ISSN 1870-4069

Creación y clasificación de un corpus criminológico en español usando características lingüísticas superficiales

Luis Gil Moreno[1], Noé Alejandro Castro[1], Juan-Manuel Torres-Moreno[2,3], Luis-Adrián Cabrera-Diego[2], Carlos-Emiliano González-Gallardo[2], Alberto Iturbe[1], Kenia Nieto[1], Arturo-Michel Gómez[1]

[1] Centro Nacional de Investigación y Desarrollo Tecnológico CENIDET, Cuernavaca, México

[2] LIA/Université d'Avignon et des Pays de Vaucluse, Avignon, France

[3] École Polytechnique de Montréal, Montréal, Canada

{luismoreno,ncastro,arturog,iturbe,kenianieto}@cenidet.edu.mx,
{juan-manuel.torres,luis-adrian.cabrera-diego}@univ-avignon.fr,
carlos.gonzalez-gallardo@alumni.univ-avignon.fr

Resumen. Este artículo propone la creación y la caracterización de un corpus especializado en criminología. El corpus está constituido por noticias en texto plano divididas en cinco clases de delitos: homicidio, asalto, secuestro, abuso sexual y extorsión. El objetivo es doble: El primero es crear y anotar manualmente el corpus. Mientras que el segundo objetivo consiste en establecer una clasificación de base usando características lingüísticas superficiales, como los sintagmas nominales y verbales. Los clasificadores utilizados son una Máquina de Soporte Vectorial (SVM) clásico y un modelo Bayesiano.

Palabras clave: Corpus, criminología, clasificación automática, sintagma, máquina de aprendizaje, extracción de información.

Creation and Classification of a Spanish Criminological Corpus using Superficial Linguistic Features

Abstract. This article proposes the creation and characterization of a specialized corpus in criminology. The corpus is composed of news in plain text divided into five classes of crimes: homicide, assault, kidnapping, sexual abuse and extortion. The corpus' objective is twofold: one to create and annotate manually the corpus. Two, to establish a basic classification using superficial linguistic features, such as noun and verb syntagms. We have used two classifiers: a classical Support Vector

Machine (SVM) and Bayesian model.

Keywords. Corpus, criminology, automatic classification, syntagm, machine learning, information extraction.

1. Introducción

De acuerdo con la investigación realizada por [9], México se encuentra entre los 20 países más violentos y peligrosos del mundo. A nivel Centroamérica y el Caribe, México ocupa el segundo lugar en esta clasificación. Por ello es frecuente encontrar en periódicos de circulación nacional, estatal y local, noticias que hagan referencia a diferentes delitos, como el secuestro, el asalto, el abuso sexual, entre otros.

En este artículo se proponen dos tareas relacionadas con el tema evocado. La primera de ellas es la creación de un corpus especializado de noticias delictivas usando diferentes periódicos del Estado de Morelos, México. La segunda tarea es la utilización de herramientas de Procesamiento del Lenguaje Natural (PLN) con el objetivo de clasificar automáticamente noticias que traten sobre delitos. Específicamente, se busca crear un clasificador de noticias que permita determinar cuáles son los delitos tratados al interior de una nota informativa, para que de esta forma se establezca un *baseline*. Los delitos a identificar por el clasificador son aquellos denominados de *Alto impacto* [15].

El artículo está organizado de la siguiente manera: en la Sección 2. se presenta el estado del arte. En la Sección 3. se encuentra la descripción del corpus textual de delitos y su estadística descriptiva. Posteriormente, en la Sección 4. se da a conocer la metodología que se siguió. En la Sección 5. se presentan los resultados obtenidos de los experimentos. Finalmente, en la Sección 6. se presentan las conclusiones y perspectivas.

2. Estado del arte

En los últimos años se han desarrollado diferentes investigaciones relacionadas al uso del PLN, y especialmente de la Extracción de Información (EI) en temáticas delictivas [12].

