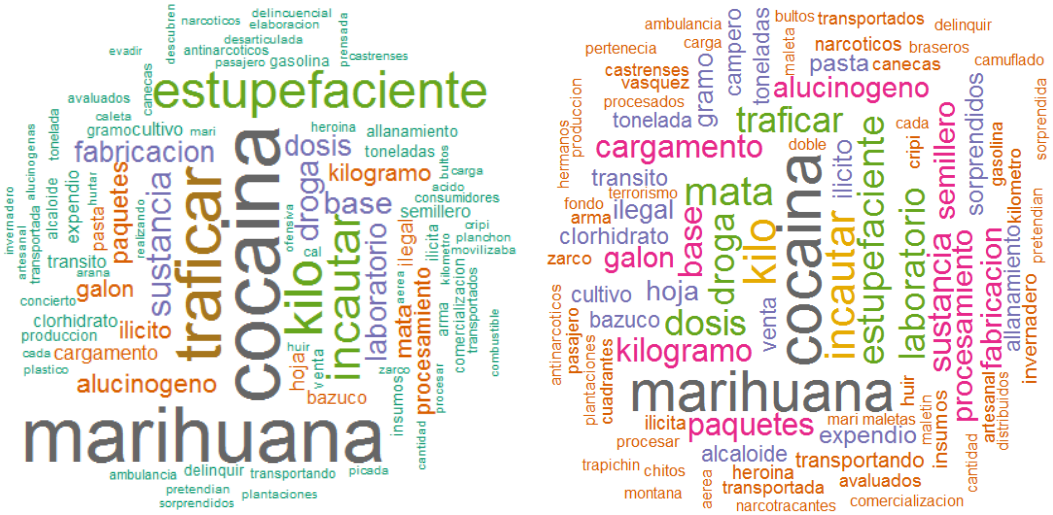


Figura No. 4.24 Tag Cloud, Frecuencia de termino vs. Contribución de Termino, Trafico de Drogas.

Las palabras más relevantes para la temática de Secuestros, la cual se caracterizó a partir de las fuentes Periódico Virtual y Notivision, se muestran en los Anexos III, Tabla No. 8.9.

El diccionario definitivo que computa las palabras más significativas de cada una de las medidas de selección de atributos es representado mediante el siguiente TagCloud:

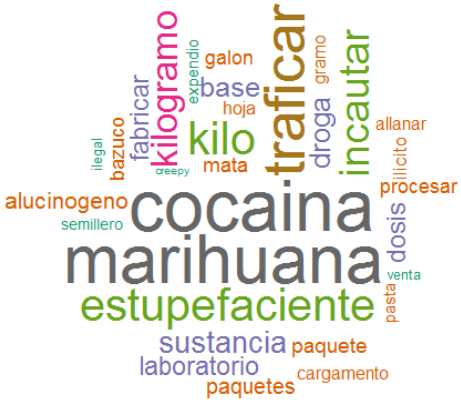


Figura No. 4.25 Tag Cloud del diccionario de Trafico de Drogas

4.4. MATRIZ DE CONFUSIÓN CLUSTERING

En la sección 4.2. se presentaron diferentes algoritmos de clustering evaluados a partir de los coeficientes de validación interna, permitiendo verificar una buena agrupación de los documentos, pero no se observó cómo estaban siendo distribuidos cada uno de los documentos, por este motivo y a raíz de que se tenía conocimiento previo de las categorías a las que pertenecía cada documento para las fuentes Periódico Virtual y Notivision, se realiza la inspección por medio de la matriz de confusión.

Se evalúa el algoritmo de K-Means con selección de atributos TC y métrica del coseno, para visualizar como se realizó la clasificación:

Tabla No. 4.16 Matriz de confusión Clustering

Real	Predicción								Recall
	Droga	**Secuestro**	**Tránsito**	**Hurto**	**Incendio**	**Homicidio**	**Sexual**	**Minería**	
Droga	110	0	0	0	1	0	0	0	99%
Secuestro	0	27	0	0	0	0	0	0	100%
Tránsito	0	0	44	0	3	0	0	0	93,61%
Hurto	0	0	0	61	8	2	0	0	85,91%
Incendio	0	0	0	0	19	0	0	0	100%
Homicidio	0	0	0	1	0	91	0	1	97,84%
Sexual	0	0	0	1	0	0	29	0	96,66%
Minería	0	0	0	0	0	0	0	22	100%
Precision	100%	100%	100%	96,82%	63,33%	97,84%	100%	95,65%	

Se aplican las medidas de evaluación obteniendo valores muy buenos, validando de esta manera la clasificación.

