Universidad Autónoma del Estado de México

PROTOCOLO DE INVESTIGACIÓN

PROYECTO:

Implementación de una nube multimodal dinámica de palabras para visualización de emociones basada en SenticNet: Caso de estudio Contexto Político Mexicano

Alumno: Keren Mitsue Ramírez Vergara

Asesor: Dr. Asdrúbal López Chau

Modalidad: Licenciatura

Participación: Individual

Área de Conocimiento: Físico-matemático

Ingeniería en Computación

Octubre 2022

Contenido

Introducción

Una de las primeras etapas en el proceso de análisis de sentimientos, consiste en un estudio exploratorio de los documentos, para ello, la nube de palabras es ampliamente usada como una herramienta para la visualización resumida de los documentos. Una de las principales desventajas de las nubes de palabras comunes, es que no permiten visualizar los cambios en la polaridad de las opiniones durante el tiempo; adicionando a lo anterior, estas nubes no consideran teorías emocionales para una mejor comprensión de los documentos.

En este proyecto se propone un método para la visualización de publicaciones en la red social Twitter, mediante una nube dinámica de palabras multimodal, que, a diferencia de las nubes de palabras tradicionales, considera la evolución en el tiempo y la multi polaridad emocional de las palabras. Se consideran tres factores para el diseño e implementación de la nube dinámica de palabras: la frecuencia de uso de palabras, las polaridades emocionales de las palabras y el tiempo. El modelo usado para extraer las polaridades se basará en SenticNet, uno de los frameworks recientes más importantes para realizar análisis de sentimientos.

Definiciones preliminares

Las nubes de palabras suelen presentarse a modo de figura abstracta, en las que son representadas de un mayor tamaño aquellas palabras que aparecen con más frecuencia, o son más importantes [8]. Para definir formalmente una nube dinámica de palabras, se presentan los siguientes términos:

- **C:** Conjunto de documentos sobre algún tema (corpus).
- **S**: Subconjunto de documentos que pertenecen al corpus.
- **S':** Conjunto de palabras perteneciente a S. Este conjunto se obtiene de pre-procesar los documentos existentes en S.
- **Nube de palabras** $\boldsymbol{N(C')}$: Es una representación gráfica de la frecuencia de uso de cada palabra en un documento d. El tamaño de una palabra es proporcional al número de veces que se repite la misma en el conjunto $\boldsymbol{C'}$
- **Documento fechado** $\boldsymbol{d_t}$: Es un documento (texto) que ha sido publicado en cierto instante t En este tipo de documentos, tanto el tiempo (fecha y hora) de publicación como el contenido, son características muy importantes. Las publicaciones en redes sociales (como Facebook) y en servicios de microblogging (como Twitter) son ejemplos de este tipo de documentos.

Basado en las anteriores definiciones, se presentan ahora las siguientes:

- **Nube dinámica de palabras** $\boldsymbol{ND_{ta}^{tb}(C') = \left\{N_{t_a}^{t_{a+1}}(S'), N_{t_{a+1}}^{t_{a+2}}(S'), \dots, N_{t_{b-1}}^{t_b}(S')\right\}}$: Es una sucesión de nubes de palabras de documentos fechados. Esta nube permite visualizar la evolución temporal de las frecuencias de uso de las palabras en documentos fechados.
- **Nube dinámica de palabras basada en SenticNet** $\boldsymbol{NDSenticNet_{ta}^{tb}(S')}$: Es una nube dinámica de palabras en la cual las frecuencias calculadas son ponderadas con las correspondientes puntuaciones de polaridad.

Objetivo general

Diseñar e implementar una Nube Dinámica de Palabras para ser usada como una herramienta para la visualización de la evolución de la frecuencia de uso de palabras, y sus polaridades emocionales respecto al tiempo, mediante el modelo sentimental del marco de trabajo SenticNet.

