ISSN 1870-4069

Análisis del comportamiento de diferentes algoritmos de aprendizaje automático para catalogar delitos en la zona metropolitana

Belém Priego Sánchez, Stephany Anaya García, José A. Reyes-Ortiz

Universidad Autónoma Metropolitana unidad Azcapotzalco,
Departamento de Sistemas,
México

{abps,jaro}@azc.uam.mx, stephany.anaya@hotmail.com

Resumen. Actualmente nos encontramos en la era de la información en donde se produce, se difunde y se emplea gran cantidad de ésta diariamente, la cual puede ser estudiada y analizada con el fin de obtener nueva información. Aunado a este hecho se encuentran las redes sociales en las cuales se puede expresar una idea, concepto, opinión o evento. El análisis de las ideas, comentarios de Twitter, por ejemplo, escritas por la comunidad a despertado el interés de muchos investigadores. En este artículo, se está interesado en catalogar información, de delitos ocurridos en la zona metropolitana, reportada en Twitter a partir de tres algoritmos de clasificación. Teniendo como objetivo presentar los resultados obtenidos tras su ejecución; estos resultados evidencian que si se tienen los parámetros correctos se puede clasificar y obtener nueva información a partir de una ya existente.

Palabras clave: Aprendizaje automático, algoritmos de clasificación, textos cortos, delitos.

Analysis about the Behavior of a Different Machine Learning Algorithms in order to Catalog Crimes in the Metropolitan Area

Abstract. Nowadays we live in the information era where a massive information is produced, disseminated and used daily, which can be studied and analyzed in order to obtain new information. In addition, we have the social networks, in which you can express an idea, concept, opinion or event. Twitter opinions analysis - twitter comments, for example- written by the community has awake the interest of many researchers. Therefore, this paper is focused on classify the information about crimes that occurred in the metropolitan area of Mexico City, reported on Twitter from three types of classification algorithms. Having in mind to present the results obtained after their execution, we observed that when the correct parameters are set, then it is possible to classify and to obtain new information from a preexistent one.

Keywords: Machine learning, classification algorithms, short texts, crimes.

1. Introducción

Indudablemente, nos encontramos en la era de la información, en donde se produce, difunde y emplea información en abundantes cantidades diariamente. El estímulo que originó esta era se dio con el auge de las tecnologías de la información y la comunicación. Un campo encargado del estudio de dicha información es el Procesamiento del Lenguaje Natural, denotado de aquí en adelante por PLN, gracias a que proporciona un análisis de los textos convenientes para detectar, recuperar y computar información subjetiva de forma metódica. Dentro del PLN se encuentra el área de la Minería de Textos la cual, actualmente, ha despuntado debido a la creciente demanda en el uso de las redes sociales. Oportunamente, las empresas y otras instituciones se han beneficiado de ello, estudiando y examinando las colosales cantidades de datos que se producen y difunden por los usuarios a través de Internet, esto diferenciando la composición del sentimiento en un texto.

La detección de las opiniones expresadas por los usuarios de las redes sociales, brinda la facultad de responder interrogantes en un gran número de dominios de aplicación. Así mismo, dentro de las redes sociales emergen los denominados *influencer*, los cuales son individuos con características y aptitudes sociales favorables, que contribuyen en la toma de decisiones de un gran número de personas, guiados por las opiniones de estos mismos. Cuando se origina y difunde una opinión pública, en la que la cual puede estar en actuación diferentes circunstancias, desde el comportamiento en la bolsa de valores de una empresa o bien la toma de decisiones en la política de un país, se analiza el comportamiento de estas opiniones dentro de las redes sociales que brinda la capacidad de informar acerca de sucesos o acontecimientos venideros. Sin embargo, el ordenamiento y estructuración de los textos capturados, resultaría irrealizable para el ser humano, de tal suerte que se cuenta con algoritmos, el propósito de disponer de estos algoritmos es conjeturar la condición objetivo, mediante el análisis del conjunto de datos capturados. Es decir, etiquetar y organizar con una valoración la información difundida en el texto, el cuál es la labor que se llevará a cabo en este trabajo de investigación.

