Propuesta recibida en octubre de 2018 y aceptada para publicación en noviembre de 2019.

Extracción y análisis de las causas de suicidio a través de marcadores lingüísticos en reportes periodísticos

Extraction and analysis of suicide causes through linguistic markers in news reports

José A. Reyes-Ortiz
Universidad Autónoma Metropolitana
jaro@azc.uam.mx

Mireya Tovar
Benemérita Universidad Autónoma de Puebla
mtovar@cs.buap.mx

Resumen

El análisis automático de información(textos) sobre el suicidio se ha convertido en un reto para el campo de investigación en lingüística computacional, cada vez más, son necesarias herramientas que ayuden a disminuir las tasas de suicidios, por ejemplo, extraer las causas para apoyar en su detección temprana. Los aspectos lingüísticos en los textos en Español, tales como frases clave o partes de la oración, pueden ayudar en dicha tarea. Por ello, en este artículo se presenta un enfoque computacional para la extracción y análisis de causas a partir de cabeceras de reportes periodísticos sobre el suicidio en español. La tarea de extracción automática de causas de suicidio es llevada a cabo mediante marcadores lingüísticos basados en verbos, conectores, preposiciones y conjunciones. Por su parte, el análisis de las causas de suicidio es realizado en dos enfoques: a) un análisis centrado en frases verbales y nominales, estudiando la presencia de la negación; b) un análisis centrado en la frecuencia de los unigramas y bigramas de palabras. Ambos análisis muestran resultados prometedores, los cuales son útiles para conocer los motivos de los suicidios reportados en México en un periodo determinado. Finalmente, se obtiene una colección de 581 causas del suicidio.

Palabras clave

análisis de causas, suicidio en reportes periodísticos, patrones lingüísticos, lingüística computacional

Abstract

The automatic analysis of suicide data(texts) has become a challenge for the computational linguistics research field, increasingly, tools are needed to help reduce suicide rates, for example, by extracting the suicide causes in order to support their early detection. Linguistic aspects in Spanish texts, such as cue phrases or parts of speech, can help in this task. Therefore, this paper presents a computational approach to the extraction and analysis of suicide causes from news reports in Spanish. The automatic extraction of suicide causes is carried out through linguistic markers based on verbs, connectors, prepositions and conjunctions. On the other hand, the analysis of the suicides causes is performed in two approaches: a) an analysis focused on verbal and noun phrases, studying the presence of the negation; b) an analysis on the frequency about unigrams or bigrams of words. Both analyzes show promising and correlated results, which are useful for recognizing the suicide causes reported in Mexico in a given period. Finally, a corpus is obtained with a collection of 581 suicide causes.

Keywords

cause analysis, suicide in news reports, linguistic patterns, computational linguistics

1 Introducción

El suicidio es definido por la Organización Mundial de la Salud (OMS), como un acto de quitarse deliberadamente la propia vida, el cual es iniciado y realizado por una persona en pleno conocimiento o expectativa de su desenlace fatal. En México, el Instituto Nacional de Estadística y Geografía (INEGI, 2015) define el suicidio como "la acción de matarse a sí mismo". Los suicidios son un problema presente en la sociedad mexicana, donde se suscitaron 5.2 suicidios por cada 100 mil habitantes en 2015 según el INEGI (2015). En ese sentido, se estima que el suicidio ocupa una de las primeras diez causas de muerte en México.

El suicidio se caracteriza como una muerte violenta o traumática (Hernández-Bringas & Flores-Arenales, 2011), el cual tiene una causa, motivo, razón o justificación. Esta causa resulta ser una de las características más importantes del evento, ya que proporciona información sobre su origen, que al extraerla podemos realizar un análisis con la finalidad de prevenir este evento.

La prevención del suicidio es un problema social muy importante ya que según Omer & Elitzur (2001) se necesitan esfuerzos en conjun-

DOI: 10.21814/lm.11.2.276

to entre organizaciones y personas para reunir la información suficiente con la cual caracterizarlo, por ejemplo sus causas. Por ello, el extraer y analizar las causas del suicidio se convierte en un reto importante que sería de gran ayuda para los analistas de noticias en dos aspectos: a) disminuyen los tiempos invertidos en la tarea tediosa de análisis manual de los textos; b) los analistas tienen conocimiento de las causas de los suicidios de manera automática para la toma de decisiones. Los textos de cabeceras de reportes periodísticos son una fuente importante para recabar dicha información. Estos textos resultan de gran utilidad ya que mediante señas lingüísticas relacionados a las causas del suicidio, se puede conocer el conjunto de ellas para conducir acciones hacia la prevención del mismo (Pestian et al., 2012). Esto se debe a que los periodistas reportan, entre otras cosas, las causas de haberse cometido un suicidio. La idea es contar con herramientas computacionales que apoyen a los analistas de noticias a realizar su actividad de manera rápida y contar con un apoyo en la detección temprana y prevención del suicidio. El problema radica en que existe una carencia de herramientas y recursos para el tratamiento de textos de suicidio en español, aunado a que resulta complicado tener acceso a una base de notas suicidas. Pero, es posible contar con los reportes periodísticos en línea, de los cuales sus cabeceras son de acceso público y pueden ser extraídas con facilidad a partir de la Web.