Sin embargo, la mayoría de las investigaciones en este campo se han realizado para el idioma inglés. Por ejemplo, [23] utiliza un modelo predictivo con un análisis semántico para inferir posibles crímenes a partir de *tweets*[1]. Otro ejemplo, es el de [10], el cual utiliza técnicas de EI y modelos cognitivos para aumentar la información obtenida de entrevistas con testigos criminales. [14] utiliza métodos de agrupamiento automático *(clustering)*, basados en *k-means*, para encontrar patrones criminales. [11] utiliza medidas de similitud y de aprendizaje automático para analizar y clasificar textos que describen crímenes, ya

[1] `http://www.twitter.com`

sean distintos, similares o los mismos. [6] utiliza una red de neuronas artificiales (*Artificial Neural Network, ANN)* para encontrar patrones de clasificación en bases de datos delictivas de la policía. Finalmente, en el trabajo desarrollado por [4], se procesan los reportes delictivos de la policía de Arizona en EE.UU, en búsqueda de elementos relevantes, por ejemplo, nombres de personas, drogas, armas o eventos criminales.

Para el idioma portugués, se encuentra el trabajo de [17] en el que los autores emplean la EI y un análisis semántico para enriquecer automáticamente la información sobre crímenes presentes en el sitio colaborativo *WikiCrimes*[2]. Por otro lado, para el idioma alemán se encontró el trabajo [22] en el cual se propone la creación de un corpus especializado del plagio con la intención de utilizarlo para la detección de plagio de documentos. No obstante, para el idioma español, según nuestro conocimiento, no existen investigaciones similares.

La importancia de la clasificación automática de textos criminales, recae en el hecho de que puede ser empleado en un análisis criminalístico [7] y para la detección de entidades criminales [5]. Incluso, la clasificación de noticias puede ayudar a encontrar patrones en reportes delictivos u otro tipo de aspectos criminalísticos [14].

2.1. Características utilizadas

Con base en la investigación presentada por [12], se puede concluir que la mayoría de las trabajos que buscan patrones criminales usan elementos específicos como armas de fuego o armas blancas. Sin embargo, existen otras investigaciones que emplean el PLN, más específicamente la EI, para llevar a cabo esta tarea.

En el trabajo realizado por [12] se realiza un estudio exhaustivo sobre los diferentes entornos de trabajo o *Frames* existentes, los cuales analizan patrones en búsqueda de tendencias criminales. Aunque la mayoría utiliza sensores en búsqueda de elementos más específicos como armas de fuego o armas blancas, existen algunos que trabajan directamente con herramientas de PLN.

Las técnicas de EI son variadas y muchas veces se realizan con base en la frecuencia de términos en los textos. En [13], se utiliza el modelo TF-IDF donde se vectorizan los textos de entrada; los términos con una frecuencia elevada así como las palabras vacías se discriminan, mientras que aquellos con frecuencia única, se consideran como Entidades Nombradas (EN).

En el trabajo de [18] se ocupa la misma técnica, sólo que esta vez el modelo TF-IDF se emplea para la extracción de características. Posteriormente, las características extraídas se utilizan para clasificar los textos usados ocupando la medida de similitud coseno.

Analizando los resultados de los trabajos mencionados, se optó por implementar dichas técnicas para la extracción de características, con la diferencia de proponer un modelo diseñado para textos cortos. Esto para permitir mantener la misma calidad en las características extraídas, pero aumentando la velocidad

[2] `http://www.wikicrimes.org`

de procesamiento. Se observó que se podría realizar una adaptación de los trabajos de [1] y [19]. En el primero de ellos, se hace una adaptación del modelo TF-IDF para textos cortos y en el segundo se toman bigramas poco comunes como candidatas a EN. Con estas técnicas se podrán realizar búsquedas de términos relevantes en las noticias y considerarlas entonces como características que pudieran describir cada clase que se tiene como objeto de estudio en esta investigación.

3. Corpus textual y estadística descriptiva

A continuación, se describe el proceso que se siguió para la conformación del corpus presentado en este trabajo. De igual manera se detalla la información estadística sobre el mismo.