Los algoritmos con los cuales se disponen, principalmente, son los de clasificación y agrupamiento (*agrupamiento*). Los algoritmos de clasificación persiguen predecir la clase objetivo por medio del análisis del conjunto de datos de entrenamiento. Por otro lado, los algoritmos de agrupamiento aspiran a reunir los elementos haciendo uso de diferentes medidas de afinidad o similitud. Es por ello que el aprendizaje automático se apoya de estos algoritmos, que se destinan para evolucionar los métodos de procesamiento y posibilitan a las computadoras a aprender a predecir acontecimientos en los cuales se involucran distintos métodos matemáticos. Al situar estos métodos a la práctica se instruyen de los modelos de entrenamiento para realizar su predicción.

En esta investigación se realiza una comparativa de tres algoritmos de aprendizaje automático, nos apoyamos en los algoritmos de clasificación y agrupamiento. Se selecciona, analiza y compara determinados modelos algorítmicos, con la finalidad de identificar la precisión y exhaustividad de estos algoritmos dentro de un mismo caso de estudio. El caso de estudio en el que se implementarán los algoritmos, será enfocado en clasificar seis tipos de delitos en el área metropolitana de México que de acuerdo con el Instituto Nacional de Estadística y Geografía [1] son los incidentes que más ocurren en México. Estos son reportados por diversas fuentes de información oficiales vía Twitter. Los *tweets* que contienen la información de los delitos fueron etiquetados manualmente, previamente, de acuerdo a la categoría a la cual pertenecen. Las categorías de etiquetado son: Homicidio, Suicidio, Asalto, Violación, Explotación y Secuestro. Con la elaboración de esta investigación se señala y demuestra el algoritmo con mayor precisión y exhaustividad dentro del análisis de textos orientado a los tweets que contengan esta información.

El resto del artículo se presenta como sigue: en la Sección 2 se da una breve motivación del porque realizar este trabajo, en la Sección 3 se describen algunos de los trabajos reportadosen el estado del arte, en la Sección 4 se presenta la metodología propuesta que da solución a la temática del artículo y en la Sección 5 se presentan los resultados obtenidos tras ejecutar dicha metodología. Fianlmene, se presentan las conclusiones obtenidas.

2. Motivación

En este trabajo de investigación se está interesado, principalmente, en analizar el comportamiento de los algoritmos de clasificación y agrupamiento, orientados a la clasificación automática de los seis delitos (homicidio, suicidio, asalto, violación, explotación y secuestro) en el área metropolitana de México. Este análisis nos permitirá determinar el mejor algoritmo en cuanto a precisión y exhaustividad.

El objetivo de realizar un estudio de diferentes algoritmos de clasificación es debido a que en la actualidad el análisis de los textos difundidos a través de las redes sociales, contribuye en la predicción de sucesos o bien permite modelar situaciones. Giovanni Cherubini, Jens Jelitto y Vinodh Venkatesan su artículo prospectivo *Cognitive Storage for Big Data* [2] proponen la noción de un sistema de "almacenamiento de datos cognitivos", en el cual se habitúa a criterios como la estimación y obsolescencia de los datos. Con la meta de desarrollar un algoritmo de aprendizaje que contribuya a la clasificación de los datos, en varias clases de relevancia y trabajar en conjunto con una arquitectura de almacenamiento multinivel para personalizar la ubicación de los datos.

El almacenamiento de los datos de forma cognitiva considera y reúne los datos, los cuales contribuyen al procedimiento de análisis de estos mismos. De esta manera, esta investigación contribuirá a sealar el algoritmo optimo a efectuar en la situación donde se cuente con datos afines a los del caso de estudio manifestado. Así mismo, se está interesado en contribuir en el análisis predictivo

de los seis principales delitos del área metropolitana de México. Lo cual, además, permitirá modelar la situación delictiva en esta región del país.

3. Estado del arte

En esta sección, se describen los trabajos reportados en la literatura relacionados con diferentes algoritmos de aprendizaje automático para el análisis de opiniones. Si bien es cierto que existen bastantes investigaciones que se dedican a la comparativa de diferentes algoritmos de clasificación, en este apartado hemos seleccionado algunas de las muchas investigaciones. Esto debido a que el objetivo es mostrar que esta temática es un punto de referencia clave a la hora de utilizar diferentes algoritmos y que es una investigación atrayente a los investigadores dedicados al análisis de opiniones.