Los reportes periodísticos se han convertido en una fuente de datos muy poderosa, ya que nos brinda datos frescos sobre lo que está aconteciendo en una región o país y con una temporalidad determinada. En este artículo se presenta un enfoque de tratamiento automático de textos en español a partir de cabeceras de reportes periodísticos con la finalidad de detectar y analizar causas de suicidios en México haciendo uso de patrones lingüísticos. Este enfoque inicia con el reconocimiento de cabeceras de reportes periodísticos en español que traten sobre suicidio, utilizando el método presentado por Reyes-Ortiz & Bravo (2018); después, se extraen las causas, de manera automática, utilizando marcadores lingüísticos formados por verbos, preposiciones, conjunciones y conectores. Por último, tres enfoques de análisis de las causas son presentados: un análisis a nivel de frases tanto verbales como nominales, un análisis del impacto de la negación en las frases verbales y un análisis centrado en frecuencia de palabras. El conjunto de cabeceras de reportes periodísticos utilizado en este artículo se compone de la siguiente manera: a) como conjunto inicial, se utilizan 9 574 cabeceras para la tarea de clasificación; b) a partir del conjunto inicial, se identifican 1 347 pertenecientes a la categoría de suicidio; c) 581 causas son extraídas para su análisis a partir de las cabeceras sobre suicidio. Los resultados del análisis en términos de frecuencias de las palabras o frases son analizados mediante una nube de palabras para encontrar las causas más frecuentes y obtener la correlación de los resultados entre el enfoque de palabras y frases. Como producto final, un corpus de una colección de causas de suicidio es construido.

Las principales contribuciones de este artículo, se pueden resumir de la siguiente manera: a) la extracción automática de causas en las cabeceras de suicidios utilizando marcadores lingüísticos; b) el análisis de causas del suicidio usando dos enfoques: frecuencias de palabras y frecuencias de frases verbales y nominales; c) la creación de un corpus de causas de suicidio. Además, el enfoque propuesto resulta de gran utilidad para los analistas de noticias, apoyándolos en la tarea de recopilación de noticias sobre el suicidio y el análisis de sus causas, mediante su frecuencia, ya sea formadas por una palabra, un par de palabras, frases nominales o frases verbales.

El resto del artículo se organiza de la siguiente manera. En la Sección 2 se muestra el estado del arte relacionado a tres temas importantes: extracción automática de causas a partir de textos, aplicaciones de patrones lingüísticos para la extracción de información en diversos dominios y la extracción de cualquier tipo de información relacionada al suicidio. En la Sección 3 se expone la caracterización del suicidio y sus causas, así como la metodología de solución propuesta para la extracción y análisis de las causas. La Sección 4 presenta los marcadores lingüísticos utilizados para la extracción de causas en suicidios. La Sección 5 conduce un análisis de causas del suicidio a partir de los textos de cabeceras de reportes periodísticos. Finalmente, las conclusiones y el trabajo a futuro son presentados en la Sección 6.

2 Trabajo relacionado

La tarea de Extracción de Información es una subdisciplina de la Lingüística Computacional que consiste en identificar elementos o entidades de interés a partir de textos. Los patrones lingüísticos para esta tarea han sido utilizados como se describe a continuación. En el trabajo presentado por González-Gallardo et al. (2016) se hace uso de patrones sintácticos para la normalización de textos multilingüe extraídos de redes sociales con la finalidad de perfilar autores. La extracción de definiciones analíticas y relaciones semánticas de hiponimia-hiperonimia con

un sistema basado en patrones lingüísticos construidos manualmente es presentada por Dorantes et al. (2017). Los contextos definitorios son definidos por Sierra (2009) como un término y su definición introducida en el discurso de un texto de especialidad, en dicho trabajo se extraen estos contextos de manera automática con el apoyo del reconocimiento de patrones lingüísticos. La recuperación de patrones a partir de textos es una tarea que en el trabajo de Da Cunha et al. (2009) es abordada mediante un proceso de aprendizaje automático con la finalidad de generar resúmenes de textos especializados, es decir, de dominio específico. Roberto Rodríguez et al. (2013) presenta dos tipos de patrones (morfosintácticos y léxicos) para la clasificación automática de textos en registros lingüísticos en español, es decir, información sobre el perfil de los usuarios y sobre el contexto en sistemas de recomendación. Los patrones lingüísticos también pueden estar enfocados en el análisis automático de sentimientos en redes sociales mediante algoritmos de clasificación usando características lingüísticas de las partículas de los textos como las conjunciones ('but') y los condicionales ('if') (Chikersal et al., 2015) o bien, analizando las relaciones entre los conceptos usando patrones lingüísticos para obtener el tipo de sentimiento o polaridad en una sentencia (Poria et al., 2015). Adicionalmente, el uso de patrones lingüísticos es utilizado por Bertin et al. (2016) para la identificación de contextos de citas discerniendo las citas negativas.

La identificación automática de la causalidad a partir de textos se ha abordado desde un punto de vista de eventos. Los eventos tienen características como sus causas y efectos, aspectos que son identificados por Borsje et al. (2010), específicamente, para eventos financieros usando patrones semánticos y con la finalidad de enriquecer una ontología de dominio. En el dominio biomédico, un enfoque para la identificación automática de marcadores discursivos de causalidad usando aprendizaje automático con características semánticas a partir de textos biomédicos es presentado por Mihăilă & Ananiadou (2013). En el trabajo de Kang et al. (2017), se detecta las características causales haciendo uso de series de tiempo entre los *n-gramas*, temas, sentimientos y su composición extraídos a partir de textos. La clasificación de emociones es una tarea dentro del análisis de sentimientos, la cual es abordada desde un punto de vista lingüístico, por Li & Xu (2014), mediante la extracción de causas de eventos basada en patrones para ayudar en la clasificación de emociones como felicidad, tristeza, ira, sorpresa y disgusto, logrando resultados prometedores.