Objetivos Específicos

1. Definir formalmente el concepto de Nube Dinámica de Palabras, en el que se consideran la frecuencia de las palabras, la temporalidad de los documentos que las contienen y su polaridad emocional.
2. Presentar una metodología para la búsqueda y recolección sistemática de documentos en la plataforma de Twitter.
3. Aplicar la metodología del objetivo anterior para recolectar documentos recientes relacionados con el contexto político mexicano.

Marco teórico

Análisis de sentimientos

El análisis de sentimientos es un campo de la investigación interdisciplinar que permite identificar y modelar emociones, interpretar estados de ánimo, opiniones y tendencias en textos, imágenes, vídeos, lenguaje corporal, etc. Con el objetivo de comprender el comportamiento humano, teniendo en cuenta la emoción y la cognición en el diseño de tecnologías relacionadas para satisfacer las necesidades humanas. El reconocimiento de la información emocional requiere la extracción de muestras significativas de los datos recogidos mediante técnicas de aprendizaje automático, que procesan los datos mediante el reconocimiento de la voz el reconocimiento de patrones, el procesamiento del lenguaje natural o la detección de expresiones faciales [5,6]. Las redes sociales son una fuente importante de información para la extracción de datos con el fin de detectar patrones emocionales mediante de procesamiento del lenguaje natural [7].

Nube de palabras

Una nube de palabras o nube de etiquetas es una representación gráfica de un texto o conjunto de textos con propiedades visuales como el tamaño de las palabras indica la frecuencia de aparición. Debido a la fácil comprensibilidad de la nube de palabras, es usada en diferentes ámbitos, tales como el político, el de salud, educación y el económico. En ámbitos políticos se utiliza esta herramienta para obtener de manera rápida la idea principal de los temas tratados en un comunicado. [8]

SenticNet

SenticNet es un sistema semántico y recurso público para el análisis de sentimiento a nivel de concepto, gracias a su formato compatible con la web semántica es muy fácil

interconectar el recurso con cualquier aplicación del mundo real que necesite extraer la semántica y el lenguaje natural.

SenticNet cuenta con una representación del modelo inspirado en el cerebro y el análisis de emociones humanas denominada "The Hourglass of Emotions" (Reloj de arena de emociones). [4]. Véase en la figura 1. Los resultados en SenticNet se proporcionan como un vector séntico que especifica el agrado, la atención, la sensibilidad y la aptitud asociados al concepto, además de un valor de polaridad, un estado de ánimo primario y secundario. Los valores de polaridad y los vectores sénticos son útiles para la recuperación y la detección de la polaridad. La polaridad puede ser un número en un rango fijo o simplemente un indicador (positivo/negativo)

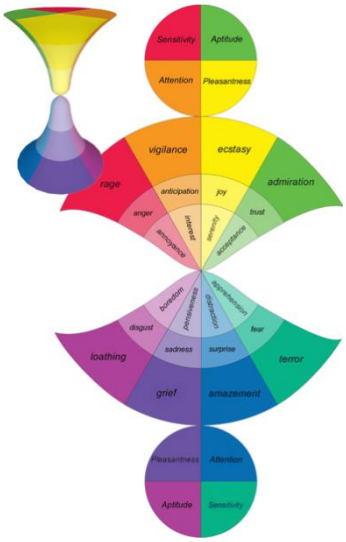

Figura 1. The Hourglass of Emotions. Fuente: sentic.net

Metodología de la investigación

La Figura 2 muestra un resumen gráfico de la metodología empleada, misma que a continuación se explica detalladamente.

