ISSN 1870-4069

Detección automática de zonas de alto riesgo de eventos delictivos a través de noticias periodísticas

Yadira Laureano de Jésus, Guillermo De Ita Luna,
Mireya Tovar Vidal

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

star95ariday@hotmail.com,{deita,mtovar}@cs.buap.mx

Resumen. Actualmente en las noticias, los eventos criminales están más presentes y cada año aumentan significativamente. Debido a esto, se genera inseguridad y miedo en las personas que día a día se informan mediante este medio de información. Por ello es importante estar al tanto de estos eventos, principalmente para las autoridades para que tomen medidas de seguridad y apliquen las políticas de prevención de estos eventos criminales. En este trabajo de investigación se desarrolla un algoritmo basado en aprendizaje profundo, para la clasificación automática sobre noticias de eventos criminales (homicidio, secuestro, asalto, suicidio y violación) y para la detección automática de zonas de alto riesgo. Para la clasificación de noticias, el entrenamiento se realiza con un corpus de encabezados de noticias de Twitter y para las pruebas se realizan con un corpus formado por noticias obtenidas de un periódico local (Milenio). Finalmente, en el proceso de clasificación se obtuvo una exactitud del 98 % y también se distinguió el estado de México por ser el estado donde suceden la mayoría de estos eventos criminales.

Palabras clave: Aprendizaje profundo, clasificación automática, eventos criminales.

Automatic Detection of High-risk Areas for Criminal Events through Journalistic News

Abstract. Currently in the news, criminal events are more present and each year they increase significantly. Due to this, insecurity and fear are generated in people who are informed daily through this information medium. Therefore, it is important to be aware of these events, mainly for the authorities so that they take security measures and apply the prevention policies for these criminal events. In this research work, an algorithm based on deep learning is developed, for the automatic classification of criminal events news (homicide, kidnapping, assault, suicide

and rape) and for the automatic detection of high risk areas. For the news classification, the training is done with a corpus of Twitter news headlines, and for the tests they are carried out with a corpus formed by news obtained from a local newspaper (Milenium). Finally, in the classification process an accuracy of 98 % was obtained and the state of Mexico was also distinguished for being the state where most of these criminal events occur.

Keywords: Deep learning, automatic classification, criminal events.

1. Introducción

Las noticias sobre crimen o violencia [3], giran en entorno a investigaciones de un delito convirtiéndose en el foco de atención de los lectores dejando a un lado otros temas. Las noticas con el tema del delito genera miedo y preocupación al lector puesto que ponen en peligro su seguridad y la seguridad de sus familiares. Este tipo de noticias que presentan un grado de criminalidad perjudica a la sociedad, debido a que su temor es más grande que su confianza y por ello en vez de ayudar a combatir el crimen, favorece a las organizaciones criminales. Por lo tanto, es de suma importancia analizar las noticias, dado que al hacer un análisis de las noticias se pueden obtener los eventos delictivos con más incidencia de ocurrencia por país, estado o entidad, al igual que el mes, año, etc. Además de que se pueden obtener las zonas con mayor índice de peligro y sobre todo el delito con más ocurrencia. Al obtener un análisis reforzado se puede alertar a las comunidades más vulnerables y así tomar estrictas medidas de seguridad por parte de las autoridades.

En la revisión de la literatura científica de trabajos relacionados a los objetivos de este tema de investigación, hasta este momento no se han encontrado investigaciones que permitan realizar la clasificación de eventos delictivos en noticias periodísticas a través de redes neuronales, con la clasificación de los eventos sobre: homicidio, secuestro, asalto, suicidio y violación, en noticias periodísticas mexicanas. En la revisión realizada se pudo observar que hay una variedad de estudios realizados en los temas de aprendizaje profundo en diferentes áreas, todas ellas abordadas desde el idioma inglés. Se ha abordado el tema de clasificación de crímenes a través de patrones [11], minería de textos [13], aprendizaje automático [10], todos estos estudios reportan buenas soluciones, todos ellos con un grado de dificultad. Con aprendizaje profundo se obtiene una solución más efectiva, pues gracias a todos los beneficios que nos ofrece con el tratamiento de grandes cantidades de datos por medio de varios niveles. Es por ello, que en este trabajo de investigación se pretende desarrollar un algoritmo basado en redes neuronales para la clasificación automática de noticias de eventos delictivos en el idioma español y la detección automática de zonas de alto riesgo en los estados y municipios de México, además de obtener datos que puedan ser de gran ayuda para la seguridad de las personas y sobre todo para la prevención del delito.

