Fecha de recepción: 30-01-2023
Fecha de aceptación: 28-02-2023
pp.: 121-140

GUAYANA MODERNA
https://revistasenlinea.saber.ucab.edu.ve/index.php/guayanamoderna/index

Choque de Tuits: Aplicación de Minería de Texto para Análisis de Sentimientos

Tweet Shock: Text Mining Application for Sentiment Analysis

Borjas Medina, Livia Carolina
UCAB Guayana-Escuela Ingeniería Informática / Ciudad Guayana, Venezuela
livacaro7@gmail.com

Resumen:

En el presente artículo se exponen modestos resultados de un proceso de análisis de sentimiento político que se realizó, aplicando minería de texto sobre las cuentas de Twitter de representantes de la "oposición" y representantes "oficialistas" afectos al gobierno Madurista en Venezuela, en el contexto de las elecciones parlamentarias venezolanas del 2015. Este estudio se ejecutó con el propósito de identificar si ambos grupos políticos representan dos clases claramente diferenciadas en su discurso lingüístico y en cuyo caso establecer y describir sus sentimientos, intereses, preocupaciones y preferencias. Para realizar este trabajo se aplicó una adaptación de la metodología CRISP-DM para datos complejos de textos conocido como proceso de minería de texto. Para el estudio se utilizaron las herramientas R, R Studio, TwitteR y WEKA. Los resultados observados en forma de modelos de clasificación, grupos y reglas de asociación involucran los términos frecuentemente usados por los tuiteros analizados, muestran que sí existen diferencias importantes en el contenido discursivo, muy modestamente se muestran los temas y sentimientos que caracterizan a los dos grupos que se disputan el poder en Venezuela. Este estudio y sus resultados evidencian la factibilidad y utilidad de aplicar minería de texto en redes sociales para realizar análisis de sentimientos en el ámbito político.

Palabras clave: Análisis de Sentimientos, Minería de Texto, Twitter, R Studio, WEKA.

Digital: ISSN 2790 5039 – Depósito legal digital: DC2021001122
Impreso: ISSN 23435658 – Depósito legal impreso: PP201202DC4092

Digital: ISSN 2790 5039 – Depósito legal digital: DC2021001122
Impreso: ISSN 23435658 – Depósito legal impreso: PP201202DC4092

Abstract:

This article presents modest results of an analysis process of political sentiment that was carried out, applying text mining on the Twitter accounts of representatives of the "opposition" and "official" representatives of the Madurista government in Venezuela, in the context of the 2015 Venezuelan parliamentary elections. This study was carried out with the purpose of identifying if both political groups represent two clearly differentiated classes in their linguistic discourse and in which case establish and describe their feelings, interests, concerns and preferences. To carry out this work, an adaptation of the CRISP-DM methodology was applied to complex text data known as the text mining process. For the study, the R, R Studio, TwitteR and WEKA tools were used. The results observed in the form of classification models, groups and association rules, involve the terms frequently used by the tweeters analyzed, show that there are important differences in the discursive content, very modestly the themes and feelings that characterize the two are shown groups vying for power in Venezuela. This study and its results demonstrate the feasibility and usefulness of applying text mining on social networks to perform sentiment analysis in the political arena.

Key Words: Sentiment Analysis, Text Mining, Twitter, R Studio, WEKA

INTRODUCCIÓN

Las redes sociales son un punto de encuentro de millones de usuarios que publican y comparten informaciones y opiniones sobre diversos temas. Estos espacios virtuales y sociales son atractivas fuentes de información donde los usuarios exponen la percepción, sentimientos y comportamiento que tienen sobre un determinado tema, acontecimiento, producto y servicio. Realizar análisis de textos sobre los comentarios que publican los usuarios, sus opiniones y sentimientos, se ha convertido en uno de los objetivos a corto plazo de las organizaciones en su área de influencia. En la última década, los esfuerzos se han centrado en la monitorización de textos largos publicados en foros o blogs. Sin embargo, dado el reciente éxito de redes de microblogging como Facebook o Twitter; las organizaciones han empezado a centrar su interés en el análisis de textos cortos. En especial, la red social Twitter es muy popular en este sentido, pues los usuarios publican sus experiencias en mensajes de hasta 140 caracteres, popularmente conocidos como tuits.

El fenómeno del uso del Twitter por líderes e influenciadores políticos y sociales, recientemente ha tenido un auge mundial. En el campo político, analizar los contenidos de los grupos políticos y sus seguidores también es una tendencia mundial que ha cobrado gran importancia por los resultados que dichos análisis han mostrado.

Un ejemplo de esto se describe en el artículo (K. Sylwester, 2015) donde se investigó las diferencias psicológicas entre individuos de las dos organizaciones políticas más importantes de los Estados Unidos, expresadas en los mensajes publicados en Twitter.

En el artículo en cuestión, se pregunta si el uso de palabras frecuentes por parte de los seguidores de los dos grandes partidos norteamericanos, los Republicanos y los Demócratas, refleja diferencias psicológicas importantes. En este tipo de análisis, a través de técnicas de minería de datos, palabras y frases son clasificadas como portando una valencia negativa o positiva. Con ello, se pueden identificar o inferir emociones, intereses y actitudes de aquellos que usan determinadas palabras. Una herramienta usada para esta clasificación es el *Linguistic Inquiry and Word Count* (LIWC), una aplicación que clasifica palabras según categorías o clases. De este modo, conforme a los intereses de este artículo, será posible correlacionar cierto tipo de palabras con rasgos de carácter (por ejemplo, la gente extrovertida o, para dar otro ejemplo, la gente religiosa, usará algunas clases de palabras con más frecuencia que otras). Con esto en mente, los autores del estudio exploran Twitter para hacer inferencias sobre las personalidades de los demócratas y los republicanos.