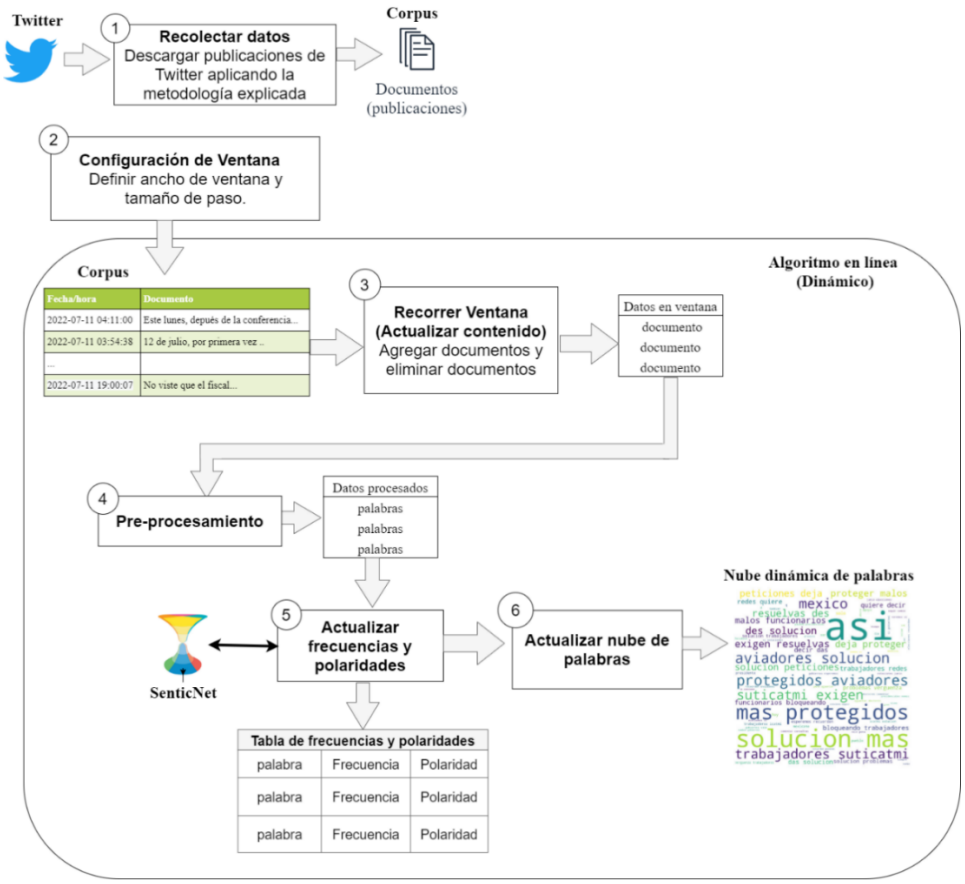

Figura 2. Resumen de la metodología aplicada. Fuente: Elaboración propia.

Primer paso: Recopilación de datos

La fuente de datos seleccionada para la recopilación de los datos es la red social Twitter, además se utiliza la biblioteca Tweepy. Para realizar la colección de datos con Tweepy, se propone el siguiente proceso:

1. Comenzar con una lista de términos de búsqueda (hashtags y usuarios) relacionados con el tema, especificando el periodo y el lugar para recolectar datos.
2. Buscar para cada elemento de la lista en Twitter y descargar los documentos.
3. Aplicar un muestreo aleatorio simple a los documentos para elegir el tamaño adecuado del subconjunto.
   El tamaño del muestreo se calcula con la siguiente fórmula:

$$n' = \frac{n}{1 + \frac{z^2 p(1-p)}{\varepsilon^2 N}}$$

   Donde:

- $n = \frac{z^2 p(1-p)}{\varepsilon^2}$
- $z = 1.644854$
- $P: porción\ de\ población\ (P = 0.6)$
- $\varepsilon: margen\ error\ (\varepsilon = 10\%)$

- $N = tamaño\ de\ población\ (número\ de\ tweets\ descargados)$

4. Leer los tweets de la muestra y seleccionar aquellos que sean relevantes para el estudio.
5. Agregar a la lista las menciones y hashtags que aparecen en los archivos seleccionados, y que aún no están en la lista.
6. Repetir los pasos del 2 al 5 tres veces.
7. Usar la lista ampliada de hashtags y usuarios creada anteriormente para encontrar datos relevantes y descargue los documentos. Este conjunto de documentos se le denomina Corpus.

