Clasificador de crimenes

Gabriel Álvarez 09-10029 - Francisco Martínez 09-10502 - Prof. Masun Nabhan Homsi

Universidad Simón Bolívar

gabrielaar11@gmail.com - frammnm@gmail.com - mnabhan@usb.ve

November 7, 2016

Tabla de contenido I

Tabla de contenido II

Introducción

Este mini-proyecto pretende desarrollar una aplicación móvil para visualizar en forma gráfica los eventos ocurridos referentes a crimen en la ciudad de Caracas. Se explicará el trabajo realizado para extraer los datos que vienen como texto en forma de tweets utilizando el API de Twitter, el preprocesamiento realizado para la obtención de términos relevantes y representantes del tema de crimen como también los pasos, algoritmos y tareas necesarias para entrenar los algoritmos para la correcta clasificación de los tweets de manera tal que se puedan identificar tweets nuevos y en tiempo real acerca del tema. Finalmente se hablarán de los resultados encontrados, los cuales consideramos como suficientemente exitosos para seguir en los siguientes pasos a trabajar.

Descripción del Problema

El problema a trabajar posee varios subtareas que deben ser resueltas para el cumplimiento de los objetivos:

- La obtención de datos.
- El preprocesamiento del texto.
- La clasificación de los tweets.

Objetivo General

- Diseñar e implementar una aplicación móvil para visualizar los eventos ocurridos en la ciudad de Caracas

Objetivos Específicos

- Aplicar los algoritmos de procesamiento de lenguaje natural para la limpieza del texto fuente (Twitters).
- Desarrollar la base de datos para llevar el control de los usuarios de la aplicación.
- Analizar y procesar el texto para detectar patrones relacionados a eventos.
- Usar algoritmos avanzados de visualización de datos.
- Modelar el usuario para brindarle una lista de eventos cercanas a su interés.
- Hallar visualmente las relaciones entre diferentes eventos.
- Implementar la interfaz del sistema.
- Presentar el sistema a través de una interfaz web amigable para el usuario.

Obtencion de datos

La obtención de los datos fue utilizando el API de twitter y el lenguaje de programación Python, a través del API de twitter se pudo especificar las cuentas de las cuales se iban a obtener los tweets y la cantidad de tweets que se iban a utilizar, alrededor de 33000 tweets. Estas cuentas fueron las que más relacionadas estuviesen con los crímenes en el área de Caracas.

- La libreria tweepy

Preprocesamiento

El preprocesamiento consistió en eliminar los signos de puntuación de los tweets, menciones, signos de exclamación, signos de interrogación, corchetes, parentesis y links. También se transformó todo el texto a minscula. Este filtro fue realizado debido a que estos elementos no presentaban información relevante para los crímenes. Luego, se utilizó la herramienta WEKA, la cual posee una variedad de opciones para el procesamiento del lenguaje natural, así como algoritmos de clasificación. En este caso, los pasos, de forma general, a seguir fueron los siguientes:

Preprocesamiento

Se cargó el archivo de los tweets, con el formato de WEKA. Este archivo en un principio se le colocó de manera aleatoria las etiquetas de yes/no. Una vez cargado el conjunto de datos, se procedió a obtener los unigramas, bigramas y trigramas. Utilizando las opciones de IDF y TF, junto con stopwords y stemming para el caso de los unigramas. Luego de obtener los resultados, se procedió a filtrar manualmente los mismos. Dejando sólo los unigramas, bigramas y trigramas que más relacionados estuviesen con crímenes. Una vez obtenida esta información, así como unas palabras extras ofrecidas por la profesora, se procedió a etiquetar correctamente el conjunto de datos.

Stopwords

Los stopwords son las palabras más comunes dentro del lenguaje, generalmente los artículos, estas palabras fueron eliminadas para la obtención de los unigramas ya que no aportan información. En este caso, se utilizó la opción de WordsFromFile de WEKA. En esta opción se le debe proporcionar un archivo con la lista de palabras que serán utilizadas para la eliminación de stopwords, en nuestro caso se utilizó una lista de stopwords en español.