Después de analizar 5.373 seguidores de tres cuentas de Twitter de los partidos Demócrata y 5.386 seguidores de tres cuentas de la Organizaciones del Congreso Partidos Republicano, se confirmó la hipótesis que existen diferencias marcadas en el lenguaje utilizado por seguidores de ambos partidos políticos.

Este estudio muestra la factibilidad de realizar análisis de sentimiento a partir de los mensajes publicados en Twitter. Inspirados en este resultado, el presente artículo expone los resultados de realizar minería de texto sobre los tuits publicados en las cuentas

más representativas de los grupos políticos que protagonizan la escena política venezolana, con el objetivo general de analizar el lenguaje usado y así predecir y/o describir palabras que usan con mayor frecuencia a fin de determinar sus intereses, sentimientos, preocupaciones y los temas que estos líderes políticos venezolanos conversan.

La hipótesis del estudio parte del hecho que es sabido que en la sociedad venezolana se observa una polarización política protagonizada por dos grandes grupos: oficialistas y opositores que involucran diversos partidos políticos. La hipótesis inicial planteada en este proyecto es que los "opositores" y los "oficialistas" representan dos clases políticas claramente diferenciadas, con un uso de lenguaje distinto y dos representaciones conceptuales de la actividad política de carácter idiosincrático, lo cual constituye un hecho inédito en la historia democrática venezolana.

A diferencia del estudio de Sylwester, Purver y Nicole (K. Sylwester, 2015)**¡Error! No se encuentra el origen de la referencia.**, el objetivo de la presente investigación no será extraer inferencias psicológicas de carácter de los tuits de los usuarios de las dos grandes tendencias políticas en Venezuela sino probar tentativamente otra hipótesis: **¿existen diferencias políticas o ideológicas profundas e interesantes entre los seguidores de las dos tendencias que, en Venezuela, se disputan actualmente el poder?** Para responder esta pregunta se aplicarán las metodologías de minería de texto presentes en la literatura y se utilizarán las bondades de las herramientas R Studio y WEKA.

Por tanto sería un interesante aporte identificar si los "oficialistas" y los "opositores" constituyen dos clases claramente diferenciadas en su discurso lingüístico predominante, pudiendo con esto establecer sus intereses, preocupaciones, preferencias. Conocer esta clasificación y características distintivas facilitará la toma de decisiones asertivas, en materia política y social, basadas en las diferencias y puntos de encuentros descubiertos.

El presente artículo, presenta los detalles del estudio realizado y se organiza de la siguiente manera: una

breve Introducción, descripción del estado del arte de la investigación en el tema de minería de texto en Twitter, una breve explicación de la metodología aplicada, seguido de los resultados, análisis y conclusiones.

ESTADO DEL ARTE

De acuerdo con (A. Bifef, 2010) se han realizado dos grandes tipos de investigación sobre bases de datos provenientes de Twitter: Minería de grafos, que se apoya en el descubrimiento de vínculos entre los mensajes y Minería de textos, basada en el análisis de los mensajes propiamente dichos.

1.1 Minería de grafos

La minería de grafos puede usarse para examinar una serie de problemas interesantes. En primer lugar se puede usar para medir la influencia de un usuario y la popularidad de sus tuits. En este sentido, existirían tres tipos de influencia: la que se evidencia al interior de la cuenta por su número de seguidores (indegree), la que se evidencia a través de los retuits de los mensajes de un usuario y la que se evidencia por las menciones. Las tres pueden estar o no correlacionadas. En segundo lugar, se puede descubrir la creación y formación de comunidades. Este caso es diferente del primero, dado que un usuario que no es influyente pudiera formar parte de una comunidad que lo es o puede llegar a serlo. Este tipo de análisis es también relevante para explicar la formación de grupos de presión política (como los que dieron lugar a la Primavera Árabe). Este aspecto de la minería se usa también mucho para rastrear grandes movimientos demográficos.

Por otro lado, también se puede examinar cómo se difunde información social o información que tiene un impacto social. Por ejemplo, información sobre catástrofes o inicios de epidemias, con lo que se puede inferir los focos iniciales, el epicentro del evento o el paciente cero. A partir de este análisis, se pueden hacer inferencias sobre los modos como la gente reacciona a determinados tipos de eventos. De acuerdo con (Gundecha, 2012) la minería de grafos se puede usar también para establecer riesgos y límites de seguridad y vulnerabilidad de usuarios en redes, que pudieran ser puestos en peligro por sus relaciones en redes sociales. Este aspecto de la minería de grafos es relevante en casos de redes con

mucha información personal, como Facebook.

