ISSN 1870-4069

Propuesta de un modelo de análisis de textos para la identificación de posibles autores de mensajes criminales

Alberto Ochoa-Zezzatti[1, 2], Guadalupe Gutiérrez[2], Jorge Ramírez[2], Nathalie González[2], Marco Álvarez[2], Alberto Hernández[3], Alhelí Román[4]

[1] Universidad Autónoma de Ciudad Juárez, México

[2] Universidad Politécnica de Aguascalientes, México

[3] Instituto Nacional de Electricidad y Energías Limpias, México

[4] FCAeI-Universidad Autónoma del Estado de Morelos, México

alberto.ochoa@uacj.mx, {guadalupe.gutierrez, jorge.ramirez, up160648, marco.alvarez}@upa.edu.mx, jose.hernandez@uaem.mx

Resumen. Uno de los aspectos más desconcertantes de la violencia contemporánea en México es que parece relativamente desprovisto de discurso. A diferencia de, por ejemplo, el terrorismo religioso o nacionalista, la violencia en México no emana de una beligerancia ideológica o política. En los medios, predomina la evidencia material del conflicto: cantidad de cadáveres, tipos de proyectiles, casas de seguridad, etc. Periodistas, fiscales y expertos en seguridad son los responsables de interpretar esa evidencia y colocarla dentro de una narrativa estándar protagonizada por "cárteles". En los últimos años, las organizaciones dedicadas al tráfico de drogas se han destacado por su violencia y brutalidad. Una de las características más comunes en los ataques cometidos son los "narco mensajes". En este trabajo se realiza un análisis de narco mensajes encontrados en mantas, redes sociales y otras bases de datos aplicando minería de datos, con el fin de proponer un modelo geo-espacial por medio del cual sea posible la identificación y distribución geográfica de los autores de los mensajes.

Palabras clave: análisis de emociones, narco mensajes, discurso narrativo, minería de textos.

Proposal of a Text Analysis Model for Identifying Possible Authors of Criminal Messages

Abstract. One of the most disconcerting aspects of contemporary violence in Mexico is that it appears relatively devoid of discourse. Unlike, for example,

religious or nationalist terrorism, violence in Mexico does not emanate from an ideological or political belligerence. In the media, the material evidence of the conflict predominates-number of corpses, types of projectiles, security houses, etc. Journalists, public prosecutors and security experts are responsible for interpreting that evidence and placing it within a standard narrative starring "cartels." In recent years, organizations dedicated to drug trafficking have been noted for their violence and brutality. One of the most common characteristics in the attacks committed are the "narcomensajes". In this work we perform an analysis of narco messages found in blankets, social networks and other databases applying data mining with WEKA, in order to propose a spatial geo model that contributes to the identification and geographic distribution of the authors of the messages.

Keywords: emotional analysis, narco messages, narrative discourse, text mining.

1. Introducción

En este tipo de explicaciones, predominan los motivos estrictamente económicos o delictivos: principalmente el control de rutas y plazas, y el castigo de la deserción o la traición. El carácter precario y fragmentario del discurso público de los narcotraficantes -así como la preponderancia de las narrativas policiales ha ocultado la dimensión estrictamente política de la violencia "criminal" en México. En términos pragmáticos, el crimen organizado y la política son más similares de lo que nos gustaría suponer. Tienen en común el objetivo de dominar territorios, recursos y poblaciones; ambos tienden a ponerse de pie como un sistema de "intermediación parasitaria". Tanto las mafias como el estado ofrecen "protección" a cambio del pago de honorarios, recompensan la lealtad y castigan la traición. Son los actos discursivos que acompañan a la violencia y la serie de procedimientos institucionales en los que están registrados, que nos permiten trazar el límite entre lo político y lo criminal, lo legítimo y lo ilegítimo, lo justo y lo injusto. En México, esa frontera ha perdido claridad.