El presente documento se organiza de la siguiente manera: en la sección 2 se presentan los trabajos relacionados con el tema de investigación propuesto.

La sección 3 presenta la propuesta de solución donde se mencionan las técnicas que se aplicaron. La sección 4 presenta los resultados experimentales que se realizaron en la clasificación, así mismo se mencionan los datos que se requirieron para la prueba. Finalmente, la Sección 5 muestra las conclusiones y el trabajo futuro.

2. Trabajo relacionado

En esta sección, se presenta un estudio detallado de los trabajos relacionados con el objetivo de conocer investigaciones anteriores en las tareas de clasificación de noticias periodísticas, los cuales se describen a continuación.

En [11] presentan un enfoque para mejorar los patrones con información morfológica y categorías POS para reconocer y extraer eventos criminales de noticias publicadas en periódicos digitales mexicanos. Se consideran seis eventos criminales, de los grupos principales del CED (Clasificación Estadística de los Delitos) los cuales son: asesinato, violación, asalto, suicidio, secuestro y explotación sexual. El corpus de capacitación está compuesto por 1,600 noticias en español para cada evento criminal. Otra forma de clasificación automática de textos es la dependencia entre variables la cual queda plasmada en forma de enlaces en grafos de palabras co-ocurrentes como se muestra en [7], clasificaron el sentido positivo, negativo o neutral de más de 1,000 mensajes de Twitter escritos en español.

En [13] se profundiza sobre el nivel de criminalidad presente en las noticias policiales las cuales son recopiladas desde julio a diciembre del año 2011, que contiene 23,726 noticias. Se distinguieron siete temáticas, las cuales son: delitos sexuales, incendios, drogas, disturbios, homicidios, tránsito y robos. Se estudiaron dos modelos de clasificación distintos: Naïve Bayes y K-NN, seleccionando el método de evaluación cruzada k-fold (con k=10). También, en [15] extrajeron información sobre desastres naturales a partir de noticias en español con Naïve Bayes, C4.5, k-vecinos más cercanos y máquinas de soporte vectorial. El dominio de extracción se basó en la clasificación de sólo cinco clases de eventos: Forestal, huracán, inundación, sequía y sismo.

Otro enfoque que se aborda en la clasificación, son las aplicaciones de técnicas de aprendizaje automático. En [10] se presenta un proyecto donde aplican estas técnicas a la seguridad, utilizando algoritmos de aprendizaje supervisado, concretamente usan algoritmos de clasificación (*Random Tree*, *Random Forest*, J48, regresión logística, SVM y Naïve Bayes). El conjunto de datos utilizado es "KDD Cup 1999" que contiene un conjunto de conexiones clasificadas como normales o como un ataque determinado.

Además de estos modelos de clasificación en [1] abordaron arboles de decisión, *Random Forest*, Máxima Entropía, SVC Kernel Lineal, NB Multinomial, para el análisis de sentimientos. Se realizaron las pruebas con dos bases de datos alojadas en la web, la primera se llama *Sentiment Labelled Sentences Data Set*, *Movie Review Data Set* y la base de datos de incidencias, proporcionada por la empresa *Cognodata Consulting*.

Posteriormente se aborda el tema de clasificación mediante aprendizaje profundo como se aborda en [9], donde se analizó el uso de modelos de redes convolucionales (CNN), Long short Term Memory (LSTM), LSTM bidireccionales (BI-LSTM) y una aproximación híbrida entre CNN y LSTM para el análisis de sentimiento en Twitter, utilizando el conjunto de datos: InterTASS ES, que permitirá predecir la polaridad de los tweets en base a cuatro niveles: P (Positiva), N (Negativa), NEU (Neutra), NONE (sin opinión).

También en [5] se estudia la aplicación del aprendizaje profundo a la tarea del resumen automático y abstractivo de textos. El conjunto de datos con el que se trabajó se llama CNN DailyMail, que consiste en aproximadamente 300,000 pares de artículos periodísticos y su resumen se analizaron las técnicas aprendizaje profundo enfocadas al procesamiento de lenguaje natural: las redes neuronales recurrentes y los modelos encoder-decoder, entre otras.