1.2 Minería de texto

La minería de texto sobre Twitter puede usarse para análisis de sentimientos y opiniones sobre distintos temas, controversiales o no. Permite clasificar la semántica de los tuits para distinguirlos en positivos o negativos y a entrenar una muestra de la data para clasificar futuros tuits como negativos o positivos. La minería de sentimientos se aplica, sobre todo, a data de productos y marcas (nuevos productos o reputación de marcas), pero también puede involucrar interrogantes relacionados con distintos aspectos de las ciencias sociales. La Minería de Texto sobre Twitter también puede usarse: Clasificación de tuits en categorías, Clustering de tuits, Detección de trending topics y análisis semántico de conceptos básicos y cambios en la percepción de aspectos de la realidad. Análisis de cambios lexicales o cambios lingüísticos. Análisis sociolingüísticos.

METODOLOGÍA

Como referencia para hacer el proceso de minería de datos se utilizó la metodología CRISP-DM (S. Moro, Marzo, 2014), (Wikimedia, 2014) la cual incluye las etapas de: Entendimiento del negocio, Comprensión y preparación de los datos; Modelado del conocimiento, Evaluación, dónde se cuantifica el rendimiento del modelo y Difusión, ver

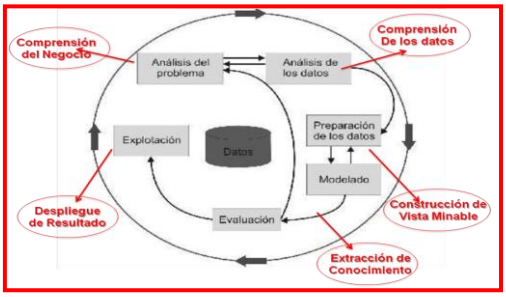

Figura 1: Metodología para Minería de Datos, CRISP-DM.

Sin embargo, considerando que el estudio trata de un tipo de minería para datos complejos como los tuits, se aplicó de manera más detallada el proceso de minería de texto descrito en la Figura 2, el cuál es una adaptación de la metodología CRISP-DM al caso de datos complejos de textos.

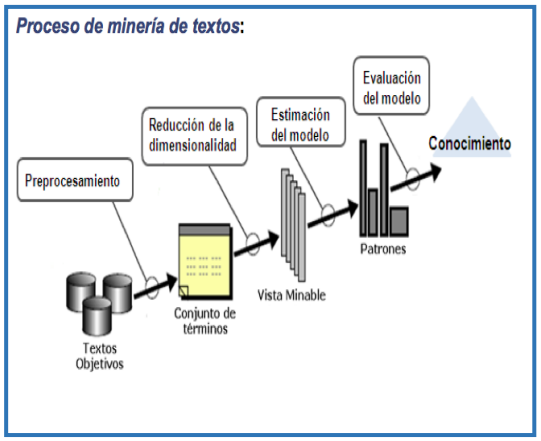

Figura 2: Proceso de Minería de Texto.

En este sentido se aplicaron las siguientes fases, que no son más que una combinación de los precitados procesos:

Comprensión del Negocio, donde se analizó los requerimientos del análisis del dominio del problema, el cual obedecía a objetivos académicos aplicados al contexto de los tuits de cuentas de líderes políticos venezolanos de interés. Aquí se seleccionaron los algoritmos de minería de datos que resuelven los objetivos de negocios y de minería de datos planteados.

En la Fase de **Comprensión de los Datos** se seleccionaron los textos objetivos: se observaron y seleccionaron las cuentas de twitter candidatas para el estudio. Se analizaron la calidad de sus tuits y la forma de uso de la red social por cada líder cuya cuenta fue seleccionada. En esta Fase se recopilaron los tuits objetivos, luego se preprocesaron los tuits recopilados y se limpiaron hasta extraer el conjunto de términos, todo esto se realizó aplicando funcionalidades de la herramienta R Studio.

Luego en la Fase de **Preparación de los Datos**, utilizando funcionalidades de la herramienta WEKA para a partir de la tabla atributo-valor construir la vista minable. Aquí se eliminaron las variables que no interesan, se realizaron las transformaciones de los datos necesarias y se redujo la dimensionalidad de la tabla, hasta construir las vistas minables a la que se le aplicaron los algoritmos de minería de datos seleccionados.

En la Fase de **Modelado**, se estimaron los modelos planificados, extrayendo los patrones luego de aplicar los algoritmos de minería de datos sobre las vistas minables. Luego se pasó a la Fase de **Evaluación del Modelo**, donde con la intervención de un experto en Filosofía se evaluaron los modelos aplicando el Criterio del Experto. Finalmente, en la Fase de **Explotación**, donde se despliega el conocimiento nuevo, se procede a difundir los resultados a través de los medios propuestos.

RESULTADOS Y ANÁLISIS

Una vez aplicados el proceso descrito con anterioridad los resultados de pueden resumir de la siguiente manera:

1.3 Negocio

El objetivo de negocio planteado es: Identificar si los "oficialistas" y "opositores" son dos clases claramente

diferenciadas en su discurso lingüístico predominante, con el propósito de establecer sus intereses, preocupaciones, preferencias. Conocer esta clasificación y características distintivas facilitará la toma de decisiones asertivas, en materia política y social, basadas en las diferencias y puntos de encuentro descubiertos. Por tanto el Objetivo General de la Minería de Datos ***es realizar un análisis de texto en los tuits escritos por tuiteros representantes de ambas tendencias políticas, tanto del oficialismo como de la oposición***. Los objetivos específicos:

1. Identificar palabras frecuentes en los tuits de seguidores claves del oficialismo y la oposición venezolana
2. Estimar un modelo para predecir si un tuitero es oficialista u opositor, en base a las palabras frecuentes presentes en sus tuits
3. Describir relaciones y/o grupos existentes entre las palabras frecuentes o de mayor uso/ importancias presentes en los tuits de oficialistas y opositores

Para ofrecer respuesta a los objetivos planteados se aplicaron las técnicas de minería de datos de clasificación para estimar un modelo predictivo clasificador y se aplicaron reglas de asociación con WEKA y R Studio para extraer modelos de tipo descriptivos que relacionen el uso de palabras frecuentes al mismo tiempo.