En [8] se proporciona una pesquisa sobre los desafíos y la visión general de algunos algoritmos de clasificación y agrupación utilizados para el análisis sentimental y de la minería de opiniones. La similitud, entre la propuesta y dicho artículo, radica en la intensión de los algoritmos, sin embargo, difiere en la implementación y las herramientas de software a utilizar. En [7] se aborda el tema de minería de opiniones y análisis de sentimientos, la cual es una tarea de procesamiento de lenguaje natural e información que identifica las opiniones de los usuarios explicadas en forma de comentarios positivos, negativos o neutrales y citas subyacentes al texto. La similitud radica al utilizar distintos algoritmos supervisados o basados en datos. No obstante, en este caso se realiza para el análisis de sentimientos y también considera la precisión de la clasificación del sentimiento. En [6] se describe un trabajo de estudio sobre la eficiencia del lenguaje R en la minería de opiniones, la similitud de este artículo con la propuesta presentada radica en el uso del lenguaje R así, mismo, como el uso de un corpus extraído de Twitter. No obstante, difiere en la aplicación de algoritmos de aprendizaje automático.

En el trabajo [3] se aplicaron técnicas de PLN e implementaron cuatro diferentes procesos de Minería de Textos. La principal afinidad que se tiene con esta investigación y con el que se propone en esta propuesta, es el uso de diferentes algoritmos como máquinas de soporte vectorial (SVM, por sus siglas en inglés), Naïve Bayes, J48 y K-vecinos más cercanos para realizar la minería de texto y se compararán los resultados generados. Sin embargo, la diferencia radica en el uso del software Weka para implementar dichos algoritmos, mientras que en el caso de esta propuesta se implementaron en el software R. En la propuesta [4] se aplicaron tres algoritmos agrupamiento a los genes activos e inactivos, asociados con tumores pediátricos de 60 pacientes para poder determinar con ello que tipo de tumores se presentaba en cada paciente. La similitud con el presente trabajo se tiene en la aplicación de algoritmos de agrupamiento y el uso del software R, este trabajo es muy similar al que presentamos pero los objetivos planteados, resultados y los datos empleados son algunas de las características diferentes que se tienen. Además del análisis y resultados completamente distintos. En [5] se diseñó un sistema para la clasificación de artículos científicos en idioma

inglés, utilizando el formato PDF y se presentaron los resultados por medio de la tecnología Web Java Servlets. La similitud con este trabajp es la clasificación de un conjunto de datos, utilizando el algoritmo "K-Means", mientras que la diferencia radica en los datos a emplear. De igual forma, existe una diferencia con las herramientas a utilizar.

En la siguiente sección, se describe la metodología propuesta que dará solución al análisis de los diferentes algoritmos que se utilizarán para la tarea de clasificación automática.

4. Metodología propuesta

El objetivo de este trabajo, de investigación, es el de clasificar seis tipos de delitos. Para lograr dicha meta, se propone una metodología que está conformada de cuatro etapas principales.

1. *Agrupamiento de los tweets.* El objetivo, de esta etapa, es conformar la colección de textos, éstos formarán el corpus. Partiremos de un aproximado de 2,500 tweets que han sido, previamente, etiquetados manualmente de acuerdo a la etiqueta tipo de delito. Ésta corresponde al delito que el tweet tiene.
2. *Preprocesamiento de los tweets.* Se cubrirá, principalmente, el análisis gramatical de la oración para leer el texto, es decir, se analizará el cómo están formados los tweets. De la misma manera, se analizan dichos textos por estructuras.
3. *Minería de textos.* Se extrae información utilizando diferentes herramientas, este proceso encuentra las similitudes entre los datos que tiene el mismo significado y así poder obtener información sobre ellos. Dentro de esta etapa, se tienen dos subetapas. Con respecto a la primera etapa, "Aplicación del algoritmo", se tienen diferentes procesos. En la Figura 1, se muestran de forma gráfica estos procesos, los cuales están definidos de forma lineal.
   Las subetapas que se llevan a cabo son:
   a) Aplicación del algoritmo. Los algoritmos que serán implementados son: *K-Means, Naïve Bayes* y *k Nearest Neighbor*. Los procesos, o pasos, por los cuales se deberá someter a cada algoritmo son:
      1) Selección del algoritmo.
      2) Construcción del modelo.
      3) Implementación del modelo.
      4) Generación del modelo y coordinación.
      5) Análisis.
   b) Modelado de la información. En este proceso, se muestra de forma gráfica la información que los algoritmos suministren, de tal forma que el análisis de esta información sea más comprensible para la comparación de los algoritmos.