3.1. Construcción del corpus

Para la conformación del corpus, fue necesario la descarga de noticias a través de diversos portales periodísticos locales en el Estado de Morelos, México. Se eligieron periódicos de circulación local, ya que son en estos medios en donde se reporta mayormente la actividad delictiva de la zona.

Para la extracción de noticias, se desarrolló un módulo ocupando la biblioteca JSoup [8]. Esta analiza la estructura HTML de una página web para ubicar y extraer específicamente la información deseada. En este caso, lo que interesa para el estudio es el cuerpo de la nota periodística.

El corpus fue constituido con noticias descargadas de los siguientes periódicos:

- La Unión de Morelos[3],
- El Diario de Morelos[4],
- La Jornada de Morelos[5].

Se descargaron en total 1 000 noticias que fueron almacenadas en texto plano en formato *utf8*. Los documentos recuperados cumplen la condición de reportar al menos uno de los siguientes delitos:

1. Homicidio,
2. Asalto,
3. Secuestro,
4. Abuso sexual.

Cabe destacar que al momento de etiquetar las noticias, los anotadores se percataron de la existencia de una clase adicional presente en un número

[3] http://www.launion.com.mx
[4] http://www.diariodemorelos.com
[5] http://www.jornadamorelos.com

significativo de notas periodísticas: la extorsión. Por tanto, se agregó esta clase como uno de los posibles delitos del corpus.

Las noticias fueron descargadas en dos períodos de tiempo, para evitar una tendencia sobre ciertos hechos específicos. Por ejemplo, la descarga de todas las noticias posteriores al asesinato de un presidente, implicaría que la mayoría de las notas descargadas serían destinadas a la clase homicidio. Así, el primer intervalo de tiempo de descarga de noticias fue del 11 al 15 de abril de 2016, y el segundo intervalo de descargas tuvo lugar del 07 al 09 de septiembre de 2016.

3.2. Proceso de anotación manual

Para la tarea de anotación manual del corpus de noticias, se seleccionaron cuatro estudiantes universitarios con nivel de maestría. El corpus completo anotado de noticias se nombró como Corpus Anotado de Delitos (CAD)[6].

A cada una de estas personas se les proporcionaron aleatoriamente y sin repetición 250 noticias. Considerando que cada noticia posee en promedio 371.6 palabras, significa que cada uno de ellos tuvo que analizar en total un promedio de 93 000 palabras.

Cada noticia fue clasificada manualmente en al menos una de las cinco posibles clases según la información contenida. En otras palabras, una noticia puede contener múltiples actos delictivos y por consiguiente, pertenecer a más de una clase. Esto fue detectado durante el proceso de etiquetado manual.

Finalmente, las cinco clases retenidas para el presente experimento fueron las siguientes:

1. Homicidio,
2. Asalto,
3. Secuestro,
4. Abuso sexual,
5. Extorsión.

Se midió el tiempo usado por los anotadores en el experimento. La tarea de anotación se llevó a cabo en 17 horas. La anotación de cada noticia necesitó de aproximadamente tres minutos.

En la Tabla 1 se muestran las estadísticas básicas del corpus CAD.

3.3. Palabras clave de las clases

Además de la anotación manual hecha por los cuatro anotadores, se les pidió que estos realizaran una lista de las palabras claves que les permitiera clasificar las noticias. De esta forma, no solamente se obtuvo una anotación manual del corpus, sino también un conjunto de términos, mono o multipalabra, que se usan de manera recurrente en el corpus. En la Tabla 2 se muestran los términos, en su forma canónica, encontrados por los anotadores para cada clase.

[6] El corpus CAD podrá ser solicitado a través del correo: `luismoreno@cenidet.edu.mx`

Tabla 1. Corpus CAD en función de los documentos.

Clases	La Unión de Morelos	El Diario de Morelos	La Jornada de Morelos	Total	Palabras (tokens)
Homicidio	139	130	125	394	146 410
Asalto	145	160	136	441	164 990
Secuestro	101	95	109	305	113 338
Abuso Sexual	45	48	62	155	55 598
Extorsión	36	69	45	150	55 740

Tabla 2. Palabras clave del corpus CAD.