Exactitud = 95,95%

Tasa de Error = 4,15%

4.5. APLICACIÓN DE CLASIFICADORES

Teniendo en cuenta los resultados obtenidos luego de aplicar los algoritmos de clustering y asumiendo que cada cluster representa una categoría, se implementan los algoritmos de clasificación supervisada Naive Bayes, K-nn, SVM y Redes Neuronales, para las fuentes Periodico Virtual y Notivision.

Para la fuente El Nuevo Liberal no se identificó una diferencia marcada en las categorías dado que los documentos presentaban contenidos variados respecto a los tipos de delitos y fueron agrupados por características similares en sus atributos, por este motivo no se implementó los métodos supervisados.

Como criterio de selección de atributos se utilizó las medidas de tf y TC, en donde a cada uno de los algoritmos se realizó una fase de entrenamiento del modelo, comprendida por el 70% de los documentos y una fase en que se testea o prueba correspondiente al 30% restante de los documentos y se evaluaron a través de la matriz de confusión.

4.5.1. Clasificador Naive Bayes

Para aplicar el clasificador Naive Bayes se tiene en cuenta la corrección de Laplace para los valores cero, tomando el valor de uno, donde se obtiene los siguientes valores:

Tabla No. 4.17 Exactitud y tasa de error de Naive Bayes

	Periódico Virtual		**Notivision**	
	tf	***TC***	***tf***	***TC***
Exactitud	21,58%	**35,2%**	13,20%	33,33%
Tasa de Error	76,92%	65,07%	86,79%	60,57%

Para visualizar como se realiza la clasificación la Tabla No. 4.18 describe la matriz de confusión para la fuente Periódico Virtual con selección de atributos *TC*.

Tabla No. 4.18 Matriz de confusión Naive Bayes, Periódico Virtual, TC

Real	Predicción								Recall
	Tránsito	Droga	Homicidio	Hurto	Incendio	Minería	Secuestro	Sexual	
Tránsito	10	0	0	0	0	0	1	0	90,90%
Droga	0	2	0	0	16	0	6	2	7,69%
Homicidio	0	0	0	0	14	0	16	0	
Hurto	2	0	0	0	5	0	17	0	
Incendio	0	0	0	0	5	0	1	0	83,33%
Minería	0	0	0	0	0	9	2	0	81,81%
Secuestro	0	0	0	0	1	0	10	0	90,90%
Sexual	0	0	0	0	0	0	0	8	100%
Precision	83,33%	100%			13,88%	100%	18,86%	80%	

4.5.2. Clasificador K-nn

Para aplicar el clasificador K-nn se tiene en cuenta el número de vecinos más cercanos k=5, puesto que constituye un buen valor conforme a la cantidad de documentos, donde se obtiene los siguientes valores:

Tabla No. 4.19 Exactitud y tasa de error de K-nn.

	Periódico Virtual		**Notivision**	
	tf	***TC***	***tf***	***TC***
Exactitud	80,15%	33,33%	**80,76%**	46,15%
Tasa de Error	19,84%	66,66%	19,23%	53,84%

Para visualizar como se realiza la clasificación la Tabla No. 4.20 describe la matriz de confusión para la fuente Notivision con selección de atributos *tf*.

Tabla No. 4.20 Matriz de confusión K-nn, Notivision *tf*

Real	**Predicción**					**Recall**
	Sexual	**Homicidio**	**Hurto**	**Secuestro**	**Droga**	
Sexual	2	0	1	0	0	66,66%
Homicidio	0	2	3	0	0	60%
Hurto	0	0	16	0	0	100%
Secuestro	0	0	4	0	0	
Droga	0	0	2	0	22	91,66%
Precision	100%	100%	61,53%		100%	

4.5.3. Clasificador SVM

Para aplicar el clasificador basado en las máquinas de soporte vectorial se realizó la configuración uno vs uno, dado que constituye una clasificación multiclase, en donde realizaron $n(n-1)/2$ clasificadores y computando los votos para cada una de las configuraciones, se aplicó un kernel Gausiano y el factor C=1, donde se obtuvo los siguientes valores:

Tabla No. 4.21 Exactitud y tasa de error de SVM.