Segundo paso: Configuración de la ventana

Con el objetivo de observar la evolución de la frecuencia de uso de palabras en el corpus y su polaridad, se propone generar nubes de palabras sucesivas a partir de subconjuntos del corpus. Cada subconjunto contiene documentos publicados entre dos tiempos (fecha y hora), a los que llamaremos t_a y t_b , donde:

- t_a tiempo de inicio de ventana: Los documentos publicados a partir del tiempo t_a y hasta $t_b$son considerados para generar la nube de palabras.
- t_b tiempo de fin de ventana: Los documentos publicados hasta el tiempo t_b pero desde t_a son considerados para generar la nube de palabras.

El ancho de ventana de se define como: $W = |t_b - t_a|$

Dado un corpus con la estructura siguiente: $C = \{(t_i, d_i): t_i\ tiempo\ , d_i\ texto, i = 1, \dots, N\}$

El subconjunto $S_m \subseteq C$, es el conjunto de documentos que pertenecen a la ventana m, está definido como: $S_m = \{d_i: d_i \in C\ , t_a \leq t_i \leq t_b\}\ m = 0,1,2, \dots$

Donde:

$$t_a = t_{1+m \times step}$$
$$t_b = t_a + W$$
$$step \subseteq \mathbb{Z}^+: desplazamiento\ de\ la\ ventana\ en\ el\ tiempo$$

Los parámetros W y $step$ son valores que se configuran manualmente por el usuario y deben de cumplir con la condición $W > step$

Tercer paso: Recorrer o desplazar la ventana

Una vez establecido el ancho de la ventana (W) y el tamaño del paso (step), se comienza con un proceso en línea de actualización de frecuencias de palabras y sus polaridades, con los elementos del conjunto actual S_m. Este último se genera cuando la ventana avanza (o retrocede) un paso. La Figura 3 ejemplifica el desplazamiento de la ventana en un paso.

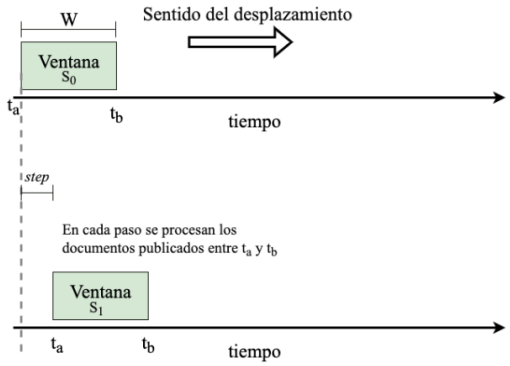

Figura 3. Desplazamiento de ventana. Fuente: Elaboración propia.

Cuarto paso: Pre-procesamiento de documentos (textos)

El procesamiento de documentos consiste fundamentalmente en una serie de transformaciones que se realiza al documento (texto), con el objetivo de obtener datos uniformes y sin elementos que puedan afectar a la calidad de los resultados. Por lo que, para cada documento fechado (d_t) del subconjunto S_m se realiza lo siguiente:

1. Eliminar caracteres especiales: Considerando que P es un conjunto de caracteres especiales $P =$ { | , “ ,’ ,$,% , & ,/ ,(,) ,= ,? ,¿,* ,- ,+ ,: , ; ,_ ,< ,> ,[,] , \ , ^ , © , ® , ~, ˁ }, se realiza la operación mostrada en la ecuación (1)

   $$d_t' = d_t - P \text{ --- (1)}$$

   Donde, d_t' : es el documento sin símbolos especiales.

2. Tokenizar: Se utiliza para dividir una frase en palabras, frases, símbolos u otros tokens significativos mediante la eliminación de los signos de puntuación [9], por lo que el documento modificado d_t' se separa cómo se observa en la ecuación (2)

   $$T_{d_t'} = \{< palabra >_1, < palabra >_2, \dots, < palabra >_k\} \text{ --- (2)}$$

   Donde, $T_{d_t'}$: Conjunto de palabras o también llamado conjunto de tokens

3. Eliminación de palabras: Consiste en remover las palabras que son usuarios, hashtags, urls, y palabras vacías (stopwords en inglés). El conjunto de palabras es modificado como se muestra en la ecuación (3):

   $$R_{d_t'} = \left(T_{d_t'} - U\right) \cup \left(T_{d_t'} - H\right) \cup \left(T_{d_t'} - Y\right) \text{ --- (3)}$$