Determinacion para conceptos relevantes (NGram)

Para determinar los términos relevantes para el tema de crimen, utilizando la herramienta Weka se uso un proceso el cual dado un texto ya libre de datos irrelevantes como signos de puntuación, enlaces, numeros, llaves, corchetes, signos matemáticos, referencias a otras cuentas de Twitter, saltos de línea, y otros caracteres que no brindan información en lo que respecta a un tema (una coma por ejemplo no me dice si un tema es de crimen, o no, al igual que los otros elementos mencionados), se divide y se crean gramas de tamaños específicos. Un grama viene siendo un conjunto de tamaño específico de palabras consecutivas en el texto, para que quede de forma clara se ejemplifica a continuación :

Dado el texto : El asesinato en las Mercedes
Dependiendo del tamaño buscado se pueden generar distintas listas de gramas:
Unigramas: (El, asesinato, en, las, Mercedes)
Bigramas : (El asesinato, asesinato en, en las ,las Mercedes)
Trigramas : (El asesinato en, asesinato en las, en las Mercedes)

Determinacion para conceptos relevantes (NGram)

Al obtener la lista de los distintos gramas cada lista fue filtrada de conceptos irrelevantes, para explicar cuáles de estos términos fueron los eliminados hay que poner en contexto que las cuentas de Twitter utilizadas no hablan nicamente de crímenes, ya que también pueden hablar de noticias económicas, tecnológicas, culturales, y sobre todo un tema recurrente es de la política, términos como Mesa de la Unidad, Capriles, biotecnología, felicidad, entre otros, que terminan siendo irrelevantes para nuestros objetivos, asi que cada lista fue filtrada por separado para finalmente tener un solo conjunto de términos resultantes que provienen de la unión de estos distintos conjuntos de gramas.

Determinacion para conceptos relevantes (NGram)

Ahora es necesario explicar como un grama dado llega a estar en la listas de gramas creadas, para esto hay que entender que no todas las palabras de la lengua española se utilizan con la misma frecuencia (como en la mayoría de los idiomas), ya que palabras como la, el, del entre otras son usadas de manera sumamente frecuente, a diferencia de palabras clave como crimen, asesinato, arma, y por esto no es suficiente que la lista de de gramas sea basado nicamente en la frecuencia de aparición de las palabras en el texto, para resolver este problema se realiza una normalización de la frecuencia llamada TF-IDF.

Etiquetacion de tweets

Para etiquetar los tweets, se utilizaron los términos más relevantes, obtenidos en el punto anterior. El proceso utilizado fue el siguiente: De los unigramas, bigramas y trigramas obtenidos, se generaron tres listas, en una de ellas se guardaron los términos que mejor explicaran qué sucedió, en otra se guardaron los términos que tuviesen que ver con el tiempo y en la ltima los términos que mejor explicaran el cómo sucedió, Estas listas fueron denominadas como What, When y How, respectivamente. Luego para etiquetar los tweets se utilizaron la lista de What y How, la lista When no fue utilizada porque por sí sola no aportaba información referente a crímenes; sí los tweets poseían términos de alguna de estas dos listas, entonces eran considerados como tweets de crimen y eran etiquetados con un Yes, en caso contrario se etiquetaban con un No.

TF-IDF

En la obtención de información tfidf (term frequencyinverse document frequency), es un valor estadístico que busca reflejar la importancia de un término en un texto dado, esta normalización busca darle más peso a aquellas palabras que son usadas de manera poco comn, y castigar a aquellas que son utilizadas de manera extensiva.

Por ejemplo en la frase la casa azul quiere ser buscado en un conjunto de documentos y devolver el documento que sea más relevante en relación a esa frase, directamente se podría contar la cantidad de frecuencia de cada término, y devolver el documento que tenga el mayor valor para la suma de estas frecuencias, pero esto devolverá documentos que utilicen más intensivamente el término "la" en vez de los terminos más relevantes "casa" y "azul", esto sucede ya que la mayoría de los terminos más repetidos llegan a ser articulos y conectores en vez de palabras de conceptos más complicados.