En la siguiente sección, se presentan los resultados obtenidos tras llevar a cabo la metodología propuesta. Los experimentos se ejecutaron mediante el entorno de R[9].

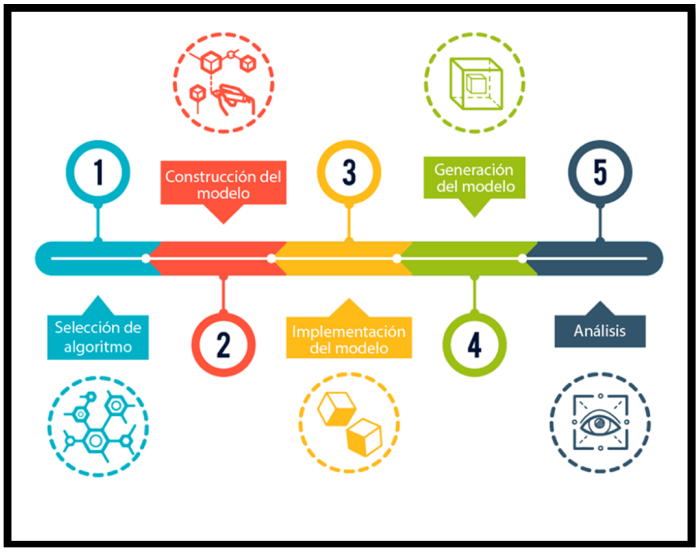

Fig. 1. Diagrama del proceso de análisis por cada algoritmo.

5. Resultados experimentales

Con el fin de mostrar los resultados obtenidos a lo largo del desarrollo de esta investigación, se ha divido esta sección en tres subsecciones principales: a)Conjunto de datos, b)Algoritmos empleados y c)Resultados obtenidos. A continuación, se describe cada una de éstas.

5.1. Conjunto de datos

En esta subsección se describe el conjunto de datos, utilizado y construido, denominado "Eventos Crimen México". Este conjunto de datos corresponde a encabezados (*headlines*) de noticias y tweets en Español sobre seguridad. Tanto los encabezados como los tweets provienen de periódicos electrónicos mexicanos como: "El universal, Excelsior, La Jornada, Noticias MVS, Reforma" y sus respectivas cuentas en Twitter.

En este corpus de noticias se tienen 1,500 encabezados de periódicos mexicanos y 1,500 tweets, los cuales fueron recolectados entre el 17 de noviembre del 2017 y el 18 de enero de 2018. Cada cabecera de noticia está compuesta de 35 palabras en promedio y cada mensaje de Twitter de máximo 140 caracteres que es lo que permite la red social.

El conjunto de datos está dividido en seis categorías: asalto, trata de personas, homicidio, suicidio, secuestro y violación. Para obtener las categorías se llevo a cabo un proceso de anotación, el cual consistió en asignar una etiqueta al encabezado o tweet de acuerdo a la categoría del evento. Este etiquetado fue realizado por humanos, quienes leen el encabezado de la noticia o tweet y

determinan la categoría. Una misma cabecera de una noticia o un tweet es asignado(a) a dos etiquetadores; si estos anotadores no se ponen de acuerdo entonces se realiza un proceso de decisión que consiste en que dicha cabecera se le asigna a un nuevo etiquetador y éste es quien decide la categoría final.Una vez finalizado este proceso, los 1,500 encabezados y 1,500 tweets han quedado distribuidos en seis categorías, mencionadas anteriormente, de acuerdo a la Tabla 1.