La idea de la prevención del suicidio se ha abordado como un análisis de información textual relacionada a los suicidios. De esta manera, diversos trabajos han abordado la extracción de información sobre suicidios a partir de textos como fuente de datos. Existen trabajos que han analizado las notas clínicas para predecir los riesgos de suicidios en los pacientes, como en el trabajo de Poulin et al. (2014) que utiliza un algoritmo de aprendizaje automático basado en programación genética para llevar a cabo esta tarea a partir de notas clínicas en inglés. El análisis de notas suicidas mediante técnicas de minería de sentimientos es un tema que ayuda en la prevención del suicidio, en la cual se han detectado trabajos que identifican de manera automática emociones (culpa, felicidad, agradecimiento, amor, información, desesperanza e instrucciones) en estas notas usando características de los textos con algoritmos de aprendizaje automático tales como: máquinas de soporte vectorial (Desmet & Hoste, 2013; Luyckx et al., 2012), campos aleatorios condicionales (Liakata et al., 2012) y un clasificador de máxima entropía (Wicentowski & Sydes, 2012). El aspecto lingüístico en el suicidio es de gran importancia. En esta línea, un análisis lingüístico de notas suicidas y textos sobre el suicidio aplicado al inglés y adaptado para el español ha sido presentado por Fernández-Cabana et al. (2015) para comparar las características sociodemográficas y forenses de ejemplos de víctimas de suicidio a partir de las notas suicidas dejadas por ellos, dicho estudio se realizó con notas suicidas en español comparando sus características de género, edad y nivel social. Los autores obtuvieron resultados sobre las características estudiadas y las diferentes frases lingüísticas utilizadas en las notas suicidas, tales como: longitud de las frases, uso de tiempos verbales, uso de pronombres y verbos. Finalmente, un estudio similar de Stirman & Pennebaker (2001) enfoca en el dominio de la poesía.

Con el estudio de los diversos temas y trabajos revisados en el estado del arte, se hace evidente que la investigación dirigida hacia el análisis lingüístico de la causalidad de los suicidios a partir de textos en español es un problema a resolver, existiendo trabajos como el de Reyes-Ortiz & Bravo (2018); Cook et al. (2016) que solo se enfocan en la clasificación del suicidio en español usando patrones o técnicas de aprendizaje supervisado. Esto muestra una carencia de herramientas y enfoques computacionales que utilicen técnicas de la Lingüística Computacional para el análisis de textos sobre las causas del suicidio en español. Este estudio, también, expone que la mayoría de los trabajos están enfocados

en el idioma inglés (Sawhney et al., 2018; Carson et al., 2019; Leiva & Freire, 2017) o en dos idiomas español-inglés (Cook et al., 2016), originando una necesidad creciente de contar con recursos de análisis de textos de causas del suicidio. Esto abre una ventana de desafíos y retos para llevar a cabo procesamiento automático de textos en español. Además, algunos trabajos revisados (Reyes-Ortiz & Bravo, 2018; Sawhney et al., 2018; Wicentowski & Sydes, 2012; Luyckx et al., 2012; Liakata et al., 2012; Poulin et al., 2014; Carson et al., 2019), se enfocan, solamente, en la identificación o clasificación del suicidio, mientras que nuestro trabajo añade la extracción y análisis de las causas del mismo. Por lo tanto, además de aportar una solución al problema de la extracción automática de causas del suicidio, brinda un panorama sobre los marcadores lingüísticos causales que son característicos de este tipo de textos en español y almacena las causas del suicidio reportadas en notas periodísticas.

3 Caracterización del suicidio y la causalidad

El suicidio como causa de muerte se encuentra dentro de la categoría de muertes violentas, ya que según Hernández-Bringas & Flores-Arenales (2011) se trata de muerte traumática, producidas por medios externos al organismo humano. Por otro lado, el Instituto Nacional de Estadística y Geografía (SSP & INEGI, 2012) ha definido al suicidio en México como "un evento que implica una conducta en la que una persona se priva de la vida por sí misma, involucrando sólo la intervención de una persona suicida", que se describe como:

- Suicida.
  Es alguien que se suicida por sí mismo.

Por lo tanto, un suicidio se caracteriza como un evento monovalente, es decir, que únicamente tiene un actor, el suicida. En la Figura 1 se muestra como se caracteriza un suicidio en el contexto del esquema actancial de la teoría de las valencias de eventos de Tesniére (1976). Donde "alguien" representa al actor que comete el suicidio.

Una causa expresa el argumento o la justificación, la cual es responsable de que suceda un evento o acción que Born (1949) formaliza como "la ocurrencia de una entidad B de cierta clase depende de la ocurrencia de una entidad A de otra clase", donde la entidad puede ser cualquier objeto físico, fenómeno, situación o evento. La Figura 2 muestra la representación de la causa de un evento suicida.

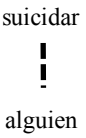

Figura 1: Esquema actancial de un evento suicida

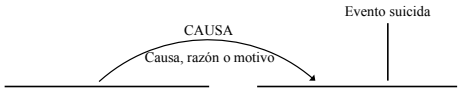

Figura 2: Caracterización de causalidad en eventos suicidas

Considerando la definición según Born (1949), un evento suicida tiene una micro-situación que expresa la causa. En este trabajo, se considera esta situación de causalidad para su análisis a partir de textos de cabeceras de reportes periodísticos en español.

En el contexto de datos textuales, la causalidad se manifiesta mediante una variedad de expresiones lingüísticas. Por lo tanto, en este artículo se aborda la extracción y el análisis de causas en eventos suicidas para el español. Este análisis se basa en un estudio sobre las construcciones de causas en las cuales intervienen marcadores lingüísticos y categorías gramaticales como, verbos, preposiciones y conjunciones. Para ello, se utilizan técnicas de la Lingüística Computacional para hacer posible el análisis automático de los textos en español relacionados al suicidio. La metodología de solución propuesta para la extracción y análisis de causas del suicidio en textos en español, es presentada en la Figura 3.