Los organismos del gobierno municipal y estatal han utilizado grupos delictivos para imponer el control político, y se ha registrado la circulación de empleados entre la policía municipal y los grupos arma-dos privados. Asimismo, en los últimos años ha habido una participación creciente de miembros o ex miembros del crimen organizado en la política electoral. Pero hay otra dimensión, tal vez más sutil, de este enfoque que tiene que ver con la dificultad que tiene el Estado para establecer y defender lo que, en principio, lo distingue de otros grupos armados. La dificultad de trazar discursivamente la frontera entre el crimen y la política.

Esta pérdida de autoridad implica que la serie de actos discursivos que constituyen la práctica diaria del Estado -desde la concesión de una licencia de conducir hasta que se resuelve una investigación judicial- han ido perdiendo eficacia lingüística: capacidad de afectar el mundo. Implica que las instancias gubernamentales encuentran cada vez más problemas cuando tratan de establecerse como fuentes fidedignas. A estos factores se agregó, alrededor de 2006, un cambio notable: los presuntos miembros de organizaciones criminales comenzaron a participar directamente en los espacios públicos regionales y nacionales, algo que hasta ese momento había sucedido muy raramente [5]. Lo hicieron con mantas atadas a cadáveres, llamadas tele-fónicas a los

medios, entrevistas, videos, comentarios en foros de Internet, confesiones anónimas y ceremonias públicas de arrepentimiento. Difícilmente se puede decir que hay una serie de demandas políticas específicas para el tráfico de drogas, como lo fue, por ejemplo, la lucha contra la extradición en Colombia. Tampoco parece existir una narrativa social o ideológica coherente y general que encuadre, defina o dé sentido al sufrimiento. No hay, por ejemplo, un discurso que permita que el dolor se convierta en un sacrificio orientado hacia el logro de un bien mayor, ya que no garantiza la supervivencia de la siguiente generación: "Me involucré en esto que mis hijos no se tengan que matar la espalda trabajando". No es suficiente formar un sujeto político como tal, un" nosotros "bien definido con sus propias demandas como en [2].

Aun así, en las expresiones públicas esporádicas y de alguna manera, infructuosas del narcotráfico; es posible delinear algo que va más allá de lo estrictamente económico o criminal y que sugiere las dimensiones ideológicas y políticas del conflicto. En este artículo se analiza un tipo particular de expresión: mensajes escritos en pedazos de tela o cartón que comenzaron a aparecer con frecuencia en las vías públicas alrededor de 2006. Los narco-mensajes son casi más medios que el mensaje: su forma va más allá del significado y su contenido [3]. En primer lugar porque muchos derivaron su visibilidad pública y fuerza discursiva del hecho de aparecer físicamente asociados con un cadáver. No solo es el contexto en el que se encuentra la manta, sino también su forma. La gran mayoría están escritos con aerosol, con abundantes errores ortográficos, insultos y declaraciones ininteligibles. La excepción a esta regla han sido las mantas de las organizaciones criminales de Michoacán, específicamente La Familia y Los Caballeros Templarios, que solían estar escritas de una manera muy intimidatoria para sus enemigos.

2. Metodología

Como resultado de la necesidad de contribuir a mejorar la seguridad en México, los métodos automatizados para analizar el contenido de los mensajes e identificar a los autores potenciales son cada vez más esenciales [4]. En este contexto, esta investigación busca hacer una contribución a la PFP (Policía Federal) para la identificación de posibles autores de estos crímenes mediante el análisis del contenido de los mensajes que utilizan el procesamiento del lenguaje natural y las técnicas de IA.

El problema existente se debe en gran parte a las siguientes características:

- Recursos humanos insuficientes.
- Gran cantidad de información disponible (características del mensaje).
- No hay un mecanismo articulado de automatización.