En otra investigación que abarca este mismo tema, se presenta en [4] donde se realiza una investigación, la cual tiene como objetivo clasificar los sílabos mediante la técnica de redes neuronales conectadas de aprendizaje profundo. El conjunto de datos con el que se realizó la experimentación consta de 2,316 sílabos. El lenguaje de programación utilizado fue Python.

Además, el modelo fue comparado con respecto a algoritmos basados máquinas de soporte vectorial (SVM), Naïve Bayes y árboles de decisión. Los resultados demostraron que el modelo de aprendizaje profundo propuesto fue superior en 1.4 % con respecto a Naïve Bayes, 6.2 % con respecto a SVM y 7.2 % con respecto a árboles de decisión.

Además, en [2] propone consolidar y preparar un cuerpo con expresiones de texto extraídos de Twitter el cual permite analizar la información a través de la ciencia de datos, con el fin de utilizarlo como insumo esencial para entrenar una red neuronal convolucional (CNN), mediante técnicas de aprendizaje profundo. Como resultado de este entrenamiento, se genera un modelo de predicción de textos que puedan o no presentar signos de cyber acoso (*cyber bullying*), en los cuales se pueda manifestar agresión verbal grave, como insultos, ataques racistas, ataques homofóbicos, etc. Finalmente, los temas abordados en esta sección utilizan las medidas de evaluación tales como: validación cruzada *Macro-*F_1, *precisión*, *recall*, *Exactitud*, *ROUGE-1*, *ROUGE-2* y *ROUGE-L*

En este trabajo se propone el uso de redes neuronales profundas para la clasificación de noticias de eventos criminales. Con los trabajos revisados anteriormente se optó por utilizar este método y se llegó a la conclusión de que el aprendizaje profundo genera buenos resultados, además de que está teniendo mucho auge actualmente. También se propone realizar un análisis de las noticias clasificadas para la obtención de zonas de alto riesgo. Considerando cinco eventos criminales, los cuales son: homicidio, secuestro, asalto, suicidio y violación. El reconocimiento automático de estos cinco eventos criminales y su respectivo análisis, es crucial para que el gobierno tome decisiones sobre la implementación de políticas y estrategias de prevención del delito para prevenir eventos violentos en un futuro cercano.

3. Propuesta de solución

Para llevar acabo el trabajo propuesto se definen dos fases. La primera fase consiste en la clasificación automática de las noticias sobre eventos criminales, pues es necesario que estas noticias estén previamente clasificadas y así llevar a cabo la segunda fase, la cual consiste en la detección de zonas de alto riesgo. Teniendo la clase a la que pertenece cada noticia se procede a obtener las características más importantes, por ejemplo: ubicación, personas, organizaciones y entidades diversas, en este caso los estados y municipios de México y así mostrar estas zonas, en donde suceden con mayor frecuencia estos eventos. Posteriormente en esta sección se describen ambas fases y en las pruebas se profundizan las implementaciones.

A continuación se describe la fase de clasificación de eventos criminales en noticias periodísticas. En la cual se incluye el desarrollo de una red neuronal convolucional basada en aprendizaje profundo. El algoritmo propuesto consiste en los siguientes pasos:

1. Entrada: corpus de entrenamiento y corpus de prueba.
2. Aplicación del algoritmo de clasificación basado en redes neuronales.
3. Evaluación del modelo obtenido con los datos de prueba.
4. Salida: noticias clasificadas.

En la Fig. 1 se pueden observar las fases que se deben seguir para llevar a cabo el algoritmo propuesto.

Fig. 1. Diseño propuesto para la primera fase de clasificación automática de eventos periodísticos.

Para llevar a cabo la primera fase, que es la aplicación del algoritmo de clasificación basado en redes neuronales de los eventos criminales (clases), se realizan los pasos del Algoritmo 1.

Algoritmo 1.1: Algoritmo de clasificación basado en redes neuronales.

Require: Corpus para entrenar, corpus para realizar las pruebas y corpus de palabras preentrenadas [14].

Ensure: Corpus de noticias clasificadas en un csv.

1: Se genera la matriz de incrustaciones de palabras, donde se realiza una representación de cada palabra, donde a cada palabra que sea similar, tendrá una representación similar.
2: Se aplican las redes neuronales CNN y LSTM.
3: En la aplicación de la red neuronal convolucional (Conv1D) se utilizó la función de activación *relu*, un Dropout (0.2, 0.25, 0.275) entre otras.
4: En la aplicación de LSTM se utilizó la función de activación *relu* y *sofmax*, un Dropout(0.25) entre otras.
5: Se aplica el optimizador *Adam*.
6: Se aplica la métrica *Accuracy*.
7: Se realiza la prueba con 10 épocas.