Clase	Palabras clave
Asalto	robo, sustraer, hurtar, amenazar, despojar, interceptar, quitar, desvalijar, desmantelar, sorprender en posesión, ocultar
Homicidio	encontrar sin vida, linchamiento, asesinar, atropellar, hallar muerto, cadáver, disparar, baleado, balazo, acribillar, atacar a tiro, persona sin vida, morir, dar disparos
Secuestro	secuestro, subir a la fuerza, forcejear, libertad, levantar persona, víctima rescatada, privación, liberar víctima, llevar a la fuerza, rescate
Abuso sexual	violación, agredir de forma sexual, íntima, estupro
Extorsión	extorsionar, golpear, obligar, intimidar, amenazar

Como se puede observar, la gran mayoría de las palabras claves encontradas en cada clase corresponden a verbos (por ejemplo: *despojar, forcejear, disparar*) y sustantivos (*privación, estupro, cadáver*). Sin embargo, también se encuentran algunas formaciones como verbo-sustantivo (*levantar-persona, liberar-víctima*), verbo-adjetivo (*hallar-muerto*) y adjetivo (*íntima, baleado*).

4. Metodología

La metodología propuesta en este artículo esté dividida en dos partes: Extracción de las características (Sección 4.1.) y Clasificación de noticias por el contenido (Sección 4.2.).

4.1. Extracción de características

Este proceso se basó en las conclusiones y resultados obtenidos por [3]. Este artículo argumenta que son los sintagmas nominales los que mejor describen la información de un texto. En este caso en particular, interesa la identificación del tipo de delito que se reporta en la nota. Para este proyecto consideramos que, los actos que se detallan en el texto, pueden analizarse mediante la identificación de los verbos. Lo anterior, coincide con las conclusiones de los anotadores, quienes observaron que los sustantivos y verbos son los que más información proveen

para la detección del tipo de delito. Por tanto, también se estudiarán sintagmas verbales.

La adecuada extracción de los sintagmas es de vital importancia para las actividades posteriores, ya que esta será la que defina el conjunto de características que constituirán una bolsa de palabras. La bolsa de palabras será la representación utilizada tanto en la fases de aprendizaje como en la fase de pruebas de los clasificadores.

El proceso que se sigue para la extracción de las características se describe a continuación:

Primeramente, cada uno de los textos se anota con etiquetas que indican la categoría gramatical de las palabras (POS o *Part-of-Speech* en inglés), usando la herramienta Freeling [16].

Después, a partir de las etiquetas POS de las palabras de cada texto, se extraen los siguientes patrones sintácticos:

- Sintagmas verbales (VP)[7],
- Sustantivos,
- Verbos.

Una vez extraídos estos patrones sintácticos, se calcula el grado de importancia de cada palabra que aparece en los sintagmas. Esto se realiza con base en las investigaciones de [2]. Para llevar a cabo esto, se multiplica el número de palabras que compone el sintagma (|VP|) por la frecuencia de cada palabra del sintagma (UF o Unigram Frequency en inglés), dicha operación se formaliza en la Ecuación 1:

$$\mathrm{UF(VP)} = \sum_{i=0}^{|\mathrm{VP}|} \text{Unigram Frequency}(w_i). \tag{1}$$

Posteriormente, según con el modelo propuesto por [2], el resultado obtenido en la Ecuación 1 se multiplica por la frecuencia del sintagma en el artículo, (VPF(VP)). El resultado se divide entre la cantidad de palabras que compone el sintagma (|VP|):

$$Score(\mathrm{VP}) = \frac{\mathrm{UF(VP)} \cdot \mathrm{VPF(VP)}}{|\mathrm{VP}|}. \tag{2}$$

Una vez que se han determinado los grupos y sus elementos, se calcula su puntuación. El cálculo de lo antes mencionado no es más que una media aritmética. Siendo más específicos, esto se hace mediante la sumatoria del *score* de cada elemento, perteneciente al grupo, sobre la cantidad de elementos que pertenecen al mismo grupo. En la Ecuación 3 se presenta la fórmula utilizada para el cálculo del *score* de cada grupo:

$$Score(Grupo) = \frac{\sum_{i=0}^{|Grupo|} Score(\mathrm{VP}_i)}{|Grupo|}. \tag{3}$$

[7] Son aquellas construcciones que se componen de un verbo y su complemento.