Tabla 1. Distribución de encabezados de noticias y tweets del conjunto de datos utilizado.

Categoría	Número de encabezados	Número de tweets	Total de textos
Asalto	346	386	732
Secuestro	210	218	428
Violación	185	164	349
Trata de personas	195	180	375
Homicidio	365	366	731
Suicidio	199	186	385

A manera de ejemplo, en la Figura 2 se muestra la estructura de un tweet para la categoría *homicidio*. En el caso de las demás categorías, se sigue el mismo diseño y patrón de formato xml.

```
<?xml version="1.0" encoding="UTF-8"?>
<tweets>
<tweet id="935272038603870213"> La fiscal Yendi Torres Castellanos <evento tipo="Homicidio">fue asesinada</evento>
                                en <espacio>Pánuco #Veracruz.</espacio> https://t.co/0si97IsW5b </tweet>
</tweets>
```

Fig. 2. Estructura de un encabezado y de un tweet.

5.2. Algoritmos empleados

Las técnicas de aprendizaje automático supervisado son capaces de aprender el proceso humano para clasificar delitos con base en las características alimentadas, al clasificador, y los parámetros que se les pueden asignar. Con el fin de tener una perspectiva del tipo de clasificador que puede tratar mejor el problema de clasificación de delitos, se han seleccionado los siguientes tres algoritmos de aprendizaje:

1. *K-Means*: es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.
2. *Naïve Bayes*: es un clasificador probabilístico basado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales.

3. *K-Nearest Neighbor*: el método de los k vecinos más cercanos es un método de clasificación supervisada que sirve para estimar la función de densidad $F(x/C_j)$ de las predictoras x por cada clase C_j.

Los resultados que se obtuvieron al tratar de clasificar delitos en la Zona Metropolitana son presentados en la siguiente subsección.

5.3. Resultados obtenidos

Para llevar a cabo la ejecución de los algoritmos seleccionados, K-Means, Naïve Bayes y K-Nearest Neighbor, se han empleado diferentes parámetros. En esta sección se presentan los resultados óptimos, es decir, los mejores resultados que se obtuvieron al variar los parámetros.

En el caso de la clasificación con el algortimo de K-Means, en la Figura 3 se presenta un enfoque con un vali de $k = 3$ que corresponde al número de agrupamientos en los cuales se van a agrupar los tweets; éstos son el resultado de la agrupación de acuerdo a la parametrización de los tweets. Se presenta esta figura, debido a que con tres agrupamiento se observó que al parametrizar dichos tweets, los elementos a clasificar (tweets) son mejor agrupados a diferencia de la utilización de más o menos agrupamientos.

Además, en la Figura 4 se presenta de forma gráfica el dendograma que corresponde al agrupamiento de la calsificación de los tweets. Como puede observarse en la misma figura, Figura 4, las palabras más frecuentes que se encuentran en el corpus de tweets han sido agrupadas de acuerdo a la frecuencia de dichas palabras. Teniendo como resultado que las palabras *asesinato* y *robo* son de las palabaras con más repeticiones en el corpus y se realiza el agrupamiento de forma individual. Realizando una comparativa con el conjunto de palabras restantes, *sexual, año, hombre, cdmx, policía, dos, mujer* y *detienen*, se agrupan de manera que conforman un sólo agrupamiento.

Con respecto, al algoritmo de clasificación Naïve Bayes el conjunto de datos utilizado correspondió al 70 % para el entrenamiento y 30 % para pruebas, tomando una frecuencia de términos igual a treinta, lo cual significa que el algoritmo toma los términos con treinta repeticiones. En la Tabla 2 se muestra la matriz de confusión, tras ejecutar dicho algoritmo, la cual representa la manera en la que fueron agrupados los tweets y la predicción que tuvo el algortimo. La exactitud, accuracy, del algoritmo correspondió a un 87.35 %, el cual corresponde a un resultado de agrupamiento aceptable. Estos resultados se muestran de manera gráfica en la Figura 5, dicha figura categoriza los tweets de acuerdo a los delitos que se tienen.