	Periódico Virtual		**Notivision**	
	tf	***TC***	***tf***	***TC***
Exactitud	95,27%	**96%**	94,33%	89,43%
Tasa de Error	4,95%	5,51%	5,66%	10,56%

Para visualizar como se realiza la clasificación la Tabla No. 4.22 describe la matriz de confusión para la fuente Periodico Virtual con selección de atributos *tf*.

Tabla No. 4.22 Matriz de confusión SVM, Periódico Virtual, *tf*.

Real	Predicción								Recall
	Tránsito	**Droga**	**Homicidio**	**Hurto**	**Incendio**	**Minería**	**Secuestro**	**Sexual**	
Tránsito	18	1	0	0	0	0	0	0	94,73%
Droga	0	38	0	1	0	0	0	0	97,43%
Homicidio	0	0	27	0	0	1	0	0	96,42%
Hurto	0	0	1	10	0	0	0	0	90,90%
Incendio	0	0	0	0	6	0	0	0	100%
Minería	0	1	0	1	0	5	0	0	71,42%
Secuestro	0	0	0	0	0	0	9	0	100%
Sexual	0	0	0	0	0	0	0	8	100%
Precision	100%	95%	96,42%	83,33%	100%	83,33%	100%	100%	

4.5.4. Clasificador con Redes Neuronales

Para aplicar el clasificador basado en redes neuronales se realiza la configuración feedforward, con dos unidades en la capa oculta, donde se obtuvo los siguientes valores:

Tabla No. 4.23 Exactitud y tasa de error de Redes Neuronales.

	Periódico Virtual		**Notivision**	
	tf	***TC***	***tf***	***TC***
Exactitud	94,45%	69,53%	**95,45%**	91,86%
Tasa de Error	5,95%	29,92%	4,54%	8,84%

Para visualizar como se realiza la clasificación la Tabla No. 4.24 describe la matriz de confusión para la fuente Notivision con selección de atributos *tf*.

Tabla No. 4.24 Matriz de confusión Redes Neuronales, Notivision, *tf*.

Real	**Predicción**					**Recall**
	Sexual	**Homicidio**	**Hurto**	**Secuestro**	**Droga**	
Sexual	1	0	0	1	0	50%
Homicidio	0	2	0	0	0	100%
Hurto	0	0	5	0	0	100%
Secuestro	0	0	0	1	0	100%
Droga	0	0	0	0	12	100%
Precision	100%	100%	100%	50%	100%	

5. Conclusiones

Las técnicas de minería de texto son una herramienta que cada vez toma más protagonismo en el procesamiento de volúmenes grandes de información, respondiendo de forma eficiente y enriqueciendo el análisis de crímenes. Este estudio ha presentado siete diferentes tipos de algoritmos que permiten realizar de forma sistemática la identificación y clasificación de tipos de crímenes, demostrando que los procedimientos y la metodología son una opción viable para resolver problemas de clasificación.

Para almacenar y registrar datasets que permitan la aplicación de técnicas de minería de texto, se debe hacer un énfasis especial en la limpieza y filtrado de la información, dado que es una tarea fundamental para obtener buenos resultados y que se pueda implementar de forma eficiente los diferentes algoritmos. Por su parte, los algoritmos de clustering aplicados a minería de texto permiten identificar de forma eficiente diferentes tipos de crímenes, contribuyendo a la construcción de diccionarios de palabras que son realmente útiles para procesar información referente a las temáticas criminales, dado que se facilitan la tarea de clasificación, donde se obtuvo 8 diccionarios de palabras en crímenes. En este sentido, los algoritmos de K-Medoids y CLARA son los más eficientes en las tareas de clustering, donde se obtuvieron desempeños altos en las tareas de agrupación $s(i)$=0.48 y clasificación, exactitud = 95,95%.

Se debe hacer un buen énfasis en la selección de atributos y en la selección de las métricas de distancia, dado que los algoritmos tanto supervisados como no supervisados, son muy sensibles a cada una de estas medidas, en donde se evidencio que la distancia del coseno es la mejor métrica de distancia. Así, los algoritmos supervisados respondieron de forma eficiente en la tarea de clasificación de documentos respecto a los tipos de crímenes, obteniendo el mejor rendimiento a partir de los algoritmos de máquina de soporte vectorial SVM y redes neuronales, con valores de exactitud = 96%.

Las herramientas de visualización como los TagCloud son realmente útiles, dado que permiten distinguir fácilmente el contenido y el grado de relevancia de cada uno de los términos de un documento.