Después de la primera fase, ya que las noticias se encuentran clasificadas en homicidio, secuestro, asalto, suicidio y violación, se procede a la segunda fase que es el análisis de las noticias. Está segunda fase consta de los siguientes pasos (ver Fig. 2):

1. Datos de entrada: corpus de noticias etiquetadas.
2. Aplicación del algoritmo de reconocimiento de entidades con nombre.
3. Datos de salida: Zonas de riesgo.

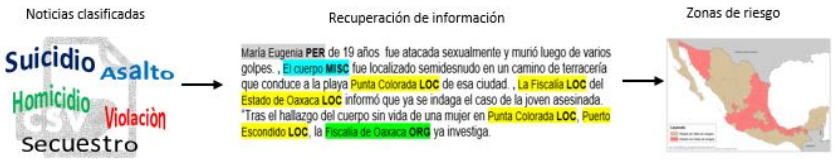

Fig. 2. Diseño propuesto para la segunda fase del análisis de las noticias etiquetadas.

4. Resultados experimentales

En esta sección se presentan los datos utilizados para este experimento. Posteriormente se muestran los resultados de clasificación y el análisis de las noticias.

4.1. Conjunto de datos

Para el conjunto de datos se cuenta con dos corpus, uno para entrenamiento y otro para prueba. El corpus para entrenamiento se recuperó de [11] el cual cuenta con encabezados de noticias obtenido de la red social de Twitter, en este artículo se mencionan seis clases de delitos los cuales son: asesinato, violación, asalto, suicidio, secuestro y explotación sexual. En la Fig. 3 se muestran los datos con lo que se cuenta para el entrenamiento. Estos datos son número del tweet, fecha de la publicación del tweet, medio en el que se transmitió, el tweet que contiene el encabezado de la noticia y la etiqueta a la que pertenece esa noticia.