Los resultados obtenidos con el algoritmo, de clasificación sueprvisada, K-Nearest Neighbor se han obtenido mediante los vecinos considerados es igual a tres. Debido a que es el modelo que más eficiencia tiene al clasificar los comentarios, de Twitter, dentro de la categoría a la cual pertenecen. En la Tabla 3, muestra la matriz de confusión, se permite visualizar el grado de efectividad de dicho algoritmo. Además, es importante mencionar que la exactitud fue del 84.36 %. Si realizamos una comparación con el mejor resultado obtenido, 87.35 %

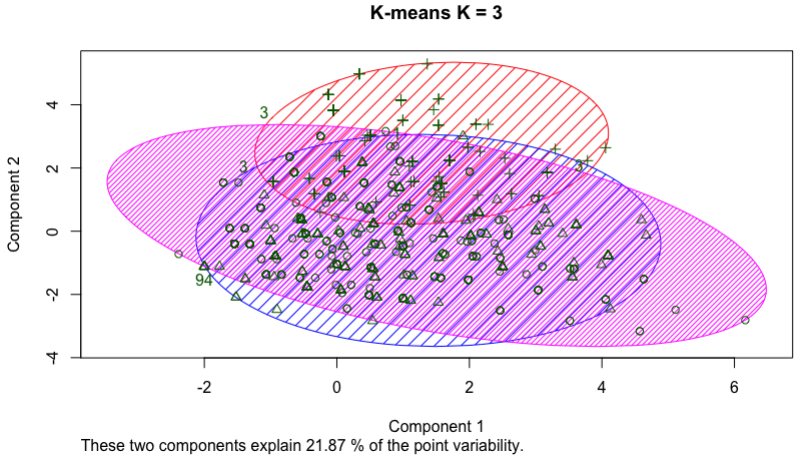

Fig. 3. Clasificación mediante el algortimo K-Means en tres agrupamientos de los delitos.

Tabla 2. Matriz de consfusión al ejecutar el algoritmo Naïve Bayes.

Predicción	Asalto	Explotación	Homicidio	Secuestro	Suicidio	Violación
Asalto	344	0	19	1	4	1
Explotación	1	0	0	0	0	0
Homicidio	23	5	388	12	14	25
Secuestro	0	0	0	7	0	0
Suicidio	1	0	7	1	12	3
Violación	1	1	0	0	0	71

resultado con Naïve Bayes, se puede observar que no existe una alta dispersión entre el resultado obtenido con K-Nearest Neighbor, sin embargo, existe.

5.4. Discusión

Cada uno de los algoritmos analizados en este artículo contiene factores a favor y factores en contra que influyen en el momento de emitir un dictamen positivo o negativo acerca de su exactitud, rapidez y facilidad de ejecución, dado que estos factores son los que concierne debatir.

El algoritmo K-means resulto ser un algoritmo, sencillo al implementar, con requerimientos mínimos de hardware, originando que el algoritmo fuese de los más veloces en tiempo de ejecución. No obstante, su desventaja radica en la exactitud que proporciona al clasificar los elementos del conjunto de datos; ya que el constante ajuste que se realiza a los centroides proporciona márgenes de error considerables. En contra parte, el algoritmo K-nearest Neighbors resultó ser el algoritmo con un tiempo de ejecución muy amplio, alrededor de quince minutos, por cada ejecucin del modelo propuesto. De igual manera, se empleó

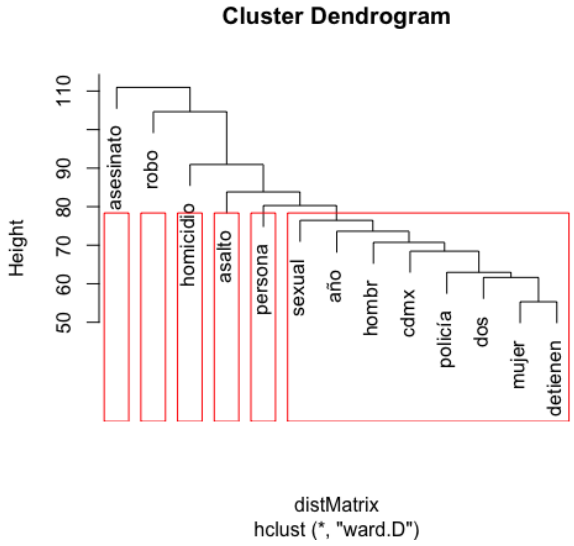

Fig. 4. Dendograma de agrupamiento mediante el algortimo K-Means en una clasificación de seis elementos.