Matemáticamente, TF-IDF es el producto de la frecuencia del término y la frecuencia inversa del término, la frecuencia inversa del término viene siendo el logaritmo de la cantidad de documentos entre la cantidad de documentos que poseen el término.
Esta técnica fue la utilizada para obtener los terminos relevantes a través de los miles de tweets que se poseen como datos en el preprocesamiento del texto.
Para apreciar mas en detalle el hecho de una mayor frecuencia de ciertos terminos en los lenguajes naturales revisar la Ley de Zipf.

KNN

De las siglas en inglés k-nearest neighbors, este algoritmo en la fase de entrenamiento consiste en colocar todos los vectores del conjunto de entrenamiento en el espacio con su respectiva clase y para clasificar un nuevo vector el proceso es el siguiente: se obtienen los k vecinos más cercanos a ese vector, y el vector es etiquetado con la clase que más aparezca en sus k vecinos. El proceso para determinar los k vecinos varía en la implementación, aunque comnmente se utiliza la distancia euclidiana para el caso de variables continuas y la distancia de Hamming para las variables discretas. El algoritmo utilizado fue el de IBK de la herramienta de WEKA, el mejor resultado obtenido fue de 73

Arbol de decision(j48)

En el algoritmo C4.5 se genera un árbol utilizando la entropía de la información, en cada nodo del árbol de decisión el algoritmo utiliza el atributo que mejor divide el conjunto de entrenamiento en subconjuntos de clases diferentes. Finalmente para clasificar, sólo se debe seguir el árbol de decisión generado.

Convolution NN(Deep Learning)

Este algoritmo funciona similar a una red neuronal donde alimenta datos a través de las capas y tiene como objetivo reducir el error del resultado calculado en la red a través de cada iteración, a diferencia de sus contrapartes las redes neuronales convolucionales tienen un arreglo distinto de las neuronas, donde las capas de neuronas se llaman campos receptivos las cuales poseen trabajos diferentes:

- Neuronas convolucionales : que se encarga de extraer las caracteristicas.
- Neuronas de reduccion de muestreo : se ocupa de la reducción de muestreo que sirve para generalizar características de manera tal que haya una tolerancia a ruido en los datos
- Neuronas de clasificacion : se encarga de poner la muestra en su etiqueta correcta luego de la depuración de la información.

Aunque este algoritmo tiene aplicaciones más que todo gráficas, se ha demostrado que tiene buenos resultados para el procesamiento de lenguaje natural, para el análisis de juego (Go), descubrimiento de drogas entre otras aplicaciones que se alejan del campo de imagenes.

Matriz de confusión

- KNN (IBK con Unigramas,Bigramas y Trigramas) :

No	**Yes**
1906.25	98.83
988.93	1042.84

- C4.5 (J48 con Unigramas, Bigramas y Trigramas) :

No	**Yes**
1959.57	45.51
133.09	1898.68

- CNN (con Unigramas, Bigramas y Trigramas) :

No	**Yes**
1960.28	44.79
161.73	1870.04

Sensibilidad

- KNN (IBK con Unigramas, Bigramas y Trigramas): 0.731
- C4.5 (J48 con Unigramas, Bigramas y Trigramas): 0.956
- CNN (con Unigramas, Bigramas y Trigramas): 0.949

Especificidad

- KNN (IBK con Unigramas, Bigramas y Trigramas): 0.787
- C4.5 (J48 con Unigramas, Bigramas y Trigramas): 0.957
- CNN (con Unigramas, Bigramas y Trigramas): 0.950

F-Score

- KNN (IBK con Unigramas,Bigramas y Trigramas): 0.717
- C4.5 (J48 con Unigramas, Bigramas y Trigramas): 0.956
- CNN (con Unigramas, Bigramas y Trigramas):0.949

Resultados del preprocesamiento

Consideramos los resultados del preprocesamiento como exitosos, pues nos permitieron obtener una lista de términos y conceptos relevantes para la clasificación de los tweets que no teníamos y que además pasarán por un proceso de subclasificación en pasos futuros, aunque debemos admitir que parte del proceso sigue siendo manual, y esto se debe a la gran cantidad de ruido en los datos, ya que las mismas fuentes de información de crímenes en Twitter no hablan nicamente de nuestro tema particular, si no son mezcladas con otros temas de periodismo y noticias, lo cual es comn. También pensamos que estas listas pueden y deben ser ampliadas, pues an se escapan términos que no se utilizan de forma tan comn en los tweets que pueden indicar un crimen y sus subclasificaciones, pero indiferentemente consideramos que lo que tenemos es un buen comienzo.