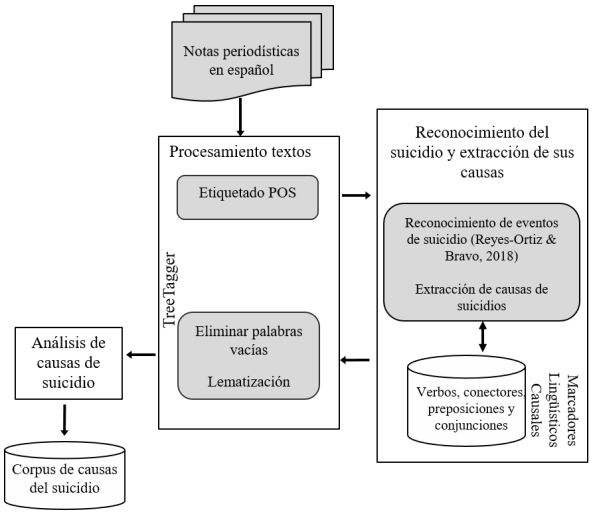

Figura 3: Metodología de solución para la extracción y análisis de causas del suicidio

Como se puede apreciar en la Figura 3 la metodología propuesta consiste de dos grandes etapas: extracción de causas del suicidio y el análisis estadístico de las mismas en México. Estas etapas son detalladas en las próximas secciones.

4 Extracción de causas del suicidio usando marcadores lingüísticos causales

En esta sección se presenta la etapa de la metodología correspondiente a la extracción automática de causas del suicidio utilizando expresiones lingüísticas tomadas de (Reyes-Ortiz et al., 2017) para reconocer las causas de suicidios a partir de cabeceras de reportes periodísticos en español. Es importante hacer notar que en (Reyes-Ortiz et al., 2017) se presentan marcadores lingüísticos para diversos tipos de eventos, sin embargo, en este trabajo se validan para el caso de eventos sobre el suicidio.

4.1 Marcadores lingüísticos causales

Un marcador lingüístico es una expresión formada por una o más palabras que tiene la función de conectar a los eventos con su argumento o justificación. En (Reyes-Ortiz et al., 2017) se presentan marcadores lingüísticos para identificar causas de cualquier evento, los cuales son llamados marcadores lingüísticos causales. Por ello, es que en este artículo, tomamos dichos marcadores y los aplicamos a los eventos de suicidios. Ellos están conformados por verbos causales, conectores conjuntivos, locuciones, preposiciones o conjunciones causales.

Para este trabajo se tomaron los marcadores lingüísticos causales presentados por Reyes-Ortiz et al. (2017) y se contrastaron con las construcciones lingüísticas presentadas en (Cano, 1981; Funes, 2010; Wunderlich, 1997). Además se consideran conectores causales obtenidos de la Real Académica de la Lengua Española (Española, 2009) para formar el siguiente conjunto de marcadores lingüísticos, los cuales serán utilizados para la extracción automática de causas del suicidio.

1. **Los verbos causales**. También llamados verbos causativos implícitos, verbos de carácter puramente causal o causativo propios, verbos básicamente causativos. Estos verbos están compuestos de una frase verbal y poseen un significado intrínsecamente causativo. La forma básica y representativa es el verbo *causar* y sus derivados como *provocar, originar, motivar, suscitar, desencadenar, promover, determinar, ocasionar, acarrear, producir, incitar, infundir, obrar* y *generar*. Su significado causal está supuesto por su forma semántica.
2. **Conectores causales**. Las causas también pueden estar representadas por conectores lingüísticos como: oraciones causales coordinadas que contiene nexos conjuntivos mediante los siguientes vocablos y locuciones *pues, pues que, porque, puesto que, a causa de, por esta razón, por eso, por ello, por esto, de esta manera, por lo cual, por lo que, debido a*; oraciones causales subordinadas, sus nexos conjuntivos son los vocablos: *de que, ya que*.
3. **Preposiciones causativas**. Las preposiciones son partículas lingüísticas que aportan gran significado a una oración. En el contexto de la causalidad, las preposiciones *tras*, *por* y *de* pueden expresar una causa. Según Funes (2010), esto se debe a que dichas preposiciones relacionan los eventos con su origen o argumentación.
4. **Conjunción causales**. Las conjunciones son partículas lingüísticas que funcionan como nexos entre segmentos de textos, las cuales pueden denotar una causalidad, como la conjunción *porque*, *pues* y *como*.

Estos marcadores lingüísticos son utilizados para la extracción automática de causas en suicidios a partir de textos en español de cabeceras de reportes periodísticos, con la finalidad de, posteriormente, llevar a cabo un análisis de las causas más frecuentes y detectar elementos relevantes.

4.2 Extracción automática de causas de suicidio

El proceso de extracción automática de causas del suicidio se compone de dos tareas: el reconocimiento de eventos reportados en cabeceras de reportes periodísticos centradas en suicidios y la extracción automática de causas basada en los marcadores lingüísticos presentados anteriormente. El contar con un conjunto de cabeceras de reportes periodísticos que contengan eventos de suicidio para la extracción de sus causas se vuelve indispensable. Para ello, se utiliza el enfoque presentado por Reyes-Ortiz & Bravo (2018) para el reconocimiento de eventos relacionados con la seguridad reportados en cabeceras periodísticas, entre ellos se encuentra el suicidio. Dicho enfoque utiliza patrones enriquecidos con información lingüística para reconocer y clasificar una cabecera de nota periodística, entre los patrones utilizados para la tarea de clasificación de una cabecera

en la categoría de suicidio se encuentran los siguientes: *X se quita la vida*, *X se suicida*, *X se ahorca*. Esta tarea considera un total de 9574 cabeceras y como resultado se obtienen 1 347 cabeceras de notas periodísticas que abordan el tema del suicidio.