Finalmente, se estableció un umbral para delimitar las palabras o sintagmas más importantes del texto procesado. Estas palabras son las que conformarán las bolsas de palabras o características que definirán cada clase de delitos. El umbral se estableció con respecto a los valores medios de cada clase, eliminando así aquellos *scores* de grupos demasiado elevados. Tal es el caso de *stopwords* o palabras que no dan ninguna descripción relevante sobre el documento. Igualmente, representan aquellas clases con *scores* muy bajos, como lo son los nombres de personas, lugares y fechas (las Entidades Nombradas).

4.2. Clasificación *baseline* del corpus

Con el propósito de establecer una medida básica de desempeño, se decidió utilizar dos clasificadores sobre el corpus anotado. El corpus CAD fue dividido en un corpus de aprendizaje (CA) y un corpus de prueba (CP). La distribución de noticias en cada subcorpus fue obtenida aleatoriamente con una distribución uniforme. Esto se llevó a cabo, para garantizar la misma distribución que en el corpus CAD. La tarea consistió entonces en determinar a qué clase pertenece cada noticia.

El corpus de aprendizaje CA está formado por un subconjunto de noticias del 70% del total del corpus CAD, y el corpus de prueba del 30% restante. Estos subconjuntos de noticias se sometieron al proceso de extracción de características descrito en la Sección 4.1.. Como datos de entrenamiento, fueron ocupadas las características extraídas del conjunto de noticias anotadas manualmente (ver Sección 3.2.). Estas características fueron usadas para analizar la clase a la que corresponden las noticias del corpus CP.

Para la clasificación del CP, se empleó la plataforma Weka[8], que permite trabajar con diversos algoritmos de clasificación. En este estudio se realizaron pruebas con un modelo Bayesiano (Naïve Bayes) y con una Máquina de Soporte Vectorial (SVM).

5. Resultados y evaluación

Los resultados que se presentan a continuación, contemplan tres experimentos. En el primero se sometió a análisis la noticia completa: Tabla 3 y Tabla 4. En el segundo, fueron utilizados únicamente el título de la noticia y el primer párrafo: Tabla 5 y Tabla 6. Finalmente, se hizo la última prueba considerando únicamente el título de la noticia: Tabla 7 y Tabla 8. En todos los casos, para la evaluación se utilizó la medida clásica F-Score, definida por la Ecuación 4:

$$\text{F-Score} = \frac{2 \times (\text{Precisión} \times \text{Recall})}{\text{Precisión} + \text{Recall}}. \tag{4}$$

En la columna "Media" de las Tablas 3–8, se indica el promedio de Precisión, *Recall* y *F-Score* de cada experimento.

[8] `http://www.cs.waikato.ac.nz/ml/weka`

Tabla 3. Resultados de la clasificación (Noticia completa - SVM).

Clases	Asalto	Homicidio	Secuestro	Abuso Sexual	Extorsión	Media
Precisión	0.7348	0.7613	0.6851	0.6000	0.8032	0.7169
Recall	0.8818	0.6146	0.6851	0.3846	0.9074	0.6947
F-Score	0.8016	0.6802	0.6851	0.4687	0.8521	0.6975

Tabla 4. Resultados de la clasificación (Noticia completa - Naïve Bayes).

Clases	Asalto	Homicidio	Secuestro	Abuso Sexual	Extorsión	Media
Precisión	0.7638	0.8033	0.6143	0.5625	0.7123	0.6912
Recall	0.8818	0.4495	0.7963	0.4615	0.9630	0.7104
F-Score	0.8186	0.5765	0.6935	0.5070	0.8189	0.6829

Tabla 5. Resultados de la clasificación (Título de la noticia y primer párrafo - SVM).