What	When	Where	How
emergencia	ayer	barlovento	arma
enfrentamiento	domingo	bolivar	armados
fallecidos	enero	caracas	armas
falleció	jueves	catia	cabeza
golpe	julio	chacao	dispararon
homicidio	lunes	lara	disparos
homicidios	martes	maracaibo	golpe
investigación	manana	margarita	granada
lucha	noche	morgue	lucha
matan	septiembre	petare	policial
mataron	sabado	sucre	tiro

Resultados de la clasificación

De la lista de resultados anteriores 95.57% de las instancias clasificadas correctamente fue el mejor resultado, proveniente de C4.5(J48 con unigramas,bigramas y trigramas) siendo además este algoritmo ganador de manera casi consistente en la mayoría de las distintas evaluaciones, se puede observar comparando cada matriz de confusión como C4.5 posee el menor nmero de evaluaciones incorrectas siendo cercana la red neuronal, pero esta ltima pierde siendo menos exacta al clasificar las instancias negativas produciendo menos clasificaciones positivas que se traduce en perdida de informacion para la futura aplicacion, además observando la sensibilidad(95.6%), especificidad(95.7%) y F-Score(95.6%) son mejores al ser comparadas con los índices de los otros algoritmos. Este resultado siendo superior al 95% se considera lo suficientemente consistente para ser utilizado para la aplicación, ya que nos permite obtener nuevos tweets en tiempo real, evaluarlos y aumentar la cantidad de datos que puede mostrar la aplicación.

Conclusiones

Concluimos que los resultados obtenidos son lo suficientemente buenos para ser utilizados en una aplicación de uso masivo, siendo estos consistentes desde el punto de vista de aprendizaje de máquina, y que estamos preparados para los siguientes pasos los cuales se discutirán en el siguiente punto. Es relevante recalcar que el proceso de aprendizaje de máquina an no se ha terminado, pues la subclasificación de los crímenes por tipo, tipo de armas utilizadas, lugar del evento al igual como el momento del evento son objetivos futuros que se desean ampliar y trabajar para ser reflejados en una experiencia para el usuario más rica y enriquecedora. También es bueno mencionar que nos sentimos satisfechos de lograr observar los frutos de esta subdisciplina de la ciencia de la computación, ya que trabajar con ideas, algoritmos, herramientas y personas en el campo nos hace apreciar de las capacidades del área, ayudándonos a comprender cómo podemos apoyar a la sociedad a tomar decisiones más inteligentes, estar mas protegidos, y utilizar la información de forma provechosa.

Trabajos Futuros

El trabajo a futuro se puede dividir en cuatro pasos relevantes :

- Subclasificación de los tweets: En esta tarea, tenemos como objetivo la subclasificación de los tweets de crimen en distintas categorías, sean estas robos, asesinatos, peleas, enfrentamientos, u otros, pero a su vez tenemos como objetivo sub clasificarlos an mas con tipos de armas utilizadas, el lugar del evento, y si es posible el momento en que sucedieron.
- Modelado de los usuarios: Aquí buscamos describir los datos del usuario, sus tareas y posibles acciones dentro de la aplicación, dependiendo del tiempo que se disponga esta tarea será mas o menos desarrollada.

- Diseno de la aplicación: Se plantea realizar una aplicación android ya que es la que posee mayor rango de disponibilidad, además de que los controles para publicar la aplicación son mucho más estrictos para los productos Apple, de ser necesario en un futuro se podría cubrir esa idea también pero posiblemente escapa al alcance de este proyecto.
- Visualización de los datos: Es uno de nuestros objetivos principales que esta informacion textual se vea reflejada gráficamente de una forma agradable e intuitiva, además de didáctica para los usuarios, facilitandoles la toma de decisiones, se plantea utilizar mapas, pero no se descartan la visualización con otro tipos de gráficos que logre resaltar patrones o relaciones entre los lugares, fechas, tipos de eventos y otros datos mostrados.

Final de la presentación