Al conjunto de cabeceras filtradas (1 347) por el evento relacionado al suicidio se aplica el proceso de etiquetado POS y posteriormente, se aplican la regla mostrada en la Figura 4 con la finalidad de extraer causas en suicidios a partir de estos textos. Esta regla está basada en la gramática de JAPE (Cunningham & Tablan, 1999) que es una Máquina de Anotación de Patrones en Java basada en expresiones regulares. Esta regla hace uso de los marcadores lingüísticos causales presentados anteriormente. La frase que viene después del marcador y que denota la causa del suicidio (cs) puede ser un sintagma verbal (SV) o un sintagma nominal (SN). El etiquetado de partes de la oración denominado (POS) realizado con la herramienta denominada *TreeTagger* (Schmid, 1995), es necesario para el reconocimiento de sintagmas verbales y sintagmas nominales. El sintagma verbal es una palabra o grupo de palabras que constituyen una unidad sintáctica cuyo núcleo es un verbo. El sintagma nominal es una palabra o grupo de palabras que tiene como núcleo a un sustantivo.

```
Rule: ExtracciónCausaSuicidio
(EventoSuicidio)
(MarcadorLingüísticoCausal)
(SV|SN):cs -->
: cs.ExtracciónCausaSuicidio = {rule = "ExtracciónCausaSuicidio"}
```

donde cs expresa la causa del suicidio, SV expresa un sintagma verbal y SN es sintagma nominal.

Figura 4: Regla JAPE para extraer causas en suicidios

Esta tarea obtiene como resultado un corpus de 581 causas compuestas por sintagmas verbales y sintagmas nominales, donde el marcador lingüístico más representativo son los verbos causales.

El corpus obtenido es utilizado para un análisis de las causas en suicidios de la siguiente manera. Por un lado, los sintagmas se dejan en su forma original para hacer un análisis a nivel de frases. Mientras que, un procesamiento es aplicado al conjunto de sintagmas verbales y nominales que representan el argumento o justificación de los suicidios, con la finalidad de llevar a cabo un análisis preciso a nivel de palabras o entradas léxicas. El objetivo es extraer los términos (palabras o frases) relacionados al suicidio.

El procesamiento de las frases consiste en las siguientes tareas:

- Eliminar palabras vacías. El objetivo de esta tarea es quitar aquellas palabras que no aportan un significado en las frases que denotan las causas de suicidios, estas palabras están compuestas, principalmente, por artículos o determinantes (*el, las, los un, unos*), preposiciones (*a, ante, bajo, cabe, con, contra, de*), conjunciones (*y, o*) y algunos verbos, como el *ser/estar*, *tiene* que si bien tienen una frecuencia alta en las frases, no aportan relevancia en el estudio de causas. Aun cuando algunas preposiciones y conjunciones fueron útiles para extraer causas de suicidio, en esta etapa se suprimen para un análisis a nivel de términos relevantes como sustantivos, verbos o adjetivos.
- Lematización. Esta tarea tiene como objetivo normalizar las palabras resultantes contenidas en las causas para agrupar las palabras que provienen de la misma raíz. Este proceso consiste en obtener el lema correspondiente de cada palabra, eliminando tiempos verbales y conjugaciones en el caso de verbos, llevándolos a su forma en infinitivo. En el caso de sustantivos, adjetivos y adverbios, el lema corresponde a su forma en masculino singular. Por ejemplo, las formas *corrieron*, *corrió*, *correrán* es transformado a la forma verbal *correr*. La lematización de las causas obtenidas se ha llevado a cabo mediante el etiquetador *TreeTagger* (Schmid, 1995), el cual ha sido exitosamente utilizado para lematizar textos en Español. Esta tarea es indispensable para aplicar la regla propuesta sobre los textos de las cabecearas lematizados ya que los verbos causales están en su forma infinitivo.

Finalmente, después de la etapa de procesamiento de las frases, se crea una nube de palabras a partir de los términos o unidades léxicas resultantes. Esta nube de palabras ayuda a extraer con claridad las causas del suicidio a partir de los datos utilizados. Para generar esta nube de palabras fue considerada la frecuencia de aparición de cada palabra con el objetivo de descartar u otorgar menor importancia a las frases generadas por el significado polisémico de las preposiciones *por* y *de*, que en ocasiones no tienen un significado causativo.

La tarea de extracción automática de causas del suicidio ha demostrado que el tipo de mar-

cador lingüístico causal más representativo para esta tarea son los verbos causales en su forma de infinitivo. De esta manera, estos verbos causales caracterizan los textos de suicidios.

5 Análisis de causas del suicidio

En esta sección se presenta el análisis de causas extraídas con los marcadores lingüísticos presentados previamente, describiendo el conjunto de datos utilizado para este análisis. Primero, se utilizan los sintagmas nominales y verbales en su forma original con el objetivo de llevar a cabo un análisis a nivel de frases. Después, los componentes léxicos (palabras procesadas) de los sintagmas son utilizados para un análisis a nivel de una nube de palabras. El objetivo de esta nube de palabras es visualizar las causas de suicidios más frecuentes reportadas por periódicos mexicanos en un lapso de tiempo determinado.

El conjunto de datos utilizado para este estudio consiste en las 1 347 cabeceras de reportes periodísticos en español relacionadas con el suicidio, las cuales son extraídos de las cuentas en la red social Twitter de los principales periódicos y páginas en México que informan sobre noticias relacionadas con la seguridad, tales como: El Universal (*@El_Universal_Mx*), Milenio (*@Milenio*), Reforma (*@Reforma*), Excelsior (*@Excelsior*), Secretaria de Seguridad Publica de México(*@SSP_CDMX*), La Jornada (*@lajornadaonline*) y Noticias de Google México (*@googlenewsmx*). Las cabeceras de reportes periodísticos pertenecen al periodo de enero de 2017 a septiembre de 2018, las cuales son extraídas de manera automática con la *API* de Java denominada Twitter4J (Yamamoto, 2008). Esta herramienta ha permitido el filtrado de los mensajes por los parámetros de ubicación e idioma que nos ha permitido centrarnos en reportes periodísticos generados en la república mexicana y escritos en español.