Se encontraron dos nuevas clases de tipos de crímenes: Secuestros y Minería Ilegal que no estaban reportadas en la literatura, debido a que el departamento del Cauca se está enfrentando estas problemáticas con un alto crecimiento.

6. Trabajo Futuro

Dada la alta variabilidad del contenido de los diferentes documentos, los algoritmos de clustering K-Means, K-Medoids y CLARA en algunas ocasiones no presentan distribuciones uniformes, reduciendo su eficiencia, por lo cual se deben aplicar algoritmos de clustering DBSCAN basados en la densidad, donde se obtienen clusters con formas geométricas arbitrarias para evitar el solapamiento entre categorías y presentando una mayor robustez frente al ruido.

Dado que la identificación y clasificación de categorías de tipos de crímenes presenta un análisis inicial en la asociación de crímenes, se debe hacer un enfoque orientado a la caracterización del “Modus Operandi”, para poder identificar y asociar los crímenes a sospechosos y grupos delictivos, facilitando el análisis a los expertos en criminalística e identificando características particulares para optimizar la táctica policial.

Realizar un análisis complementario de los tipos de delito, en donde se tiene en cuenta los factores espacio-temporales para desarrollar herramientas de visualización y georreferenciación (SIG) de cada uno de los crímenes.

Desarrollar un software donde se aplique de forma automática cada uno de los enfoques y algoritmos propuestos.

7. Referencias

[1] Ministerio de Defensa, Viceministro para las Políticas y Asuntos Internacionales, “Logros de la Política de Defensa y Seguridad. Todos por un Nuevo País”, Dirección de Estudios Estratégicos – Grupo de información Estadística, Junio 2016.

[2] Nael T. Elyezjy, Alaa M. Elhaless, “Investigating Crimes using Text Mining and Network Analysis”, The Islamic University of Gaza, septiembre 2015.

[3] C. Aguilar, “Curso de Procesamiento del Lenguaje Natural”, Pontificia Universidad Católica de Chile, 2012, tomado de:

“http://cesaraguilar.weebly.com/uploads/2/7/7/5/2775690/pln_uc_09.pdf”.

[4] Torres D. A., “Diseño y Aplicación de una Metodología para el Análisis de Noticias Policiales Utilizando Minería de Textos”, Universidad de Chile, Junio, 2013.

[5] P. Chapman (NCR), J. Clinton (SPSS), R. Kerber (NCR), T. Khabaza (SPSS), T. Reinartz (DaimlerChrysler), C. Shearer (SPSS) and R Wirth (DaimlerChrysler), “CRISP-DM 1.0 Step-by-step data mining guide”, Technical Report, CRISP-DM Consortium, 2000.

[6] Vega, María Guadalupe de la, “La Noticia Policial: Un género Periodístico Diferente: un análisis del cómo y porqué del lenguaje periodístico usado por Clarín y Crónica en el caso García Belsunce”, Universidad de San Andrés, Argentina, Buenos Aires, 2014.

[7] Feldman R., Sanger J., “The Text MIning Handbook”, Cambridge University Press, 2007.

[8] G. S. A. W. y C. Y. , “A Vector Space Model for Automatic Indexing”, Association for Computing Machinery, 1975.

[9] Shantanu Godbole, Indrajit Bhattacharya, Ajay Gupta, Ashish Verma, “Building Re-usable Dictionary Repositories for Real-world Text Mining”, IBM Research, India, 2010.

[10] S. M. W. T. Z. y N. I. , “Fundamentals of Predictive Text Mining”, London: Spinger, 2015.

[11] M. B. , ”Principles of Data Mining”, London: Springer, 2007.

[12] Vallejo, D. F., “Clustering de Documentos con Restricción de Tamaño”, Universidad Politécnica de Valencia, 2015-2016.

[13] Charu C. Aggarwal • ChengXiang Zhai, “Mining Text Data”,Springer, 2012.

[14] Robertson, S., “Understanding Inverse Document Frequency: On theoretical arguments for IDF”, Microsoft Research, University of Cambringe, City University, London.

[15] Pino, J., “Tutorial de R-Text Mining Solution”, Universidad de Málaga, Andalucia Tech, Escuela Técnica Superior de Ingeniería Industrial, España, 2016.

[16] Álvarez, P. A., Vega, I. F., Fernández, E., “Análisis Comparativo de las Medidas de Semejanza Aplicadas al Contenido de Documentos Web”, Universidad Autonoma de Sinaloa, Mexico, 2007.