numero	fecha	medio	tweet	etiqueta
9.32E+17	Fri Nov 17 22:38:04 CST 2017	NTelevisa_com	#EnPunto Peligra atención médica en zona serrana de #Chihuahua tras el secuestro del médico Blas Godínez& https://t.co/j4	Secuestro
9.32E+17	Fri Nov 17 23:07:04 CST 2017	NTelevisa_com	Se metieron a robar un cajero automático con todo y camionetas a un centro comercial de Ecatepec en el #Edomex& https://	Asalto
9.32E+17	Sat Nov 18 00:27:23 CST 2017	NTelevisa_com	Con camionetas y cuerdas intentaron robar un cajero del centro comercial Plaza Aragón, en Ecatepec, #Edomex.& https://t.c	Asalto
9.32E+17	Sat Nov 18 12:01:01 CST 2017	SSP_CDMX	Detuvimos a tres sujetos acusados de robar 24 mil pesos a un hombre de 28 años que ingresaba a su casa en la col. I& https:/	Asalto
9.32E+17	Sat Nov 18 14:07:34 CST 2017	elsolde_mexico	Comando ejecuta a dos personas en un bar en Nuevo León https://t.co/Cw9b7mQWJk https://t.co/k6QWyYurBg	Homicidio
9.32E+17	Sat Nov 18 17:00:01 CST 2017	lajornadaonline	#Morrissey cuestiona ola de denuncias de acoso sexual en #Hollywood https://t.co/vXG0zNgnS1 https://t.co/O88iogIbNK	Violación
9.32E+17	Sat Nov 18 18:16:00 CST 2017	Milenio	El #Vaticano investigará abusos sexuales en escuela de #Roma después de la publicación del libro 'El pecado origina& https:/	Violación
9.32E+17	Sat Nov 18 18:40:00 CST 2017	Pajaropolitico	La violencia aumentó en el país aún con la gendarmería: entre 2015 y 2017 los homicidios dolosos se han incrementad& https	Homicidio
9.32E+17	Sat Nov 18 19:36:00 CST 2017	Milenio	'Naturista' envenena a 14 y mata a otra en #Oaxaca https://t.co/GW8l3R9xbG https://t.co/JO7DfLe2pb	Homicidio
9.32E+17	Sat Nov 18 20:10:00 CST 2017	El_Universal_Mx	Los homicidios ocurrieron en los municipios de Naucalpan, Ecatepec y Chalco https://t.co/oPaZSNf8uy	Homicidio
9.32E+17	Sat Nov 18 20:34:01 CST 2017	Pajaropolitico	A un año de que @EPN termine, @anairamzepol y @ccastan3da, reflexionan sobre el aumento de homicidios en México.& h	Homicidio
9.32E+17	Sun Nov 19 00:17:00 CST 2017	Pajaropolitico	Los niños indígenas en México no tienen muchas opciones: son reclutados por el narco o víctimas de homicidio o desa& http:	Homicidio
9.32E+17	Sun Nov 19 04:30:00 CST 2017	NoticiasMVS	.@Uber encara demanda colectiva en #EEUU por agresiones sexuales de conductores https://t.co/zIlb5Yk21p	Violación
9.32E+17	Sun Nov 19 11:55:25 CST 2017	Notimex	15 personas murieron y 5 resultaron heridas durante una avalancha humana en #Marruecos, mientras se repartía ayuda& htt	Homicidio
9.32E+17	Sun Nov 19 13:25:00 CST 2017	El_Universal_Mx	La PGJ cumplimentó las órdenes de aprehensión en su contra por los delitos de robo a casa habitación y robo de vehí& https:	Asalto
9.32E+17	Sun Nov 19 13:45:00 CST 2017	El_Universal_Mx	El asesinato fue perpetrado por hombres armados en el llamado Triángulo Rojo de robo de hidrocarburo https://t.co/aEkdYb	Asalto
9.32E+17	Sun Nov 19 15:55:02 CST 2017	GoogleNewsMX	Silencioso aumento: precio del gas se disparó en más de 30% durante el año https://t.co/hkT3pdT2uL	Suicidio
9.32E+17	Sun Nov 19 16:49:08 CST 2017	Reforma	El director de Izzi, Adolfo Lagos, fue baleado por asaltantes que le robaron una bici en Teotihuacán https://t.co/GnR44RRByD	Homicidio
9.32E+17	Sun Nov 19 18:26:38 CST 2017	NTelevisa_com	Los homicidios contra mujeres y los #feminicidios se duplican en #Zacatecas; en lo que va del año se han registrado& https:/,	Homicidio
9.32E+17	Sun Nov 19 19:12:30 CST 2017	Milenio	#PGR colaborará en caso de asesinato de directivo de #Izzi: @EPN https://t.co/wIdLZ2wo28 https://t.co/SNFuQoEMfE	Homicidio
9.32E+17	Sun Nov 19 19:30:01 CST 2017	elsolde_mexico	#Almomento | @EPN condena asesinato del vicepresidente de @Televisa; @PGR_mx cooperará en investigación& https://t.	Homicidio
9.32E+17	Sun Nov 19 19:51:32 CST 2017	Siete24Noticias	La @PGR_mx participará en la invetigación del homicidio del vicepresidente de @Televisa https://t.co/SVcwHTHZ8f	Homicidio
9.32E+17	Sun Nov 19 20:45:29 CST 2017	El_Universal_Mx	Después de intentar robar a una familia, los delincuentes se escondieron en en una vecindad, lo que provocó un inte& https	Asalto
9.32E+17	Sun Nov 19 22:12:00 CST 2017	Milenio	#PGR colaborará en caso de asesinato de directivo de #Izzi: @EPN https://t.co/wIdLZ2eMDy https://t.co/qogt5Xhve3	Homicidio
9.32E+17	Sun Nov 19 22:40:01 CST 2017	NTelevisa_com	Los homicidios contra mujeres y los #feminicidios se duplican en #Zacatecas; en lo que va del año se han registrado& https:/,	Homicidio
9.32E+17	Mon Nov 20 00:03:56 CST 2017	Reforma	Charles Manson, quien ordenó una serie de asesinatos en 1969, murió hoy a los 83 años https://t.co/oEBdHNuDF5	Homicidio

Fig. 3. Encabezado de noticias de twitter [11].