1.4 Fase: Comprensión de los Datos

Se creó una aplicación en Twitter utilizando el lenguaje de programación R, una herramienta open-source especializada en el manejo de datos con relevancia estadística, y R Studio, un IDE también libre ("integrated development environment" o ambiente integrado de desarrollo) para R. Se usó el paquete o librería especializada que Jeff Gentry de R para gestionar el API de Twitter: 'twitteR', (Gentry, 2015). El API de Twitter nos da acceso a su base de datos con la creación de un mecanismo de autenticación del dueño de la aplicación. Los pasos ejecutados en esta fase se describen a continuación:

1) Generación de Datos: se trata de un conjunto de tuits, donde el tipo de dato es "Texto". Para 50 tuiteros emblemáticos de los diputados e influenciadores de cada tendencia, se recopilaron 500 tuits, se generaron un total de 50.000 tuits. Luego de un pre-procesamiento, limpieza y tratamiento

estadístico se seleccionaron las palabras frecuentes de cada tuitero.

a. Elección de las cuentas de usuarios relevantes: Aplicando el conocimiento de dominio se eligieron 100 cuentas representativas: 50 de simpatizantes o miembros del Oficialismo y 50 de la oposición. Para que la elección de las cuentas garantizara ser representativas de los seguidores de ambas tendencias, se consideraron 50 cuentas de diputados electos: 25 pertenecientes al Oficialismo y 25 de la Oposición. Los otros 50 constituyen los llamados "influenciadores": periodistas o creadores de opinión pública muy citados en los medios de comunicación y muy influyentes.
b. Elección de Tuits: Decidimos trabajar con una proyección de 50.000 tuits para el análisis. El criterio del API de Twitter, cuando el investigador le pide que extraiga un número determinados de tuits de una línea de tiempo, es el de traer a la interfaz sólo aquellos que han sido escritos por el usuario, y no los que son meras retransmisiones. Dado que muchos usuarios representativos no escriben muchos tuits, se decidió modificar la función del API para extraer un número más o menos elevado de tuits por usuario. De este modo, se manipuló la función de extracción de tuits para que pidiese 500 tuits por cada usuario. Así, por ejemplo, en el caso de Diego Arria: 94 en total, entendiéndose que el resto de los tuits en la línea de tiempo del usuario "Diego Arria" (uno de los "influenciadores") son retuiteos o retransmisiones de mensajes de terceros.
c. El Minado de texto: Siguiendo las recomendaciones de (Zhao, 2015), se adaptaron los tuits extraídos al tamaño de un documento Word. Se creó un corpus de términos del usuario analizado, así como se limpió una matriz de datos de modo que sólo las palabras frecuentes sobrevivan al proceso de filtrado. Usando el paquete 'tm' (por text mining) de R, cuyas funcionalidades incluyen un listado de "stopwords" en español. Una vez realizado este trabajo de

"limpiado", el paquete twitteR (Gentry, 2015) está listo para crear dos matrices de términos por cada usuario: una de términos frecuentes y otra de asociaciones frecuentes. Para crear una matriz de términos frecuentes del usuario con una frecuencia mínima elegida por el investigador, fue necesario calibrar la matriz de palabras frecuentes de modo que se pueda extraer del timeline del usuario que tuitea un mínimo de número de palabras frecuentes. Se aplicó el criterio que por cada usuario, debe crearse una matriz de alrededor de 7 palabras frecuentes aproximadamente.

2) Recopilación e Integración: se construyeron varias matrices atributo valor, donde se exponen las palabras frecuentes de los tuiteros de cada grupo político, representando si la palabra frecuente esta (1) o no (2) asociado a ese tuitero. Dichos términos frecuentes luego sirvieron para construir diversas matrices atributo-valor cuya forma tiene:

CuentaUserTw	Nro. Tweets	$Palabra_1$	$Palabra_2$	...	...	...	$Palabra_n$
$@userTw_1$	245	1	2	...	...	...	2
...	30	2	2	...	...	...	1
$@userTw_{25}$	400	1	1	...	...	...	2

Tabla 1: Tabla Atributo-Valor

Se obtuvieron cuatro (4) tablas de palabras frecuentes:

Tabla 2: Tabla Atributo-Valor de Influenciadores de la Oposición

Documentos	Nº de tuits	academicas	adelante	años	asamblea	BCV	bolivares	cabello
ArmandoInfo	234	2	2	2	2	2	2	2
avedefenix2	328	2	2	2	1	2	2	2
DiarioTalCual	244	2	2	2	2	2	2	2
Diego_Arria	487	2	2	2	1	2	2	1
Econ_Vzla	102	2	2	1	2	2	2	2
efernandezVE	484	2	2	2	2	2	2	2
garciabanchs	357	2	2	2	2	2	2	2
hcapriles	237	2	2	2	2	2	2	2
HenkelGarcia	210	2	2	2	2	2	1	2
HenriFalconLara	499	2	2	2	2	2	2	2
jesuscasique1	476	2	2	1	2	1	2	2
JOSEVICENTEHARO	252	2	2	2	1	2	2	2
Leonardo_Padron	30	2	2	2	1	2	2	2
LiborioGuarulla	199	2	2	1	2	2	2	2
luisoliveros13	91	2	2	2	1	2	2	2
MariaCorinaYA	105	2	2	2	1	2	2	2
NoMasGuiso	384	2	2	2	1	2	2	2
NTN24ve	48	2	2	1	2	2	2	2
odelopez	91	2	2	1	1	2	2	2
Pr1meroJusticia	25	2	2	2	2	2	2	2
RCR750	326	1	2	2	1	2	2	2
robertoenriq	94	2	1	2	2	2	2	2
RonaldBalzaG	36	2	2	2	2	2	2	2
Sangarccs	143	2	2	2	1	2	2	2
Ulises_UD	354	2	2	2	2	2	2	2

1.5 Fase: Preparación de los Datos

Estas tablas Atributo-valor se aglutinaron en dos (2) matrices generales de diputados e influenciadores de ambas tendencias: MudPsuvDip y MudPsuvInf. Seguidamente se crearon los archivo csv que posteriormente se convirtieron en dos archivo arff. Se construyeron varias vistas minables adecuadas para cada tipo de técnica que se planificó aplicar. Para construir las vistas minables se eliminaron los siguientes atributos: Atributo que añade ruido y no es de interés para el estudio: Nro. Tweets y Atributos identificadores: CuentaUserTw. Se aplicó la transformación para cambiar el rango de las variables de las palabras frecuentes, para obtener una vista minable con valores nominales binarios adecuados para aplicar la técnica de reglas de asociación. Se realizó una selección de atributos basada en el conocimiento del dominio, para obtener el conjunto de las palabras frecuentes más relevantes para el estudio que involucren tuiteros de ambas tendencias. Particularmente se aplicaron varios algoritmos de selección de atributos, para diferentes pruebas de aplicación de algoritmos.

Vistas Minables: las vistas minables se construyeron a partir de las matrices MudPsuvDip y MudPsuvInf. Estas vistas minables son:

MudPsuvDip.arff. Esta vista representa las palabras frecuentes de los diputados de ambas tendencias políticas. Consta de 47 instancias y 48 palabras frecuentes que la herramienta WEKA puede tratar como atributos con valores nominales 1 (sí aparece), 2 (no aparece) en los dos grupos que conforman las clases “Opositores” (clase azul) y “Oficialistas” (clase roja).

Al aplicarle selección de atributos se observan los siguientes resultados:

Digital: ISSN 2790 5039 – Depósito legal digital: DC2021001122
Impreso: ISSN 23435658 – Depósito legal impreso: PP201202DC4092

Tabla 3: Tabla Selección de Atributos en MudPsuvDip.arff

Algoritmo de Selección	Atributos Elegidos
Evaluador: CfsSubsetEval Buscador: BestFirst	Bloque, Cambio, Chávez, Chavista, CNE, Constitución, Crisis, Derecha, Diputados, Estudiantes, Fuerza, Mayoría, Patria, Paz, Pueblo, Revolución, Trabajo, Venezuela, Victoria
Evaluador: InfoGainAttributeEval Buscador: Ranker	Chavista, Pueblo, Revolución, Venezuela, Bloque, Chávez, Patria, Diputados, Derecha, CNE, Cambio, Maduro, Paz, Trabajo, Victoria, Bolívar, País
Evaluador: ChiSquareAttributeEval Buscador: Ranker	Chavista, Pueblo, Revolución, Venezuela, Bloque, Chávez, Patria, Diputados, Derecha, CNE, Maduro, Cambio, Trabajo, País, Paz, Bolívar, Victoria, Adeco, Constitución, Agua, Corrupción, Crisis

Con estos tres evaluadores, hemos construido tres vistas minables para aplicar clasificación y agrupación. Estas vistas minables, difieren en sus conjuntos de atributos, con el propósito de evaluar cuál es el conjunto de atributos o palabras frecuentes que permiten clasificar de mejor manera. Estas Vistas Minables son: MudPsuvDip (1CfsSE), MudPsuvDip (1ChiSAE) y MudPsuvDip (1IGAE). Se les eliminó el atributo clase, fueron usadas para análisis de agrupamiento y tienen las siguientes formas:

Tabla 4: Tabla Modelo de Vista Minable de MudPsuvDip.arff para agrupamiento.

$Palabra_1$	**$Palabra_2$**	**…**	**…**	**…**	**$Palabra_n$**
1	2	…	…	…	1
…	…	…	…	…	…
2	1	…	…	…	2

1. MudPsuvInf.arff: Esta vista representa las palabras frecuentes de los influenciadores de ambas tendencias políticas. Consta de 50 instancias y 105 palabras frecuentes que la herramienta WEKA puede tratar como atributos con valores nominales 1 (sí aparece), 2 (no aparece) en los dos grupos que conforman las clases "Opositores" (clase azul) y "Oficialistas" (clase roja).