[17] Perlibakas, V., “Distance measures for PCA-based face recognition”, Kaunas University of Technology, Image Processing and Multimedia Laboratory, Lithuania, 2003.

[18] Venkatesan, R., “Cluster analysis for segmentation”, Universidad de Virginia, Darden Business Publishing, 2007.

[19] Kaufman, L., Rousseuw, P. J., “Clustering by means of medoids. Statistical Data Analysis Based on the L1–Norm and Related Methods”, North-Holland, 1987.

[20] Kaufman, L., Rousseuw, P. J., “Finding Groups in Data: An Introduction to Cluster Analysis”, 1990.

[21] Rousseeuw, P. J., “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. Journal of Computatational and Applied Mathematics, University of Fribourg, 1987.

[22] Kassambara, A., “Practical Guide To Cluster Analysis in R”, 2017.

[23] Garcia, C., Gomez, I., “Algoritmos de Aprendizaje: Knn y Kmeans”, Universidad Carlos III de Madrid.

[24] J. Weston, C. Watkins., “Multi-class Support Vector Machines”, 1998.

[25] W. Powers, D., “Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation”, School of Informatics and Engineering, Flinders University of South Australia, 2007.

[26] Sanger, T., “Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network”, Massachusetts Institute of Technology, 1989.

[27] Keyvanpour , M., Javideh, M., Ebrahimi , M.,“Detecting and investigating crime by means of data mining: a general crime matching framework”, Alzahra University, Tehran, Iran, 2010.

[28] Ananyan, S., “Crime Pattern Analysis Through Text Mining”, Megaputer Intelligence, 2004.

[29] Kianmehr, K., Alhajj, R., “Effectiveness of support vector machine for crime hot-spots prediction”, University of Calgary, Global University, Canada, Lebano, 2008.

[30] H. CHEN, W. CHUNG, J. XU, G. WANG, Y. QIN and M. CHAU. “Crime data mining: a general framework and some examples”, 2004.

8. Anexos

ANEXO I

Tabla No. 8.1 Palabras no significativas de la fuente Periodico Virtual.

acuerdo	grupo	presunta	dar	aproximadamente
adolescente	grupos	pues	operativo	area
bajo	individuos	responsables	garantias	calle
bus	informo	traves	operaciones	colombiano
cargos	investigan	ultimas	control	comunidades
casa	operación	operativos	adolescentes	dejado
establecimiento	expreso	finalmente	llamado	lider
llevabada	poblacion	ubicación		

Tabla No. 8.2 Palabras no significativas de la fuente Notivision.

caucana	convivencia	division	inseguridad	libertad
clase	coordinacion	encontrando	integrantes	logra
color	cuadrante	estructura	inteligencia	marco
conjunto	dedicaban	funcion	intervercion	miembros
contenian	derechos	importante	investigativo	operación
controles	destino	individuo	invita	oportuna
pinzon	plan	produjo	prueba	reaccion
realizado	recibir	seguridad	seguros	

ANEXO II

Las Figuras No. 8.1 – 8.2 – 8.3 a la izquierda representan el agrupamiento donde cada punto representa un documento del dataset, que se refiere a una noticia criminal, cada cluster representado por un único color equivale a una categoría o tipo de crimen. A la derecha el grafico de los índices de Silhouettes para cada uno de los documentos del dataset, separados por cada cluster, en donde se traza una línea horizontal que representa el coeficiente promedio de Silhouettes en donde el valor promedio más alto de Silhouettes se obtuvo en igual medida de los algoritmos K-Medoids y CLARA.