   Donde:
   $R_{d_t'}: conjunto\ de\ palabras\ relevantes$
   $U = \{u |\ u \in usuarios\ \}$
   $H = \{h\ |h\ \in hashtags\}$
   $Y = \{y\ |y\ \in stopwords\}$

Finalmente, al conjunto de palabras del S_m , es expresado como se muestra en la ecuación (4):

$$S_m{}' = \left\{ \left(R_{d_t'}\right)|_{i=1}^{N} \; R_{d_t'} \; \in C \right\} \text{ --- (4)}$$

Es importante mencionar que los documentos que se encuentran en la ventana (W) deben de pasar por un pre-procesamiento. Ya que cuando los documentos del subconjunto S_m sean pre-procesados, se obtendrá un conjunto de palabras $S_m{}'$.

Quinto paso: Actualización de frecuencias y polaridades

Consiste en obtener de cada palabra del conjunto $S_m{}'$ una frecuencia y una polaridad, es decir, obtener una estructura numérica de la palabra, para el algoritmo implementado pueda comprender y generar nubes de palabras de acuerdo con las emociones del modelo de SenticNet. En este proceso existen tres aspectos a considerar:

1. Para generar las nubes de palabras emocionales sólo se consideran las palabras que cuentan con una emoción proporcionada por SenticNet, por lo que inicialmente el conjunto $S_m{}'$ es reducido.
2. Si $m = 0$ en el conjunto $S_m{}'$:
   Se obtienen todas las frecuencias y polaridades del conjunto $S_m{}'$
3. Si $m \neq 0$ en el conjunto $S_m{}'$:
   Se actualizan las frecuencias y polaridades del conjunto $S_m{}'$

Para el cálculo de las polaridades se utilizó la propiedad 'polarity_value' del concepto de la palabra que proporciona SenticNet, este valor va de -1 a 1, siendo los valores negativos una polaridad negativa y los valores positivos una polaridad positiva. Así mismo, el cálculo de las frecuencias absolutas da información acerca de las veces que se repite una palabra en la ventana actual $(S_m{}')$. Sea:

$$E = \{e | e \; \in modelo \; sentimental \; de \; SenticNet \; \}$$

Para cada e^i se obtiene un conjunto de tuplas, que contendrán la información respectiva de las palabras ubicadas en la ventana m, su estructura es la siguiente (ecuación (6)):

$$D_i(W) = \left\{ \left(w_j^i, f_j^i, p_j^i \right) \; \in \; e^i \right\} \text{ --- (5)}$$

Donde:

W: ancho de la ventana
e^i: sentimiento i-ésimo del conjunto E
w_j^i: palabra j de la ventana m correspondiente al sentimiento e^i
p_j^i: polaridad asignada a la palabra w_j^i de la ventana m correspondiente a e^i
f_j^i: frecuencia asignada a la palabra w_j^i de la ventana m correspondiente al sentimiento e^i

Conforme transcurre el tiempo, la ventana m va incrementando, por lo que las frecuencias y las polaridades de $D_i(W)$ se deben actualizar, es decir, existirán palabras que serán removidas o añadidas a la ventana. La figura 4 ejemplifica está actualización.

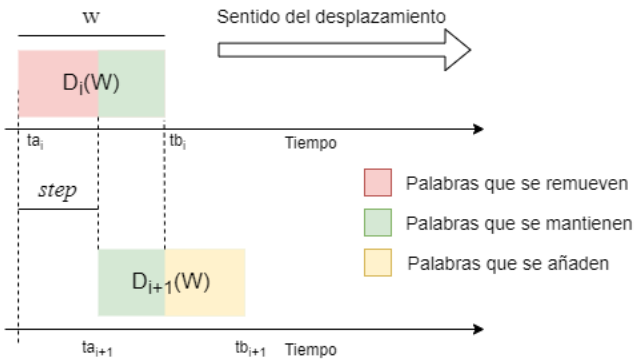

Figura 4. Actualización de palabras. Fuente: Elaboración propia.