Esta Vista Minable fue usada con modesto éxito para análisis de clasificación y tienen la siguiente forma:

Palabra$_1$	Palabra$_2$	...	...	...	Palabra$_n$	Clase
1	2	...	...	...	1	1
...	...	...	...	...	...	...
2	1	...	...	...	2	2

Tabla 5: Tabla Modelo de Vista Minable de MudPsuvInf.arff para Clasificación.

Algoritmo de Selección	Atributos Elegidos
Evaluador: CfsSubsetEval Buscador: BestFirst	Celac, Gente, Gobernación, Liliantintori, Nicolas, Patria, Venezuela, Vía
Evaluador: PrincipalComponents Buscador: Ranker	Académicas, Adelante, Anticrisis, Años, Asamblea, BCV, Bloque, Bolívar, Bolívares, Cabello, Camarada, Cambio, CELAC, Chávez, Comparto, Ciudadanos, Constitución, Control, Coño, Corrupción, Crisis, Cuba, Cultura, Decreto, Democracia, Derechos, Economía, Ecuador, EEUU, Emergencia, España, Estudiantes, Fenómeno, Fortalecer, Fuerza

Tabla 6: Tabla Selección de Atributos en MudPsuvInf.arff.

Adicionalmente, luego de eliminar el atributo lase, una vez aplicada la transformación de valores al tipo binario, se le aplicó selección de atributos con **los** siguientes resultados:

Esta Vista Minable fue usada para extraer *"relaciones"* y determinar las palabras frecuentes que se usan juntas y son "temas". La misma tiene la siguiente forma:

Palabra$_1$	**Palabra$_2$**	**…**	**…**	**…**	**Palabra$_m$**
0	0	…	…	…	1
…	…	…	…	…	…
1	1	…	…	…	0

Tabla 7: Tabla Modelo de Vista Minable de MudPsuvInf.arff para Asociación.

Adicionalmente, el trabajo de minado de texto permitió crear una serie de matrices susceptibles de ser analizadas en el sentido propuesto por nuestra investigación, que serán de mucha utilidad para complementar el análisis.

1.6 Fase de Modelado:

El primer objetivo específico planteado es *Identificar palabras frecuentes en los tuits de seguidores claves del oficialismo y la oposición venezolana*. Este resultado se logró con las fases de Comprensión y Preparación de los Datos anteriormente descritas. No requiere de una vista minable, aunque si se realizó minado de texto para extraer el conjunto de términos desde los tuits o textos objetivos.

Para lograr el segundo objetivo específico de minería planteado, *estimar un modelo para predecir si un tuitero es oficialista u opositor, en base a las palabras frecuentes presentes en sus tuits,* se aplicó la técnica de clasificación sobre las vistas minables: MudPsuvDip.arff y MudPsuvInf.arff, arrojando resultados satisfactorios sobre la vista minable de los influenciadores. Se aplicó el algoritmo IRip, J48 y IBK de WEKA, sobre MudPsuvInf.arff, siendo más preciso los resultados del JRip, los cuales fueron:

Precisión	**Exactitud**	**Sensibilidad**
65%	70%	84%

Tabla 8: Medidas de Clasificaciones

Clasificaciones Correctas	35	70%
Clasificaciones Incorrectas	15	30%

Tabla 9: Matriz de Confusión del Clasificador

	a	b	
a	21	4	a=1
b	11	14	b=2

Tabla 10: Medidas de Evaluación del Clasificador

El Modelo Clasificador es:

=== Classifier model (full training set) ===

JRIP rules:

==========

(celac = 1) => clase=2 (11.0/0.0)

=> clase=1 (39.0/14.0)

Number of Rules : 2

El clasificador obtenido, ver figura 3, es modesto y sus reglas no son concluyentes en relación a los términos frecuentes asociados a las clases. Adicionalmente no se pudo construir un clasificador con la vista minable de los diputados. Se aplicó el algoritmo de aprendizaje automático supervisado como J48 sobre MudPsuvDip.arff, y las versiones de esta vista minable luego de la selección de

atributos. Estos clasificadores arrojaron un resultado con una precisión del 100%, utilizando la palabra "chavista" para discriminar entre los dos grupos. Dado que se trata de un resultado previsible y no muy interesante no se ahondará más detalles al respecto.

Para responder al objetivo de describir *relaciones y/o grupos existentes entre las palabras frecuentes o de mayor uso/importancia presentes en los tuits de oficialistas y opositores,* se aplicaron algoritmos de agrupamiento y de asociación sobre las vistas minables de disputados e influenciadores. Para describir grupos de términos entre las palabras frecuentes se aplicaron técnicas de agrupamiento en las vistas minables: MudPsuvDip.arff y sus variantes obtenidas después de aplicar diversas selecciones de atributos: MudPsuvDip (1CfsSE).arff, MudPsuvDip (1ChiSAE).arff y MudPsuvDip (1IGAE).arff. En principio se aplicó el algoritmo de agrupación de Weka *SimpleKMeans sobre* MudPsuvDip.arff, sin aplicar selección de atributos. Los resultados fueron:

#Cluster	0	1
	"bloque", "Bolívar", "Chávez", "chavista", "derecha", "diputados", "Maduro", "país", "patria", "paz", "pueblo", "revolución", "trabajo", "Venezuela", "victoria"	**"Cambio", "CNE"**
Clustered Instances	**20 (43%) (clase roja y negativa)**	**27 (57%) (clase azul y positiva)**