Los seres humanos son seres de hábito y patrones únicos de comportamiento, lo que nos lleva a conjeturar que ciertas características (palabras comunes, errores ortográficos o firmas de autógrafos, entre otros) serán constantes. Por lo tanto, se determinará el tipo de mensaje y el enfoque de la contribución de su impacto: categorización del autor. Con base en la semántica del mismo es posible determinar el propósito del mensaje (amenazar, reclamar territorio, venganza, entre otros). Por lo cual es importante analizar, diseñar e implementar un mecanismo para la PFP, con la finalidad de apoyar a la identificación de la distribución geográfica de los autores de mensajes utilizando

técnicas como el procesamiento de lenguaje natural y la minería de textos. Para ello es necesario obtener un conjunto de características a partir de una serie de mensajes para su análisis, ordenar un conjunto de mensajes por medio de similitudes para asignarlos a un autor específico y posteriormente generar mapas de distribución donde operan esos autores. Con esto la PFP tendrá un modelo que le permitirá automatizar eficientemente el análisis del contenido de los mensajes, así como identificar a los autores y agruparlos geográficamente.

3. Análisis de texto y geo-localización

Para el análisis de texto y geolocalización se propone un modelo de tres fases, las cuales se describen enseguida.

Fase 1: Generación de repositorio.

- Se obtiene una muestra de 100 mensajes de narco y se analizan con técnicas de minería de datos social.
- Se realiza un procesamiento de imágenes (Selección de imágenes con texto legible) para determinar la ubicación de éstas en un mapa de la ciudad.
- Se aplican OCR (reconocimiento óptico de caracteres) con Matlab para convertir una imagen textual digitalizada en un documento de texto.

Fase 2: Identificación de la autoría utilizando la herramienta Weka.

- En esta fase se incorporan los grupos criminales en el repositorio con la finalidad de extender el corpus del mensaje. A través de medidas de similitud (p. Ej., Distancia de Mahalanobis) en los mensajes, se identificará al posible "grupo criminal" del mismo, generando aglomeraciones.

Fase 3: Generación de mapas de distribución de grupos criminales para determinar escenarios similares de robo-violencia.

- Los datos obtenidos y almacenados en el repositorio se procesarán utilizando el lenguaje R para determinar la frecuencia de los elementos relacionados con el mismo grupo delictivo.
- Los resultados serán discutidos a la luz de estudios similares, para proponer una política pública social para apoyar a las personas que requieren más protección.

4. Aplicación de herramientas

Se usó la herramienta de minería de datos sociales WEKA [7] para analizar los datos del corpus, con base al siguiente proceso, primero se desarrolló un modelo que permite explicar el comportamiento de tres muestras de personas, y cómo afecta su estilo de discurso relacionado con los narco mensajes en narcomantas. Entre los resultados obtenidos con WEKA se descubre una relación existente entre los parámetros de hipóstasis y parataxis, utilizados por los diferentes lectores de este mensaje, los hablantes se comunican con el grupo Criminal [8].

Tabla 1. Distribuciones de demandas por categoría y ordenadas por tres muestras analizadas.

	Muestra 1	**Muestra 2**	**Muestra 3**
Lenguaje	**Inflencia**	**Desafíos**	**Amenazas**
N	212	190	185
Imperativos	12%	36%	26%
Declaraciones de directivas	5%	6%	7%
Directivas de simulación	11%	4%	5%
Directivas de interrogativas	2%	0%	1%
Postcripts de interrogativas	35%	16%	28%
Directiva conjunta	15%	3%	11%
Preguntas explosivas	2%	11%	4%
Preguntas de información	16%	22%	17%
Mecanismos de atracción de la atención	2%	2%	1%
Total	100%	100%	100%

Asimismo, se encuentra en ambos casos que los usuarios y lectores de estos mensajes mostraron una mayor hipóstasis y parataxis más baja con respecto a los hablantes de español. Esto se puede explicar por el uso del habla informal de personas relacionadas con "narcocorridos", canciones relacionadas con grupos delictivos, porque intentan asimilarse más fácilmente a personas con antepasados comunes (varias personas en este grupo delictivo son hablantes nativos del misma región en México), y la decisión de compra está muy influenciada por la comunidad de idioma.