Clases	Asalto	Homicidio	Secuestro	Abuso Sexual	Extorsión	Media
Precisión	0.9167	0.6320	0.5200	0.5161	0.8636	0.6897
Recall	0.7000	0.7248	0.4815	0.4103	0.7037	0.6040
F-Score	0.7938	0.6752	0.5000	0.4571	0.7755	0.6403

Tabla 6. Resultados de la clasificación (Título de la noticia y primer párrafo - Naïve-Bayes).

Clases	Asalto	Homicidio	Secuestro	Abuso Sexual	Extorsión	Media
Precisión	0.9310	0.6737	0.5192	0.5278	0.6719	0.6647
Recall	0.7364	0.5872	0.7297	0.4872	1.1316	0.7344
F-Score	0.8223	0.6275	0.6067	0.5067	0.8431	0.6813

Tabla 7. Resultados de la clasificación (Título de la noticia - SVM).

Clases	Asalto	Homicidio	Secuestro	Abuso Sexual	Extorsión	Media
Precisión	0.8971	0.4167	0.5833	0.7200	0.8000	0.6834
Recall	0.5545	0.8716	0.6034	0.1915	0.6207	0.5683
F-Score	0.6854	0.5638	0.5932	0.3025	0.6990	0.5688

Se puede observar, que no existe una gran diferencia entre los resultados dados por los SVM ni por los modelos Bayesianos. Esto se puede deber a que los SVM no fueron optimizados sobre el corpus CAD. A pesar de ello, la Media de F-Score usando SVM es la más elevada, con F-Score = 0.6975. Cambiando los parámetros del SVM se podrían obtener todavía mejores resultados.

A partir de lo expresado por los anotadores, quienes en diversas ocasiones

Tabla 8. Resultados de la clasificación (Título de la noticia - Naïve Bayes).

Clases	Asalto	Homicidio	Secuestro	Abuso Sexual	Extorsión	Media
Precisión	0.9405	0.4316	0.6271	0.7273	0.6897	0.6832
Recall	0.7182	0.7523	0.7400	0.2553	0.8000	0.6532
F-Score	0.8144	0.5485	0.6789	0.3780	0.7407	0.6321

encontraron más de un delito en una misma noticia, se considera que una forma de mejorar los resultados (en Precisión, Recall y F-Score) podría consistir en analizar las noticias con un clasificador multiclase. Un clasificador de este tipo permitiría incluir una misma noticia en dos o más clases. Incluso, se podría mejorar el F-Score en noticias ligadas a "Homicidio", "Secuestro" y "Abuso sexual" que presentan los porcentajes más bajos, debido al gran recubrimiento que existe entre estas.

6. Conclusiones

En este artículo se ha introducido el Corpus Anotado de Delitos en México, CAD. Se ha caracterizado este corpus y se han presentado algunas medidas *baseline* para la clasificación automática de cinco categorías de delitos. El corpus CAD puede ser utilizado en tareas de clasificación de delitos usando herramientas de PLN. Estas herramientas de análisis podrían ser de utilidad para las diferentes instancias de gobierno (policía, institutos de la juventud, etc.) así como para organizaciones descentralizadas (comisiones de derechos humanos) o no gubernamentales.

Algunos ejemplos de posibles herramientas, son los mapas donde se indican los delitos de alto impacto, los buscadores de noticias delictivas, las herramientas de documentación y la generación de síntesis de delitos [20] entre otros.

A futuro, se estudiará si este esquema de clasificadores clásicos u otros, incluyendo Redes de Neuronas Artificiales (ANN) de tipo incremental [21], podrían funcionar con textos cortos (resúmenes de noticias, *tweets*, denuncias por Internet, etc.). Igualmente, se planea aumentar el número de noticias del corpus CAD, para abarcar diferentes periodos de tiempo.

Agradecimientos. Este trabajo fue financiado parcialmente por el *Programa de Becas Mixtas de CONACYT* (México), el *Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET)* (México) y por el *Laboratoire Informatique d'Avignon (LIA)* de la *Université d'Avignon et des Pays de Vaucluse* (Francia).