Tabla 11: Modelo de Agrupamiento, MudPsuvDip.arff

Para ampliar el estudio se aplicó el algoritmo KMeans sobre las vistas minables resultantes de aplicar la selección de atributos, se obtuvo los siguientes resultados:

MudPsuvDip(1CfsSE).arff, luego de aplicar CfsSubsetEval, ver tabla IV. (19 atributos, 47 Instancias)		
#Cluster	**0**	**1**
	"bloque", "chávez", "chavista", "derecha", "diputados", "patria", "paz", "pueblo", "revolución", "trabajo", Venezuela", "victoria"	**"Cambio", "Constitución"**
Clustered Instances	**22 (47%)**	**25 (53%)**
MudPsuvDip(1CfsSE).arff, luego de aplicar InfoGainAttributeEval, ver tabla IV. (17 atributos, 47 Instancias)		
#Cluster	**0**	**1**
	"bloque", "Bolívar", "chávez", "chavista", "derecha", "diputados", "maduro", "país", "patria", "paz", "pueblo", "revolución", "trabajo", Venezuela"	**"Cambio", "CNE"**

Clustered Instances	**21 (45%)**	**26 (55%)**
MudPsuvDip(1ChiSAE).arff, luego de aplicar ChiSquared, ver tabla IV. (22 atributos, 47 Instancias)		
#Cluster	**0**	**1**
	"bloque", "Bolívar", "chávez", "chavista", "derecha", "diputados", "maduro", "país", "patria", "paz", "pueblo", "revolución", "trabajo"	**"Cambio", "CNE"**
Clustered Instances	**21 (45%)**	**26 (55%)**

Tabla 12: Modelo de Agrupamiento con selección Atributo

El algoritmo clasifica correctamente a los dos grupos de términos frecuentes, gracias a los centroides que ofrece el grupo Oficialista, esto es interesante: el algoritmo distingue entre los dos grupos porque puede clasificar correctamente al grupo del Oficialismo. El algoritmo modela correctamente dos grupos de palabras de uso frecuente, según la figura 4:

Figura 4: Modelo Grupos de Términos Frecuentes

Finalmente se aplicó técnicas de reglas de asociación sobre las vistas minables: MudPsuvDip.arff y MudPsuvInf.arff; así como sobre y las vistas minables que se derivaron luego de seleccionar atributos. Los resultados más interesantes son:

MudPsuvDip.arff	Algoritmo: **Apriori**
1.rojos=1 46 ==>regimen=1 46 <conf:(1)> conv:(0.98) 2.regimen=1 46 ==>rojos=1 46 <conf:(1)> conv:(0.98) 3.venezolanos=1 46 ==>unidad=1 46 <conf:(1)> conv:(0.98) 4.unidad=1 46 ==>venezolanos =1 46 <conf:(1)> conv:(0.98) 5.futuro=1 45 ==>gracias=1 45 <conf:(1)> conv:(0.96) 6.gobierno=1 45 ==>gracias =1 45 <conf:(1)> conv:(0.96) 7.presospoliticos=1 45 ==>unidad =1 45 <conf:(1)> conv:(0.96) 8.presospolitico =1 45 ==>venezolanos =1 45 <conf:(1)> conv:(0.96) 9.vecinos=1 45 ==>unidad =1 45 <conf:(1)> conv:(0.96)	

MudPsuvInfB.arff	Algoritmo: Apriori
1.estudiantes=1 49 ==> academicas =1 49 <conf:(1)> conv:(0.98) 2.academicas =1 49 ==> estudiantes=1 49 <conf:(1)> conv:(0.98) 3.democracia=1 49 ==> ciudadanos=1 49 <conf:(1)> conv:(0.98) 4.ciudadanos=1 49 ==> democracia=1 49 <conf:(1)> conv:(0.98) 5.bloque=1 48 ==> anticrisis=1 48 <conf:(1)> conv:(0.96) 6.adelante=1 estudiantes=1 48 ==> academicas =1 48 <conf:(1)> conv:(0.96) 7.academicas =1 adelante=1 48 ==> estudiantes=1 48 <conf:(1)> conv:(0.96) 8.anticrisis=1 estudiantes=1 48 ==> academicas =1 48 <conf:(1)> conv:(0.96) 9.academicas =1 anticrisis=1 48 ==> estudiantes=1 48 <conf:(1)> conv:(0.96) 10.BCV=1 estudiantes=1 48 ==> academicas =1 48 <conf:(1)> conv:(0.96)	
MudPsuvInfC.arff	**Algoritmo: Apriori**
1.gobernacion =1 49 ==> gente =1 48 <conf:(0.98)> conv:(0.49) 2.gente =1 49 ==> gobernacion =1 48 <conf:(0.98)> conv:(0.49) 3.liliantintori =1 49 ==> gente =1 48<conf:(0.98)> conv:(0.49) 4.gente =1 49 ==> liliantintori =1 48 <conf:(0.98)> conv:(0.49) 5.Nicolas =1 49 ==> gente =1 48 <conf:(0.98)> conv:(0.49) 6.gente =1 49 ==> Nicolas =1 48 <conf:(0.98)> conv:(0.49) 7.totalitarismo=1 49 ==> gente=1 48 <conf:(0.98)> conv:(0.49) 8.gente=1 49 ==> totalitarismo=1 48 <conf:(0.98)> conv:(0.49) 9.liliantintori=1 49 ==> gobernacion=1 48 <conf:(0.98)> conv:(0.49) 10.gobernacion=1 49 ==> liliantintori=1 48 <conf:(0.98)> conv:(0.49)	