Tabla 3. Matriz de consfusión al ejecutar el algoritmo K-Nearest Neighbor.

Predicción	Asalto	Explotación	Homicidio	Secuestro	Suicidio	Violación
Asalto	259	0	7	0	2	2
Explotación	0	3	0	0	0	0
Homicidio	77	3	451	11	23	17
Secuestro	0	0	1	13	0	0
Suicidio	0	0	2	0	10	0
Violación	1	0	0	0	1	57

la denominada distancia euclidiana, para calcular la distancia de vecindad entre un elemento y el conjunto de elementos ya agrupados o bien conjunto de datos de entrenamiento. En el lado intermedio, de ello, se encuentra el algoritmo Naïve Bayes, el cual no resulta ser el más rápido en tiempos de ejecución, sin embargo, tampoco resulta ser el algoritmo que más demora para obtener los resultados de la clasificación.

En términos de exactitud, el algoritmo que presenta mayor exactitud (*accuraccy*) es el algoritmo de Naïve Bayes, con una exactitud de clasificación de los Tweets muy aproximada a los valores reales expresados. En comparación con los dos algoritmos presentados, K-nearest Neighbors y K-means. En el algoritmo Naïve Bayes se consideran espacios probabilísticos ligados a dos eventos, en los cuales se calcula la probabilidad condicional de que ocurra el evento A dado el evento B.

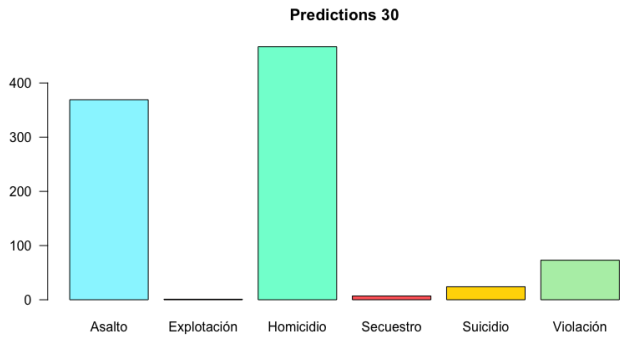

Fig. 5. Estructura de un encabezado y de un tweet.

A base de los elementos y hechos mencionados se dictamina que el algoritmo que presenta mejores resultados es el algoritmo de Naïve Bayes. Esto porque se demostró que presenta una adaptabilidad al modelo de datos presentados; resultando que el aprendizaje del algoritmo fuese rápido y certero al clasificar los Tweets. Los dos algoritmos restantes se adaptaron de forma correcta al modelo de datos, sin embargo, presentaron deficiencias en funciones esenciales como el tiempo de ejecución y de la clasificación correcta de los elementos.

6. Conclusiones y perspectivas

En este artículo se presentaron los resultados obtenidos tras ejecutar tres algoritmos de clasificación, K-Means, Naïve Bayes, K-Nearest Neighbor, para el agrupamiento de delitos sucedidos en la Zona Metropolitana y reportados en Twitter. Los resultados evidencian que cuando se configuran, o encuentran, los parámetros adecuadamente de los algoritmos empleados, éstos nos permiten obtener metainformación relevante de un conjunto de datos. Para nuestro caso, el conjunto de datos utilizado son los comentarios reportados en Twitter de delitos y de ellos se obtuvo, a manera de ejemplo, la semejanza existente entre las palabras utilizadas en uno u otro delito. Debido a la disperción existente entre las palabras de los tweets se observó más sencillez de implementación en los algoritmos de K-Means y Naïve Bayes con respecto a la K-Nearest Neighbor, esto debido al funcionamiento diferente de los algoritmos. El mejor resultado obtenido fue de 87.35 % para el caso del algoritmo,de clasificación, Naïve Bayes; de lo cual podemos concluir que dicho algoritmo es el que mejor responde para la tarea que tenemos como objetivo en este artículo, la clasificación de delitos.