Como se observa en la figura 4, las palabras w_j^i del conjunto $D_i(W)$ que se encuentran entre el periodo $|ta_{i+1} - ta_i|$ son las palabras que serán removidas de la ventana m. Por lo que las frecuencias de dichas palabras son disminuidas; matemáticamente se expresa:

$$f_2 = f_1 - k$$

Donde:

k: $número\ de\ veces\ que\ se\ repite\ w_j^i\ en\ el\ periodo\ |ta_{i+1} - ta_i|$

f_2: $f_j^i \in D_{i+1}(W)$

f_1: $f_j^i \in D_i(W)$

Posteriormente de remover las frecuencias, se debe verificar si alguna palabra w_j^i tiene una frecuencia igual a 0 ($f_j^i = 0$), si es el caso, la tupla $\left(w_j^i, f_j^i, p_j^i\right)$ será totalmente removida del conjunto $D_{i+1}(W)$. De igual manera, las palabras w_j^i del conjunto $D_i(W)$ que se encuentran entre el periodo $|tb_{i+1} - tb_i|$ son las palabras que serán añadidas de la ventana m. Por lo que las frecuencias de dichas palabras son aumentadas; matemáticamente se expresa:

$$f_2 = f_1 + k$$

Donde:

k: $número\ de\ veces\ que\ se\ repite\ w_j^i\ en\ el\ periodo\ |ta_{i+1} - ta_i|$

f_2: $f_j^i \in D_{i+1}(W)$

f_1: $f_j^i \in D_i(W)$

Considérese que para cada palabra nueva w_j^i en $D_{i+1}(W)$, se debe calcular su polaridad.

Sexto paso: Actualizar la nube de palabras

Una vez obtenido el conjunto $D_i(W)$, cada frecuencia f_j^i para w_j^i será ponderada con su polaridad p_j^i, siguiendo la ecuación (6). El objetivo es ponderar la frecuencia de w_j^i con la intensidad emocional de la palabra para obtener una representación visual tanto en frecuencia como en polaridad.

$$fit\left(w_j^i\right) = \left|p_j^i\right| * f_j^i \quad \text{--- (6)}$$

Al final del cálculo de las ponderaciones se obtiene un conjunto de valores con la siguiente estructura:

$$FIT_i\left(D_i(W)\right) = \left\{\left(w_j^i, fit\left(w_j^i\right)\right) \in e^i\right\}$$

Donde:

$D_i(W)$: conjunto de tuplas de la ventana m correspondiente al sentimiento $e^i w_j^i$: palabra j de la ventana m correspondiente al sentimiento e^i

$fit\left(w_j^i\right)$: peso asignado a la palabra w_j^i

Finalmente, se utilizó la biblioteca wordcloud de Python para la creación de las nubes de palabras emocionales, indicando el color de las palabras basado en base el "Hourglass of emotions" de SenticNet, con los resultados de $FIT_i\left(D_i(W)\right)$ [10].

Análisis de procedimientos experimentales, computacionales y su desarrollo

En este proyecto se analizan los resultados obtenidos tras aplicar la metodología descrita en el proyecto, inicialmente se recolectan los datos utilizando el API Rest de Twitter, por lo que se han descargado 232,446 tweets del 04/08/2022 al 15/08/2022 como se ilustra en la Tabla 1.

Tabla 1. Número de Tweets recopilados por cada palabra clave. Fuente: Elaboración propia.

Palabra clave	Tweets recolectados
4T	37,767
amlo	39,578
morena	29,125
obrador	22,627
RenovaciónProfunda	13,841
Nacional	26,428
México	37,908
LigaDeGuerreros	3,756
AmloNarcoDictador	3,327
13Ago	11,496
MorenaCuevaDeDelicuentes	3,132
AmloTraidorALaPatria	1,520
ConferenciaPresidente	1,941
TOTAL	**232,446**