Referencias

1. Bollen, J., Mao, H., Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM 11, 450–453 (2011)

2. Bracewell, D.B., Ren, F., Kuriowa, S.: Multilingual single document keyword extraction for information retrieval. In: 2005 International Conference on Natural Language Processing and Knowledge Engineering. pp. 517–522. IEEE (2005)
3. Bracewell, D.B., Yan, J., Ren, F., Kuroiwa, S.: Category classification and topic discovery of japanese and english news articles. Electronic Notes in Theoretical Computer Science 225, 51–65 (2009)
4. Chau, M., Xu, J.J., Chen, H.: Extracting meaningful entities from police narrative reports. In: Proceedings of the 2002 annual national conference on Digital government research. pp. 1–5. Digital Government Society of North America (2002)
5. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M.: Crime data mining: a general framework and some examples. Computer 37(4), 50–56 (2004)
6. Dahbur, K., Muscarello, T.: Classification system for serial criminal patterns. Artificial Intelligence and Law 11(4), 251–269 (2003)
7. Estivill-Castro, V., Lee, I.: Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In: Proc. of the 6th International Conference on Geocomputation. pp. 24–26. Citeseer (2001)
8. Hedley, J.: Java html parser. `http://jsoup.org/` (2016)
9. Institute for Economics & Peace: Índice de Paz en México 2015, Un análisis de la dinámica de los niveles de paz en México. Report IEP 31, México (2015)
10. Ku, C.H., Iriberri, A., Leroy, G.: Crime information extraction from police and witness narrative reports. In: Technologies for Homeland Security, 2008 IEEE Conference on. pp. 193–198. IEEE (2008)
11. Ku, C.H., Leroy, G.: A decision support system: Automated crime report analysis and classification for e-government. Government Information Quarterly 31(4), 534–544 (2014)
12. Kumar, A.S., Gopal, R.K.: Data mining based crime investigation systems: Taxonomy and relevance. In: Communication Technologies (GCCT), 2015 Global Conference on. pp. 850–853. IEEE (2015)
13. Lee, S., Kim, H.j.: News keyword extraction for topic tracking. In: Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on. vol. 2, pp. 554–559. IEEE (2008)
14. Nath, S.V.: Crime data mining. In: Advances and Innovations in Systems, Computing Sciences and Software Engineering, pp. 405–409. Springer (2007)
15. Observatorio Nacional Ciudadano Seguridad, Justicia y Legalidad: Reporte sobre delitos de alto impacto Junio 2016. Reporte Año 3, No. 5, México (2016)
16. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, Istanbul, Turkey (May 2012)
17. Pinheiro, V., Furtado, V., Pequeno, T., Nogueira, D.: Natural language processing based on semantic inferentialism for extracting crime information from text. In: Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on. pp. 19–24. IEEE (2010)
18. Quinteiro-González, J.M., Martel-Jordán, E., Hernández-Morera, P., Ligero-Fleitas, J.A., López-Rodriguez, A.: Clasificación de textos en lenguaje natural usando la wikipedia. Iberian Journal of Information Systems and Technologies (8), 39–52 (2011)
19. Shinyama, Y., Sekine, S.: Named entity discovery using comparable news articles. In: Proceedings of the 20th international conference on Computational Linguistics. p. 848. Association for Computational Linguistics (2004)
20. Torres-Moreno, J.M.: Automatic Text Summarization. Wiley and Sons (2014)

21. Torres-Moreno, J.M., Gordon, M.: Efficient adaptive learning for classification tasks with binary units. Neural Computation 10(4), 1007–1030 (1998)
22. Torres-Moreno, J.M., Sierra, G., Peinl, P.: A German Corpus for Similarity Detection Tasks. International Journal of Computational Linguistics and Applications 2(5), 9–22 (2014)
23. Wang, X., Gerber, M.S., Brown, D.E.: Automatic crime prediction using events extracted from twitter posts. In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. pp. 231–238. Springer (2012)