En este caso para llevar a cabo la clasificación de nuestro interés se tomaron cinco etiquetas, las cuales son: homicidio, secuestro, asalto, suicidio y violación. Esta información fue utilizada para entrenar el algoritmo, como se muestra en la Tabla 1, la cual contiene: los temas de encabezados de noticias de Twitter con las etiquetas respectivas de los eventos delictivos que se clasificarán, la cantidad de noticias y su respectivo vocabulario por tema, además se realiza un recuento total de noticias y su vocabulario

En el caso del corpus para probar, la fuente de información por la que se optó es Milenio una página de noticias local, como se mencionó anteriormente, debido a que esta página nos proporciona un apartado de noticias Policíacas en la cual contiene las noticias de eventos criminales, además de que cuenta con noticias de toda la república mexicana. También se revisaron noticias de otros periódicos digitales, tales como: el Sol de Puebla, el Sol de México y la Prensa, pero solo se proporciona información de temas en general, por lo que no fueron considerados.

Con ayuda del lenguaje de programación Python, se crea un raspado web con las bibliotecas scrapy [8] y BeautifulSoup [12].

Tabla 1. Datos para entrenar.

Tipo de noticias	Cantidad	Total de vocabulario
Homicidio	4,352	95,734
Suicidio	334	5,844
Asalto	2,997	59,568
Secuestro	365	7,183
Violación	666	13,704
Total	8,714	182,033

En el Algoritmo 2 se muestran los pasos que se siguieron para la obtención de las noticias periodísticas por medio de la página local: Milenio. El algoritmo consiste en la identificación de los elementos más importantes de la página de milenio los cuales son: título, contenido, fecha y lugar de la noticia, posteriormente recupera la información de estos elementos en modo texto, se estructura y se genera un archivo csv con la información recuperada de la noticia. Cada noticia está estructurada de la siguiente forma: contienen un título, encabezado, autor, fecha, lugar y el cuerpo de la noticia. En la Fig. 4 se muestran los elementos principales extraídos, con ayuda de la función *add_xpath* se recupera el texto de las etiquetas título, fecha y lugar y para la etiqueta de contenido se recupera el texto con la función *soup.find* de la página html.

Algoritmo 1.2: Algoritmo de raspado web.

Require: Página local Milenio
Ensure: Archivo estructurado csv que contiene título, contenido, fecha y lugar de la noticia
1: Definir las etiquetas a extraer de la página HTML:
2: Título
3: Contenido
4: Fecha
5: Lugar
6: Obtener la URL de la sección "Policía" de la página local (Milenio)
7: Obtener el xpath de la etiqueta título, fecha y lugar de la página HTML
8: Obtener el identificador de la etiqueta contenido de la página HTML
9: Ejecución del código para la recuperación del texto con las funciones *add_xpath* y *soup.find*, de las etiquetas título, contenido, fecha y lugar
10: Eliminar noticias duplicadas.

Después de aplicar el algoritmo de extracción de noticias de la página local Milenio se lograron recuperar 179 noticias del 06 de agosto del 2019 al 27 de noviembre del 2019. En la Tabla 2 se puede observar el tipo de delito (clase), así como el total de vocabulario y la cantidad de noticias en cada delito.

Fig. 4. Variables extraídas de la página local (Milenio).

4.2. Resultados

A través del uso de la red neuronal convolucional (CNN) y una red neuronal llamada memoria a corto y largo plazo (LSTM), se realizó el entrenamiento con 8,714 encabezados de noticias y la prueba se realizó con 179 noticias. Los resultados de la clasificación se pueden ver en la Tabla 3, la medida de evaluación que presentó mejores resultados fue la de exactitud, donde se logró obtener un 98 % de clasificación.

En la segunda fase, sobre el análisis de las noticias, se aplicó un algoritmo de reconocimiento de entidades con nombre, a través de la librería de Python en Spacy [6], donde se utilizó el modelo *es_core_news_sm* que proporciona Spacy para el idioma español, el cual asigna vectores de token específicos de contexto, etiquetas POS, análisis de dependencia y entidades con nombre. Admite la identificación de entidades como nombres de personas (PER), organizaciones (ORG), la ubicación definida política o geográficamente (LOC) y entidades diversas (MISC). En la Fig. 5 se presenta un ejemplo del resultado de la aplicación del algoritmo de reconocimiento de entidades con nombre a una noticia de la página local Milenio.

Tabla 2. Datos para la prueba.

Tipo de noticias (clase)	Total de vocabulario	Cantidad de noticias
Homicidio	40,240	116
Suicidio	0	0
Asalto	7,799	23
Secuestro	12,211	30
Violación	4,046	10
Total	64,296	179

Tabla 3. Resultado de la prueba.