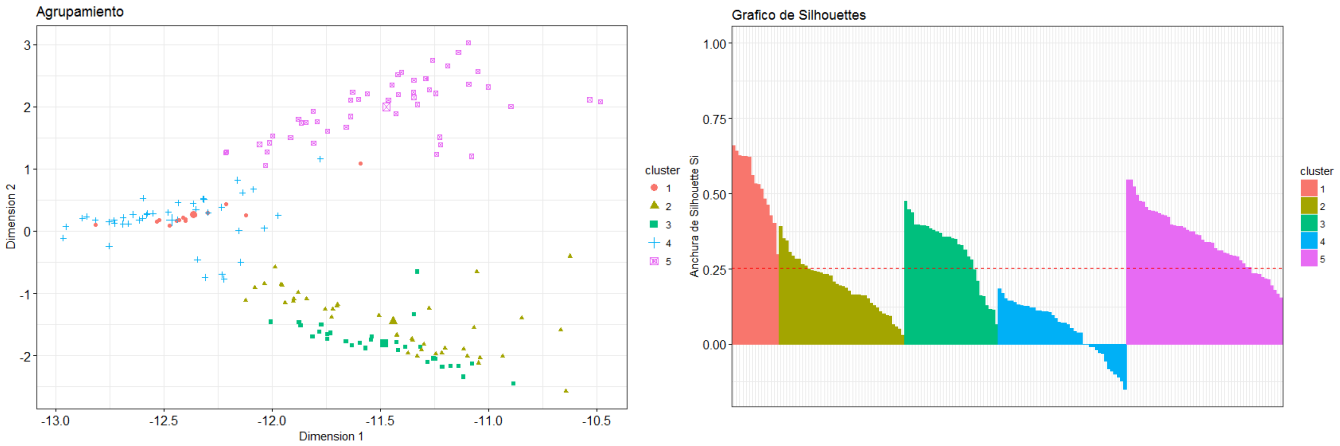


Figura No. 8.1 Agrupamiento con K-Means, tf y distancia del Coseno de la fuente Notivison

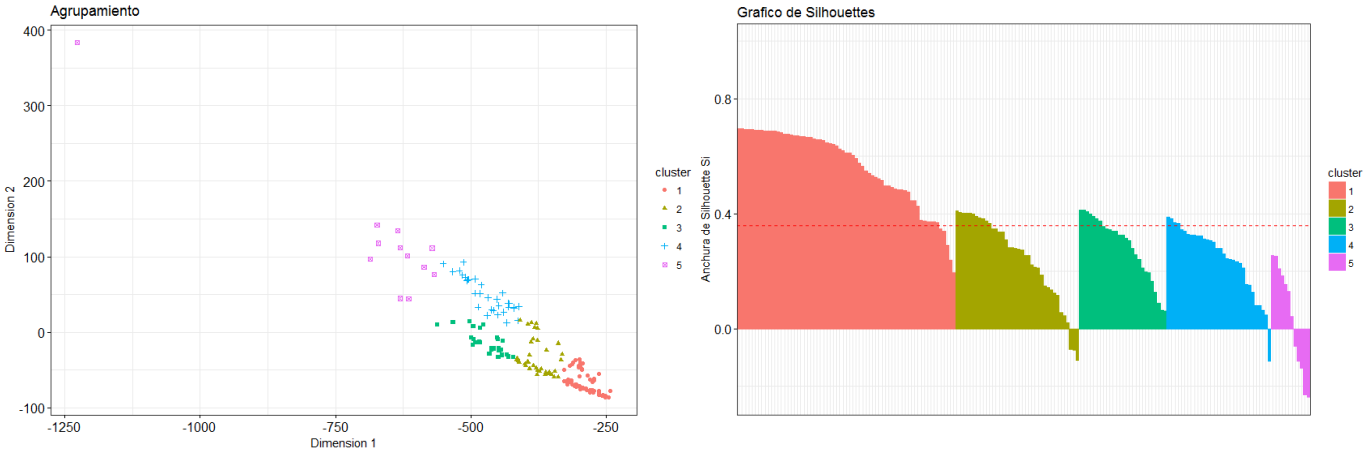


Figura No. 8.2 Agrupamiento con K-Medoids, tf y distancia Manhattan de la fuente Notivision.