Es importante destacar la efectividad de cada algoritmo, al realizar el aprendizaje automático, debido a que cada uno de ellos cuenta con parámetros diferentes para tomar en cuenta al momento de realizar la clasificación, estos parámetros

contribuyen a obtener diversos resultados. El parámetro que establece el algoritmo Naïve Bayes para calcular la probabilidad de un evento, corresponde a la proporción de ocurrencia del evento fraccionado por el número total de casos admisibles. De forma distinta el algoritmo k-nearest neighbors reúne los casos disponibles con la finalidad de clasificar los casos nuevos identificando afinidades entre los casos disponibles y los nuevos casos, proyectando puntos de los datos no clasificados en los conjuntos definidos. Por otro lado, K-means agrupa los puntos de datos en clases o clústers homogéneos. No obstante poniendo en comparación los algoritmos Naïve Bayes y K-Nearest Neighbor se puede diferenciar y destacar la efectividad en torno a la clasificación de los Tweets del algoritmo Naïve Bayes frente a K-Nearest Neighbor. Debido a que ambos algoritmos se basan en el aprendizaje supervisado. Por otro lado el algoritmo k-Means tiene un buen comportamiento frente a los dos algoritmos mencionados, sin embargo, el algoritmo que tiene una mejor presencia y exactitud al clasificar los textos es Naïve Bayes.

Como persectivas, se tiene que es posible probar otros algoritmos de clasificación y con un conjunto de datos diferente pero que corresponda al mismo dominio, los delitos. Ademaás, hoy en día nos encontramos en una turbulencia financiera, por ejemplo, esto debido a las situaciones políticas y económicas mundiales. Por lo cual se considera de suma importancia la capacidad de poder predecir acontecimientos financieros en torno a los mercados y bolsa de valores, los algoritmos de aprendizaje automático son una herramienta sustancial que nos permitirán poder realizar una predicción aceptable en este rubro y poder con ello contribuir de manera activa en el avance financiero del país. Entonces, los algoritmos empleados en este artículo podrían ser explotados en este campo dado que hemos observado que su método de predicción es fiable y aceptable.

Referencias

1. CED: Clasificación Estadística de Delitos. www3.inegi.org.mx, 2017. [Online]. Disponible: http://www3.inegi.org.mx/sistemas/clasificaciones/delitos.aspx. [Accedido: 01- Nov- 2017].
2. Cherubini, G., Jelitto, J., Venkatesan, V.: Cognitive Storage for Big Data, 49(4), pp. 40–51, (2016)
3. Paniagua, J.: Algoritmos de aprendizaje automático para el análisis de opiniones a partir de textos en español. Proyecto terminal, División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana Azcapotzalco (2017)
4. Zinzun, J.: Comparativa de la clasificación de tumores obtenida por medio de los algoritmos k-Means, PAM, AGNES. Proyecto terminal, División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana Azcapotzalco (2016)
5. López, J.: Sistema para la clasificación de artículos científicos mediante el algoritmo KMeans utilizando características semánticas. Proyecto terminal, División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana Azcapotzalco (2015)
6. Khanna, P.: Sentiment Analysis: An Approach to Opinion Mining from Twitter Data Using R. International Journal of Advanced Research in Computer Science, 8(8), pp. 1–5 (2017)

7. Khairnar, J., Kinikar, M.: Machine Learning Algorithms for Opinion Mining and Sentiment Classification. (IJSRP), 3(6) pp. 1–6 (2013)
8. Sneka, G.: Algorithms for Opinion Mining and Sentiment Analysis: An Overview. International Journal of Advanced Research in Computer Science and Software Engineering, 6(2), pp.1–5 (2016)
9. R: Past and Future History: Disponible: https://cran.rproject.org/doc/html/interface98-paper/paper.html (2017)