Tipo de noticias	Cantidad de noticias	Clasificadas por el sistema
Homicidio	116	117
Suicidio	0	5
Asalto	23	32
Secuestro	30	17
Violación	10	8
Total	179	179

Como puede observarse se identificaron estados tales como el de Oaxaca (LOC), organizaciones como la fiscalía de Oaxaca (ORG), personas como María Eugenia (PER) y entidades diversas tales como: El cuerpo (MISC).

El cuerpo de una mujer, que había sido reportada como desaparecida por sus familiares fue hallado en las inmediaciones de Puerto Escondido **LOC**, Oaxaca **LOC**. De acuerdo con el reporte de la Comandancia Regional de la Policía Estatal **MISC**, la víctima identificada como María Eugenia **PER** de 19 años fue atacada sexualmente y murió luego de varios golpes. , El cuerpo **MISC** fue localizado semidesnudo en un camino de terracería que conduce a la playa Punta Colorada **LOC** de esa ciudad. , La Fiscalía **LOC** del Estado de Oaxaca **LOC** informó que ya se indaga el caso de la joven asesinada. "Tras el hallazgo del cuerpo sin vida de una mujer en Punta Colorada **LOC**, Puerto Escondido **LOC**, la Fiscalía de Oaxaca **ORG** ya investiga bajo los parámetros establecidos en el protocolo de feminicidios" **MISC**. Tras el hallazgo del cuerpo sin vida de una mujer en Punta Colorada **LOC**, Puerto Escondido **LOC**, la FISCALIA_GobOax **LOC** ya investiga

Fig. 5. Ejemplo de reconocimiento de entidades con nombre.

Después de la aplicación del algoritmo de entidades con nombre, procedemos a la detección de características particulares en el texto tales como el lugar en el que ocurren estos eventos delictivos. Enfocándonos solo en ubicaciones, en la Fig. 6 se presentan las ubicaciones que pueden ser estados y municipios de la república mexicana.

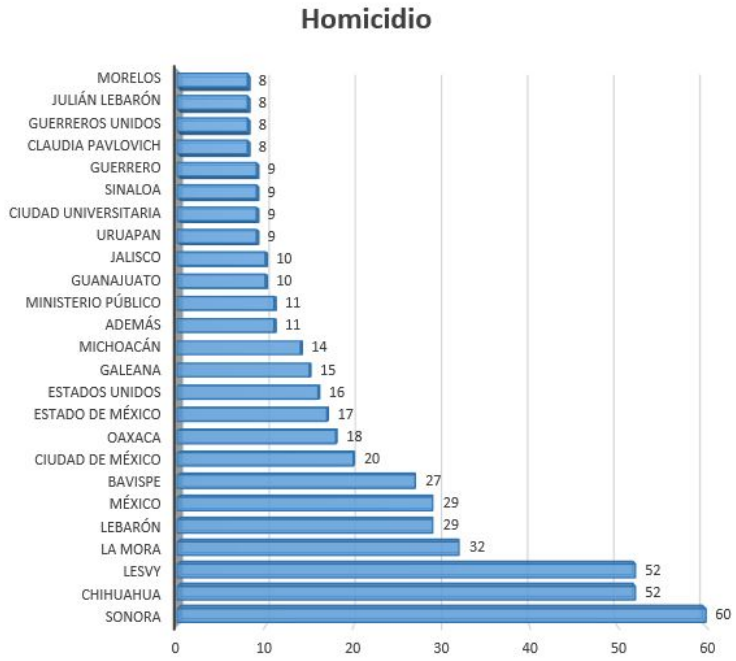

Fig. 6. Gráfica de homicidio.

En la Fig. 6 se observa que el estado de Sonora cuenta con el mayor número de homicidios, seguido del estado de Chihuahua, además se ubica el nombre Lesvy, entre otros. El algoritmo de etiquetado de entidades con nombre muestra algunas ubicaciones las cuales no son necesariamente estados, comunidades o municipios, en este caso el etiquetador cuenta con algunos errores de la extracción de la ubicación.

Las noticias clasificadas en asalto, secuestro, suicidio y violación sexual, presentan un comportamiento muy similar con respecto a los estados obtenidos en la clase de homicidio. Las zonas de alto riesgo identificadas fueron Sonora, Chihuahua, Ciudad de México, Oaxaca, Estado de México, entre otros. Estos estados fueron identificados con el mayor índice de delitos tanto en homicidio como en asalto, secuestro, suicidio y violación sexual. Con los resultados obtenidos en el etiquetador de entidades con nombre se observaron algunos errores, por ejemplo: "La Mora" corresponde con un apellido, el nombre de un rancho y una localidad.