A partir de estas 1 347 cabeceras de reportes periodísticos sobre suicidios, se lograron extraer 581 frases nominales y verbales, mediante la aplicación de los marcadores lingüísticos presentados. Entonces, se conservan esas frases para un primer análisis y después se eliminan palabras vacías y se lematizan para generar la nube de palabras.

5.1 Análisis de causas de suicidio centrado en frases

El primer análisis centrado en frases es realizado con los sintagmas nominales y verbales extraídas en su forma original. Se lleva a cabo una clasificación de frases en verbales y nominales cuyo resultado se muestra en la Tabla 1.

	Cantidad de frases	**Porcentaje**
Verbal	303	52.2 %
Nominal	278	47.8 %

Tabla 1: Resultados de la clasificación de frases.

Este análisis muestra una superioridad no muy marcada en la presencia de frases verbales para las causas de los suicidios. Las cinco frases verbales más frecuentes en las causas de suicidio son:

1. *ser víctima de violencia intrafamiliar*
2. *ser víctima de bullying*
3. *ser abusadas*
4. *ser separado de su familia*
5. *ser acusado de*

Es importante notar que la frase verbal *ser acusado de* tiene algunas variantes, entre las que destacan: *ser acusado de acoso sexual* y *ser acusado de violación*.

Las cinco frases nominales más frecuentes en las causas de suicidio son:

1. *bullying*
2. *problemas psicológicos*
3. *depresión*
4. *soledad*
5. *desesperación*

En el caso de las frases verbales (FV) se ha detectado la presencia de la negación, es decir, una frase verbal antecedida por una partícula semántica de negación, constituyendo una causa formal, como: $\neg FV$. Dada esta presencia, se incluye un estudio del impacto de este tipo de frases verbales utilizando su frecuencia en el conjunto de datos. De esta manera, dos ejemplo de una frase verbal negada y que han sido extraídas como causas de suicidios son: *no quería vivir con alzheimer* y *no recibir el regalo que quería*. La Tabla 2 muestra la distribución de este tipo de frases verbales presentes en el conjunto de causas obtenidas para el suicidio.

Las causas de suicidios tiene una gran tendencia a estar en forma afirmativa. Sin embargo, las causas expresadas con una frase verbal negativa no pueden ser descuidadas en una tarea de caracterización.

	Cantidad de frases	Porcentaje
$\neg FV$	32	10.6 %
FV	271	89.4 %

Tabla 2: Resultados de la distribución de la negación en frases verbales.

5.2 Análisis de causas de suicidio centrado en palabras

Este análisis de las causas de suicidios es llevado a cabo a nivel de palabras y mediante la generación de una nube de palabras. Para generar esta nube de palabras fue considerada la frecuencia de aparición de cada palabra en las causas, con ello se visualiza su importancia en la caracterización de causas de suicidios. Las palabras fueron obtenidas de las frases tanto verbales como nominales que representan las causas de suicidios. Estas palabras fueron lematizadas, y a partir de este conjunto, se eliminaron las palabras vacías. Dos escenarios son considerados para la generación de dos nubes de palabras: a) unigramas, causas como términos formados por una palabra; b) bigramas, causas como términos formados por un par de palabras. Esto con la finalidad de analizar la frecuencia de las causas con una y dos palabras.

En el primer caso se consideran unigramas de palabras a partir de las causas de suicidios detectadas. El resultado de esta nube de palabras es mostrado en la Figura 5, donde se aprecia una clara predominancia de la causa *bullying*.

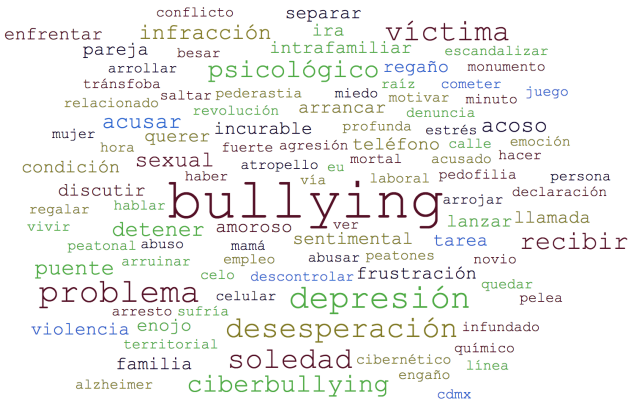

Figura 5: Nube de palabras de las causas de suicidios con unigramas

Para el segundo escenario se forman bigramas de palabras, es decir caso de causas de suicidios con número de palabras igual a dos. El resultado de esta nube de palabras es mostrado en la Figura 6.

Figura 6: Nube de palabras de las causas de suicidios con bigramas

Este análisis muestra que para el caso de los unigramas de palabras, las causas de suicidio más frecuentes son: *bullying*, *problemas*, *depresión*, *desesperación*, *soledad* y *ciberbullying*. Es importante notar que la causa *bullying* y *ciberbullying* en realidad tienen la misma naturaleza, la diferencia es que para el segundo caso, los ataques y ofensas (*bullying*) se llevan a cabo en Internet, en sitios como redes sociales, blogs o sistemas de mensajería electrónica. En el caso de los bigramas, las causas más frecuentes son: *violencia familiar* y *problemas psicológicos*.

El análisis presentado en esta sección revela que las frases verbales, frases nominales, unigramas y bigramas más frecuentes en las causas de suicidios pueden ser utilizadas como características lingüísticas para el análisis automático del suicidio a partir de cabeceras de reportes periodísticos en español. El caso de los unigramas revelan el uso de sustantivos formados por una palabra que son utilizados en las causas. Sin embargo, los bigramas serían de mayor utilidad en el caso de sustantivos compuestos, es decir, causas de suicidios expresadas con dos palabras, como es el caso de *problemas psicológicos*, que en términos de unigramas no aportarían un significado relevante a las causas. Si se desea realizar una caracterización de textos de suicidio, el uso de un solo tipo de característica no sería suficiente para abarcar todos los casos de causas de suicidios, lo más recomendable es un enfoque lingüístico que combine los diversos tipos de características.