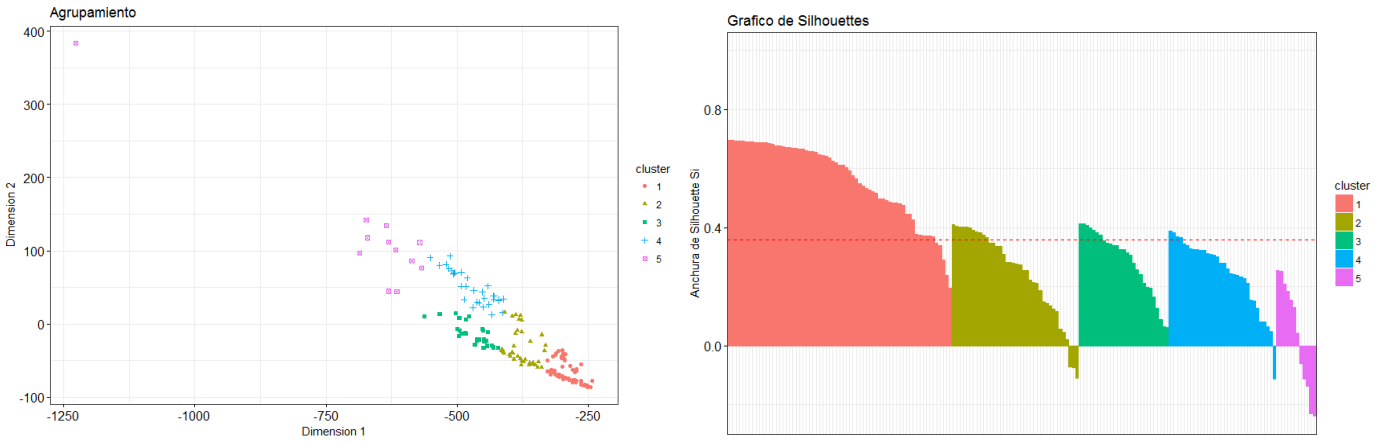


Figura No. 8.3 Agrupamiento con CLARA, tf y distancia Manhattan de la fuente Notivision.

ANEXO III

Tabla No. 8.3 Palabras más significativas de Delitos Sexuales

Periódico Virtual		**Notivision**	
tf	***TC***	***tf***	***TC***
abusar	sorprendido	sexual	acceso
sexual	promiscuo	abusar	carnal
acceso	bebe	acceso	emitida
carnal	violar	carnal	linea
violar	sexual	extorsion	plaza
carnalmente	acceso	promiscuo	responsable
promiscuo	carnal	acusado	sindicado
hurtar	acudir	emitida	ambulante
infantil	tocamientos	violencia	agresores
agravado	incautar	agresion	conmociono
agresores	incapacidad	aprovechando	ilicito
probarble	expidio	arma	robar
bebe	interpuso	capturando	violada
arma	previa	cometidos	violencia
monstruo	embarazo	planes	acusado
resistir	resistir	recaian	avanzada
aprovechaba	ausencia	carnalmente	promiscuo
agresiones	padrastro	monstruo	agravado
probartorios	infantil	placer	cometia
atentar	monstruo	violar	secuestro
homicidio	aprovechaba	acariciar	abusar
acusado	arrestaron	agresiones	detencion
sorprendido	extremas	amenazas	sexual
agresor	linchado	indebido	obscenos
atentar	agresor	infantes	salud

Tabla No. 8.4 Palabras más significativas de Homicidios

Periódico Virtual		Notivision	
tf	***TC***	***tf***	***TC***
asesinar	homicidio	asesinar	acusado
homicidio	arma	homicidio	homicidio
arma	disparo	arma	asesinar
disparo	bala	principal	agresion
bala	sicarios	comunitario	desconocidos
sicarios	atacado	ilegal	escalofriante
cadaver	fallecido	tentativa	tortura
atacado	fallecio	agresion	tentativa
cuerpos	objeto	reclaman	acusadas
desconocidos	permiso	doble	arma
reconocido	hurtar	acusado	doble
fallecio	cadaver	permite	penitenciaria
agravado	muerto	transitaban	defensor
levantamiento	cuerpos	muerto	afectuar
violencia	elegido	uniformado	agresor
hurtar	sucedidos	agresor	bala
ocurrio	independencia	bala	agredida
mineria	atacadas	esmeraldas	altercado
movilizaba	comunero	lideres	fallecio
agresor	agresion	mineria	golpeada
agresion	movilizaba	cuerpos	gritos
autor	suegra	municiones	robar
traficar	fallecieron	mortal	companero
discusion	levantamiento	defensor	golpes
cometido	agravado	entes	ilegal

Tabla No. 8.5 Palabras más significativas de Hurtos

Periódico Virtual		**Notivision**	
tf	***TC***	***tf***	***TC***
hurtar	automotores	hurtar	granada
arma	apartamenteros	arma	reportados
agravado	robar	agravado	recuperados
calificado	arma	calificado	delincuencia
robar	asalto	delincuencial	extorsion
apartamenteros	calificado	mercancia	piratas
automotores	agravado	robar	terrestres
asalto	hurtar	automotores	robar
delincuencial	delincuencial	pirateria	pirateria
fabricacion	recuperados	traficar	arma
traficar	homicidio	fabricacion	hurtaban
ilegal	huir	huir	hurtar
delinquir	cacha	telefono	agravado
homicidio	intimidan	ilicito	mercancia
huir	pistola	apartamenteross	calificado
asaltantes	asaltantes	celular	atracado
ladrones	asaltaron	huida	fabricacion
antecedentes	desmantelan	concierto	celular
atraco	antecedentes	delinquir	traficar
recuperados	fabricacion	movilizaba	delincuencial
concierto	liebres	dedicados	asesinar
municiones	atraco	inmueble	peligrosas
pretendian	delinquir	asalto	apartamenteros
mercancia	haladores	desarticulada	fleteo
atraco	ladrones	atraco	asalto