Resultados

En la figura 5, se muestran algunas nubes de palabras emocionales correspondientes a cada evolución. El corpus con el que se realizó está nube dinámica de palabras es el conjunto de datos de la tendencia #4T, cuyos parámetros a especificar fueron el tamaño de la ventana con 4 días y el paso de la ventana con 2 días. En total se generaron 4 evoluciones donde se identifican las emociones: {#anger, #calmness, #disgust, #eagerness, #fear, #joy, #pleasantness, #sadness}

Figura 5. **Nube dinámica de palabras basada en SenticNet del corpus #4T. Fuente: Elaboración propia.**

Con base en los resultados, se observa que las publicaciones sobre el contexto político mexicano contienen palabras negativas que predominan en la nube dinámica de palabras. Algunos ejemplos son judas, terror, mafia, mediocre y cáncer. Por otra parte, las palabras positivas que predominan son: favor, social, moral, defender y especial. Las emociones negativas predominan tanto en frecuencia como en polaridad.

Impacto y aportación del proyecto

Este proyecto tiene un impacto y aportación positiva en diferentes aspectos como el tecnológico, científico y social. A continuación, se explican cada uno de ellos.

Aspecto tecnológico

a) Servirá como herramienta tecnológica para visualizar las polaridades de documentos publicados en redes sociales de una manera novedosa.
b) La implementación de la nube dinámica de palabras está disponible en un sitio en Internet, los interesados podrán hacer uso del código fuente para usarlo, analizarlo, modificarlo o mejorarlo.

Aspecto científico

c) La implementación de la nube dinámica de palabras podrá ser descargada libremente para su utilización por la comunidad científica, tecnológica y el público en general.
d) Se propone un método sistemático para recolección de documentos en una plataforma de red social, en particular, para Twitter.
e) Desde el punto de vista de ciencia básica, en el área de físico matemático, permite expresar formalmente el concepto de Nube Dinámica de Palabras, ofreciendo nuevas posibilidades de explicación de los fenómenos sociales y políticos.
f) Se propone un método formal para generación de nubes dinámica de palabras. Las aportaciones anteriores se sometieron en una revista indizada.

Aspecto social

g) La implementación de la nube dinámica de palabras permite mejorar la comprensión de fenómenos actuales de tipo social y político del país. Aunque también puede aplicarse a otros contextos.
h) La Nube Dinámica de Palabras propuesta puede ser utilizada por los interesados en el quehacer político de México, sin necesidad de tener experiencia en el área de físico matemático o computación.

Referencias

[1] A. López-Chau, D. Valle-Cruz, and R. Sandoval-Almazán, “Sentiment Analysis of Twitter Data Through Machine Learning Techniques,” 2020, pp. 185–209. doi: 10.1007/978-3-030-33624-0_8.
[2] Mejía González Kevin, “Enriquecimiento del modelo basado en reglas vader a través de lexicones,” 2018.
[3] Cambria Erik, Olsher Daniel, and Rajagopal Dheeraj, “SenticNet 3: A common and common-sence Knowledge base for cognition-driven sentiment analysis,” pp. 2–6, 2014.
[4] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, “SenticNet 5: Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings,” pp. 1795–1802, 2018.
[5] Cambria, E. 2016. Affective computing and sentiment analysis. IEEE intelligent systems. 31, 2 (2016), 102–107.
[6] Cambria, E. et al. 2020. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. Proceedings of the 29th ACM International Conference on Information & Knowledge Management (2020), 105–114.
[7] Liu, B. 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. 5, 1 (2012), 1–167.
[8] Chávez Ortíz César David and Zbala Acosta Isabella, “Tendencias y dinámicas en los mercados de capitales en colombia: una aplicación mediante wordclouds,” 2019.
[9] Luque Sanchez Maria Alejandra and Cortés Diaz Luis Felipe, “Análisis etiquetado de textos para predicción de la polaridad, enfoque semi supervisado y etiquetado automático,” pp. 1–44, 2020.

[10] D. Valle-Cruz, V. Fernandez-Cortez, A. López-Chau, and R. Sandoval-Almazán, “Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods,” Cognitive Computation, vol. 14, no. 1, pp. 372–387, Jan. 2022, doi: 10.1007/s12559-021-09819-8.