6 Conclusiones

Este artículo ha presentado un enfoque para la extracción automática y, posterior, análisis de causas del suicidio en México utilizando los textos de cabeceras de reportes periodísticos en español correspondientes a un periodo determinado. El enfoque propuesto utiliza técnicas de la Lingüística Computacional para realizar el tra-

tamiento automático de los textos. El proceso completo consiste en una etapa de extracción automática de causas de suicidios a partir de texto en español y posteriormente, un análisis presentando estadísticas de frecuencia. El proceso de extracción utiliza marcadores lingüísticos causales basados en verbos, preposiciones, conectores y conjunciones para extraer frases verbales o nominales, negadas o afirmadas de los textos. Por su parte, el análisis es divido en dos escenarios, uno a nivel de frases y otro a nivel de palabras. Como salida se obtiene un corpus de 581 causas de suicidios extraídas y analizadas a partir de los textos.

El análisis centrado en frases verbales y nominales, muestra una ligera mayoría de frases verbales con un 52.2 % de las causas de suicidios. La presencia de la negación se ha encontrado en las frases verbales con una incidencia del 10.6 % de los casos.

El análisis centrado en palabras que se lleva a cabo mediante la formación de unigramas y bigramas de palabras exhibe cierta similitud a los resultados de frases, coincidiendo en el caso de las causas como el *bullying*, *problemas* y *depresión*.

En este artículo se presentaron las siguientes aportaciones: a) el método de extracción automática de causas del suicidio a partir de textos en español utilizando marcadores lingüísticos causales; b) el análisis de las causas de suicidio en los diversos escenarios: a nivel palabras o frases; c) la construcción del corpus con 581 causas de suicidios. Estas aportaciones disminuyen la carencia de recursos de análisis para textos en español, además de exponer las causas más frecuentes, ya sea formadas por una palabra (unigramas), un par de palabras (bigramas), frases nominales o frases verbales, negadas o afirmadas.

Con la tarea de extracción automática de causas se demuestra que el tipo de marcador lingüístico más representativo son los verbos causales, que fueron utilizados para extraer las frases verbales, frases nominales, unigramas y bigramas más frecuentes en las causas. Por lo tanto, estos marcadores se pueden utilizar para caracterizar textos sobre suicidios en español en notas periodísticas.

Aun cuando el análisis presentado es para un periodo determinado y utilizando cabeceras de reportes periodísticos, es de gran ayuda para los analistas de noticias, ya que tienen conocimiento de las causas del suicidio en México con la ayuda de un enfoque computacional, el cual es de gran utilidad al reducir el tiempo invertido en el análisis de noticias y reducir el esfuerzo humano. Con estos resultados los analistas pueden tomar decisiones como enfocar políticas de prevención del suicidio y conocer datos estadísticos sobre las causas del suicidio para evitarlas.

Como trabajo a futuro para este artículo, resulta enriquecedor extender el enfoque a periodos de tiempo mayores con reportes periodísticos completos de diversas fuentes en español. Además, la creación de un sistema informático para visualizar y consultar las causas del suicidio en México, sería de gran utilidad para los analistas de reportes periodísticos.

Agradecimientos

Este artículo ha sido apoyado por la Universidad Autónoma Metropolitana, unidad Azcapotzalco con el proyecto de investigación SI001-18. Los autores agradecen, también, a la Benemérita Universidad Autónoma de Puebla por el apoyo recibido y al CONACyT bajo el proyecto 257357.

Referencias

Bertin, Marc, Iana Atanassova, Cassidy R. Sugimoto & Vincent Lariviere. 2016. The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics* 109(3). 1417–1434. 10.1007/s11192-016-2134-8.

Born, Max. 1949. *Natural philosophy of cause and chance*. The Clarendon Press primary source ed.

Borsje, Jethro, Frederik Hogenboom & Flavius Frasincar. 2010. Semi-automatic financial events discovery based on lexico-semantic patterns. *International Journal of Web Engineering and Technology* 6(2). 115–140. 10.1504/IJWET.2010.038242.

Cano, Rafael. 1981. *Estructuras sintácticas transitivas en el español actual*, vol. 310. Gredos 1ª ed.

Carson, Nicholas, Brian Mullin, Maria Jose Sanchez, Frederick Lu, Kelly Yang, Michelle Menezes & Benjamin. Lê Cook. 2019. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. *PloS one* 14(2). e0211116. 10.1371/journal.pone.0211116.

Chikersal, Prerna, Soujanya Poria, Erik Cambria, Alexander Gelbukh & Chng Eng Siong.

2015. Modelling public sentiment in Twitter: using linguistic patterns to enhance supervised learning. En A. Gelbukh (ed.), *International Conference on Intelligent Text Processing and Computational Linguistics*, 49–65. Springer International Publishing. 10.1007/978-3-319-18117-2_4.

Cook, Benjamin L., Ana M. Progovac, Pei Chen, Brian Mullin, Sherry Hou & Enrique Baca-Garcia. 2016. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Computational and mathematical methods in medicine* 2016. 1–8. 10.1155/2016/8708434.

Cunningham, Diana, Hamish Maynard & Valentin Tablan. 1999. JAPE: a java annotation patterns engine.

Da Cunha, Iria, Juan Manuel Torres-Moreno, Patricia Velázquez-Morales & Jorge Vivaldi. 2009. Un algoritmo lingüístico-estadístico para resumen automático de textos especializados. *Linguamática* 1(2). 67–79.

Desmet, Bart & Véronique Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications* 40(16). 6351–6358. 10.1016/j.eswa.2013.05.050.