Tabla No. 8.6 Palabras más significativas de incendios Intencionales

Periódico Virtual	
tf	**TC**
incendio	incinerar
bomberos	quemar
llamas	incendio
quemar	voraz
conflagracion	arma
incinerar	conflagracion
arma	automotores
maquina	bomberos
consumio	bodega
forestal	llamas
fuego	forestal
controlar	tejares
voraz	plaza
plaza	consumida
consumido	fuego
extendieran	consumio
socorristas	aerea
hectareas	controlar
serviaseo	socorristas
almacen	carton
hurtar	combustion
consumida	descuido
controlada	calor
automotores	ola
calor	fusil

Tabla No. 8.7 Palabras más significativas de Minería Ilegal

Periódico Virtual	
tf	***TC***
mineria	mina
ilegal	derrumbe
maquina	realizando
retroexcavadora	muertos
explotacion	maquina
oro	explotacion
ilicita	oro
mina	incautar
recursos	naturales
incautar	recursos
naturales	ilicita
ambiental	retroexcavadora
yacimiento	orillas
contaminacion	ambientales
arma	atrapadas
derrumbe	autorizacion
ambientales	cuerpos
rios	extraer
aerea	fallecieron
dragas	crc
kilo	geografia
agencia	macizo
crc	yacimiento
fauna	aerea
ilicito	ambiental

Tabla No. 8.8 Palabras más significativas de Secuestros

Periódico Virtual		**Notivision**	
tf	***TC***	***tf***	***TC***
secuestro	soldado	secuestro	libero
liberar	liberar	extorsivo	frustrado
secuestradores	ilegal	agravado	operatividad
arma	arma	liberar	profugo
rescate	examenes	arma	secuestro
soldado	secuestradores	extorsion	desaparecidos
ilegal	rescate	plagio	desaparicion
examenes	bebe	profugo	desconoce
guerrilla	terroristas	escondia	desparecido
plagio	llamadas	judicializadas	extorsivo
sindicato	extorsivas	desplegaron	plagio
guerrilleros	agravado	retenido	asesinar
montanas	asesinar	amenazas	delictivos
plagiado	cautivo	homicidio	extorsion
asesinar	pago	asesinar	abandonado
agravado	abandonada	desaparecidos	plagiado
terroristas	extorsivo	desaparicion	trasladarlo
estudiantes	guerrilla	desaparecio	ilegal
interceptados	plagiados	frustrado	secuestradores
kilometro	desaparicion	operatividad	desaparicion
cautivo	plagiados	cautivo	cautivo
plagiados	plagio	amenazas	rescate
antecedentes	captores	privadas	pago
desaparicion	extorsion	plagiado	soldado
extorsion	salud	salud	liberar

Tabla No. 8.9 Palabras más significativas de Trafico de Drogas

Periódico Virtual		**Notivision**	
tf	***TC***	***tf***	***TC***
cocaina	cocaina	marihuana	cocaina
marihuana	marihuana	estupefaciente	kilo
traficar	kilo	traficar	marihuana
kilo	incautar	sustancia	kilogramo
estupefaciente	mata	incautar	incautar
incautar	estupefaciente	cocaina	base
sustancia	droga	fabricacion	estupefaciente
droga	laboratorio	kilogramo	traficar
base	dosis	kilo	cargamento
fabricacion	traficar	dosis	dosis
laboratorio	base	paquetes	bazuco
dosis	cargamento	expendio	laboratorio
alucinogeno	kilogramo	base	transportando
paquetes	sustancia	alucinogeno	droga
ilicito	paquetes	gramo	sustancia
kilogramo	galon	bazuco	gramo
galon	alucinogeno	cargamento	venta
mata	procesamiento	allanamiento	desarticula
procesamiento	semillero	droga	expendio
hoja	fabricacion	ilicita	paquetes
cargamento	hoja	transito	alucinogeno
ilegal	ilicito	ilegal	allanamiento
pasta	gramo	venta	fabricacion
bazuco	ilegal	transportando	incauta
expendio	sorprendidos	ilicito	creepy