5. Análisis de texto y geo-localización

Se considera una muestra de 587 segmentos de mensajes (212 mensajes de Influencia, 190 mensajes de Desafíos y 185 Mensajes de Amenazas) relacionados con grupos delictivos recuperados en los tres últimos años, conformados por tres muestras (muestra 1 con mensajes de influencia, muestra 2 con mensajes de desafíos y muestra 3 con mensajes de amenazas), así como conversaciones en redes sociales, para identificar diferentes comportamientos (ver Tabla 1).

El uso de minería de datos en aspectos sociales ha demostrado ser parte clave para corroborar las tendencias lingüísticas de un grupo establecido dentro de una red social común, no obstante, es posible encontrar ciertas variaciones dependiendo de la intención del mensaje y el recurso lingüístico utilizado en diferentes idiomas, ver la Tabla 2.

Finalmente, con la ubicación de cada narcomensaje, se propone un modelo geo espacial para representar cada escenario y determinar las situaciones futuras relacionadas con este tipo de grupos delictivos. En la figura 1, se presenta este modelo en un mapa de Cuernavaca en México.

Tabla 2. Contribuciones realizadas al discurso por una red social de acuerdo con diferentes palabras, se incluyen los giros utilizados por el lenguaje.

Volume of Speech			
Tipo de mensaje	**Palabras emitidas**	**Repeticiones**	**Promedio de palabras a su vez**
Influencia	788	127	5.9
Desafíos	567	93	6.1
	492	88	4.2

Fig. 1. Modelo geo espacial con las ubicaciones de cada narcomensaje en Cuernavaca y los lugares donde es más específico que se produzca un nuevo mensaje.

6. Conclusiones

Hay una cantidad importante de preguntas que merecen una investigación adicional. Uno de ellas sería encontrar nuevas fuentes de información sobre el uso de estos tres idiomas y otras ciudades con problemas similares de situaciones criminales como Ciudad Victoria en Tamaulipas (en la cual en un rango de 720 días tuvo poco más de 100 narcomantas) [9]. Un área con gran potencial es el uso electrónico de medios, específicamente, música digital [1]. En [6] se muestra un sistema que aprende de las preferencias del usuario en función de la música escuchada, después de que las canciones se seleccionan para jugar en un entorno físico compartido, basado en las preferencias de todas las personas presentes, este software tiene un guion narrativo para realizar recomendaciones a otros usuarios en un texto libre [10].

Agradecimientos. Queremos agradecer a la Procuraduría General de la República por su apoyo para evaluar la Minería de Datos Sociales como parte de este tipo de análisis multivariable, así como por permitir el uso de Bases de Datos relacionadas con este tipo de situaciones criminales y la emulación de este experimento social de aislamiento.

Referencias

1. Terveen, L., McMackin, J., Amento, B., Hill, W.: Specifying preferences based on user history. In: Proceedings of the (SIGCHI) conference on Human factors in computing systems, pp. 315–322 (2002)
2. Smith, M.A., Fiore, A.T.: Visualization components for persistent conversations. In: Proceedings of the (SIGCHI) conference on Human factors in computing systems, pp. 136–143 (2001)
3. Padméterakiris, A., Gyllenhaal, J., Ochoa, A.: Implementing of a Data Mining Algorithm for discovering Greek ancestors, using simetry patterns. In: Central Asia CCBR (Data Mining Workshop) (2005)
4. Pitkow, J. et al.: Life, death, and lawfulness on the electronic frontier. In: Proceedings of the Conference on Human Factors in Computing Systems, (CHI '97), pp. 383–390 (1997)
5. Tabrizi-Nouri, H., Tañón, O., Ianevski, S., Ochoa, A.: Explain mixtured couples support with Gini Coeficient. In: CACCBR (Data Mining Workshop) (2005)
6. Terveen, L., Hill, W.: Beyond recommender systems: Helping people help each other. HCI in the New Millennium, pp. 1487–509 (2001)
7. Frank, E., Hall, M.A., Witten, H.I.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Fourth Edition (2016)
8. Winograd, T.: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests (1997)
9. Minqing, H., Bing L.: Department of Computer Science, University of Illinois at Chicago, pp. 60607–7053.
10. Okaa de Vel, K.: Information Technology Division Defense Science and Technology Organization (2016)