Dorantes, Miguel Alejandro, Alejandro Pimentel, Gerardo Sierra, Gemma Bel-Enguix & Claudio Molina. 2017. Extracción automática de definiciones analíticas y relaciones semánticas de hiponimia-hiperonimia con un sistema basado en patrones lingüísticos. *Linguamática* 9(2). 33–44. 10.21814/lm.9.2.257.

Española, Real Academia. 2009. *Nueva gramática de la lengua española*, vol. 2. Espasa Calpe 2ª ed.

Fernández-Cabana, Mercedes, Julio Jiménez-Féliz, María Teresa Alves-Pérez, Raimundo Mateos, Ignacio Gómez-Reino Rodríguez & Alejandro García-Caballero. 2015. Linguistic analysis of suicide notes in spain. *The European Journal of Psychiatry* 29(2). 45–155. 10.4321/S0213-61632015000200006.

Funes, Maria Soledad. 2010. La alternancia de las preposiciones por y de en las construcciones causales. *Revista de Estudios Hispanos* 1(1). 5–14.

González-Gallardo, Carlos, Juan-Manuel Torres-Moreno, Azucena Montes-Rendón & Gerardo Sierra. 2016. Perfilado de autor multilingüe en redes sociales a partir de n-gramas de caracteres y de etiquetas gramaticales. *Linguamática* 8(1). 21–29.

Hernández-Bringas, Héctor Hiram & René Flores-Arenales. 2011. El suicidio en méxico. *Papeles de población* 17(68). 69–101.

INEGI. 2015. Estadísticas de mortalidad. Accedido 05-09-2018. http://www.beta.inegi.org.mx/proyectos/registros/vitales/mortalidad/default.html.

Kang, Dongyeop, Varun Gangal, Ang Lu, Zheng Chen & Eduard Hovy. 2017. Detecting and explaining causes from text for a time series event. En *Conference on Empirical Methods in Natural Language Processing*, 2758–2767. 10.18653/v1/D17-1292.

Leiva, Victor & Ana Freire. 2017. Towards suicide prevention: Early detection of depression on social media. En *International Conference on Internet Science*, 428–436. 10.1007/978-3-319-70284-1_34.

Li, Weiyuan & Hua Xu. 2014. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications* 41(4). 1742–1749. 10.1016/j.eswa.2013.08.073.

Liakata, Maria, Jee-Hyub Kim, Shyamasree Saha, Janna Hastings & Dietrich Rebholz-Schuhmann. 2012. Three hybrid classifiers for the detection of emotions in suicide notes. *Biomedical informatics insights* 5. BII–S8967. 10.4137/BII.S8967.

Luyckx, Kim, Frederik Vaassen, Claudia Peersman & Walter Daelemans. 2012. Fine-grained emotion detection in suicide notes: a thresholding approach to multi-label classification. *Biomedical informatics insights* 5. BII–S8966. 10.4137/BII.S8966.

Mihăilă, Claudiu & Sophia Ananiadou. 2013. Recognising discourse causality triggers in the biomedical domain. *Journal of Bioinformatics and Computational Biology* 11(6). 1343008 (15 pages). 10.1142/S0219720013430087.

Omer, Haim & Avshalom C. Elitzur. 2001. What would you say to the person on the roof? a suicide prevention text. *Suicide and Life-Threatening Behavior* 31(2). 129–139. 10.1521/suli.31.2.129.21509.

Pestian, John P., Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K. Bretonnel Cohen, John Hurdle & Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights* 5. BII–S9042. 10.4137/BII.S9042.

Poria, Soujanya, Erik Cambria, Alexander Gelbukh, Federica Bisio & Amir Hussain. 2015. Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine* 10(4). 26–36. 10.1109/MCI.2015.2471215.

Poulin, Chris, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley V.. Watts, Laura A. Flashman & Thomas W.. McAllister. 2014. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one* 9(1). e85733. 10.1371/journal.pone.0085733.

Reyes-Ortiz, José A. & Maricela Bravo. 2018. Enhancing patterns with linguistic information for criminal event recognition. *Journal of Intelligent and Fuzzy Systems* 34(5). 3027–3036. 10.3233/JIFS-169487.

Reyes-Ortiz, José. A., Maricela Bravo, Azecena Montes & Mireya Tovar. 2017. Event ontology enrichment with causal relations from spanish text. *International Journal of Computational Linguistics and Applications* 8(1). 1–16.

Roberto Rodríguez, John, Maria Salamó Llorente & Maria Antònia Martí Antonín. 2013. Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo. *Linguamática* 5(1). 59–67.

Sawhney, Ramit, Prachi Manchanda, Raj Singh & Swati Aggarwal. 2018. A computational approach to feature extraction for identification of suicidal ideation in tweets. En *ACL 2018, Student Research Workshop*, 91–98. 10.18653/v1/P18-3013.

Schmid, Helmut. 1995. Treetagger: a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung* 43. 1–28.

Sierra, Gerardo. 2009. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática* 1(2). 13–37.

SSP & INEGI. 2012. Clasificación estadística del delito en méxico. Accedido 05-09-2018. http://www3.inegi.org.mx/sistemas/clasificaciones/delitos.aspx.

Stirman, Shannon Wiltsey & James W. Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine* 63(4). 517–522. 10.1097/00006842-200107000-00001.

Tesniére, Lucien. 1976. *Eléments de syntaxe structurelle*, vol. 2. Klincksieck 2ª ed.

Wicentowski, Richard & Matthew R. Sydes. 2012. Emotion detection in suicide notes using maximum entropy classification. *Biomedical informatics insights* 5. BII–S8972. 10.4137/BII.S8972.

Wunderlich, Dieter. 1997. Cause and the structure or verbs. *Linguistic Inquiry* 28(1). 27–68.

Yamamoto, Yusuke. 2008. Twitter4j. [Web; accedido el 19-10-2018]. http://twitter4j.org/en/index.