5. Conclusiones

En este trabajo de investigación se desarrolló un algoritmo que consta de dos fases. En la primera fase se desarrolló un algoritmo para la clasificación automática

de eventos delictivos en noticias periodísticas basado en redes neuronales. En la segunda fase se desarrolló un algoritmo para el análisis de las noticas clasificadas, este análisis consiste en la identificación de entidades con nombre, por ejemplo: el estado en la que ocurre el delito. En base a los resultados experimentales en la fase de clasificación se obtuvo una exactitud del 98 %, con esta métrica se obtuvieron los mejores resultados de evaluación. En la fase del análisis de las noticias clasificadas se obtuvieron los estados con mayor frecuencia, resaltando el estado de Sonora, Chihuahua, Ciudad de México, Oaxaca, Estado de México, donde existe un mayor índice de delincuencia con respecto a homicidio, asalto, secuestro, suicidio y violación sexual. Como trabajo futuro se observó que es necesario hacer una mejora en los resultados obtenidos por el algoritmo de entidades con nombre, además de incorporar otras técnicas utilizadas en la literatura como TF-IDF o añadir un recurso para identificar los estados y sus municipios con la finalidad de identificar la frecuencia de los estados presentes en cada noticia, todo ello para mejorar los resultados obtenidos en el etiquetador de entidades con nombre y obtener resultados precisos en cuanto a zonas de alto riesgo.

Agradecimientos. Esta investigación es parcialmente apoyada por el Fondo Sectorial de Investigación para la Educación, con el proyecto CONACyT CB/257357 y por el proyecto VIEP-BUAP 100409344-VIEP2019.

Referencias

1. Álvaro., G.G.: Machine Learning en Bases de Datos de Lenguaje Natural. Master's thesis, Universidad Autónoma de Madrid. Departamento de Ingeniería Informática (2016)
2. Armijos, P.D.C.: Predicción de ataques de cyber bullying mediante técnicas de aprendizaje profundo apoyandose en un corpus de entrenamiento para la clasificación de texto en español. Master's thesis, Universidad Internacional SEK (2018)
3. Beatriz Magaloni, A.D.C.y.V.R.: La raíz del miedo: ¿por qué es la percepción de riesgo mucho más grande que las tasas de victimización? las bases sociales del crimen organizado y la violencia en méxico, josé antonio aguilar ed. mexico city: Secretaría de seguridad pública (2012)
4. Carrera, J.M.A.: Clasificación de sílabos académicos en base a redes neuronales de aprendizaje profundo (2018)
5. Hernández, A.J.A.: Deep Learning aplicado al resumen de textos. Master's thesis, Universidad Complutense de Madrid (2018)
6. Honnibal, M., Montani, I.: Spacy 2.0: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
7. Juan Pablo Cárdenas, G.O., Alfaro, R.: Clasificación automática de textos usando redes de palabras. Revista Signos 4(86), 346–364 (2013)
8. Kouzis-Loukas, D.: Learning Scrapy. Packt Publishing Ltd (2016)
9. R. Montañés, R. Aznar, y.R.D.H.: Aplicación de un modelo hibrido de aprendizaje profundo para el análisis de sentimiento en twitter. CEUR Workshop Proceedings 2172, 3027–3036 (2018)

10. Rama, J.M.R.: Aplicación de técnicas de machine learning a la detección de ataques. Master's thesis, Universitat Oberta de Catalunya (2017)
11. Reyes-Ortiz, J.A., Bravo, M.: Enhancing patterns with linguistic information for criminal event recognition. Journal of Intelligent and Fuzzy Systems 34(5), 3027–3036 (2018)
12. Richardson, L.: Beautiful soup documentation. April (2007)
13. Silva, T., Alejandro, D.: Diseño y aplicación de una metodología para análisis de noticias policiales utilizando minería de textos. Master's thesis, Universidad de Chile (2013)
14. Tatman, R.: Pre-trained word vectors for spanish (2019), https://www.kaggle.com/rtatman/pretrained-word-vectors-for-spanish
15. Valero, A.T.: Extracción de Información con Algoritmos de Clasificación. Master's thesis, Instituto Nacional de Astrofísica, óptica y Electrónica (2005)