Tabla 13: Modelo de Asociación

Este modelo de asociación obtenido, muestra el cumplimiento del nivel de confianza mayor al 90%. Esto significa que los términos relacionados en cada regla extraída aparecen juntos más del 90% de las veces. Estas reglas facilitan el estudio sobre posibles "temas" presentes en los tuits. En vista de la riqueza del modelo y considerando que el alcance del presente estudio, en el sentido de asociaciones, no es tan detallado, se propone cómo reto futuro el análisis exhaustivo del modelo a la luz del juicio de expertos filósofos o politólogos interesados en el tema. Sin embargo, para introducir un poco en el detalle de esta reflexión, se hace especial mención por su poderoso significado las siguientes reglas:

liliantintori==>gobernación <conf:(0.98)>

gobernacion==>liliantintori<conf:(0.98)>

R1: El 98% de las veces que es usado el término "liliantintori" es usado el término "gobernación" y viceversa.

rojos===>regimen<conf:(1)>

regimen==>rojos <conf:(1)>

R2: El 100% de las veces que aparece el término "rojo" aparece el término "régimen" y viceversa.

venezolanos==>unidad <conf:(1)>

unidad==>venezolanos <conf:(1)>

R3: El 100% de las veces que aparece el término "venezolanos" aparece el término "unidad" y viceversa.

bloque=1 48 ==> anticrisis=1 48 <conf:(1)> conv:(0.96)

R4: El 96% de las veces que aparece el término "bloque" aparece el término "anticrisis"

Todas estas reglas sugieren temas que relacionan los términos que aparecen juntos con cierta frecuencia. Este es un estudio de mucho valor, que amerita otros modelos, análisis y estudios con el juicio de expertos, que se deja para próximas entregas.

1.7 Evaluación de Modelos:

Los Criterios de Evaluación, basados en el "Criterio de Experto":

- Las palabras de uso frecuentes descubiertas para cada tuitero analizado deben tener una frecuencia de uso mayor o igual al 30% de las veces
- Estimar un modelo predictivo del tipo de tuitero con un mínimo de asertividad del 60% de las veces
- Que el modelo descriptivo descubra relaciones que cumplan una frecuencia mínima del 60 %

En este sentido se confirma que las palabras frecuentes fueron "minadas" cumpliendo el criterio de frecuencia según el número de tuits escritos por los usuarios seleccionados. Estos términos frecuentes seleccionados cumplen un mínimo de frecuencia de 7 veces en adelante. Esta frecuencia es sobre un total de 20 tuits, lo cual representa más del 30% de la frecuencia de uso esperada.

En relación con el clasificador, no es objetivo de este estudio usarlo para clasificar tuits, este objetivo queda como trabajo futuro. Sin embargo, se puede observar que el clasificador tiene una medida de evaluación basada en su capacidad predictiva buena, con valores de precisión, exactitud y sensibilidad por encima del 60%, lo cual es un buen pronóstico que indica que el criterio de evaluación exigido tiene posibilidades de ser alcanzado en el momento de usar el clasificador.

En relación con el modelo de asociación, es importante destacar que las reglas extraídas todas cumplen una confianza por encima del 90 %, lo cual es mayor de lo esperado. Finalmente, al evaluar el modelo de agrupamiento con el criterio del juicio de experto y basado en el conocimiento del dominio, los dos

grupos extraídos son los naturales muy cerca de la realidad.

Se requiere complementar esta evaluación con el análisis de los modelos obtenidos a través del juicio de un experto, en este caso profesionales de Filosofía y Política, los cuales podrán confirmar su utilidad futura, por ejemplo en análisis oportunos en el contexto de las elecciones planificadas para diciembre del corriente.

1.8 Fase de Difusión de los Modelos:

Se pretende representar el modelo en un lenguaje adecuado a la terminología del dominio del problema. Escribir artículos y planificar conferencias para compartir los resultados. Los hallazgos obtenidos luego del análisis, evaluación e interpretación de los resultados serán publicados en un artículo y conferencias al público de interés.

CONCLUSIONES

El análisis de los datos reveló que sí existen diferencias importantes en el contenido discursivo que caracteriza a los dos grupos que se disputan el poder desde hace algunos años en Venezuela y se observa un mayor "cohesionamiento" entre los representantes del grupo perteneciente al Oficialismo en contraste con los representantes de las fuerzas de la Oposición.

Existen una serie de palabras de uso frecuente, que han terminado siendo características representativas de la pertenencia a cada uno de los grupos, las cuales también revelan los sentimientos, preferencias e intereses de ambos grupos. Siendo este último aspecto una realidad que requiere ser estudiada a mayor profundidad en el futuro a fin de establecer temas, tendencias filosóficas y de comportamiento de ambos grupos.

El minado de Twitter reveló anomalías en el discurso político, tales como una preferencia por significados provenientes de la teoría económica (un área, en principio, ajeno al ámbito de persuasión de la política tradicional) por parte de la Oposición, y el uso probable, por parte de dos políticos importantes, uno de la Oposición y otro del Oficialismo, de administradores de redes sociales con el fin de mantener vigente su imagen y su nombre en ese ámbito.

Se puede concluir que, pese a la poca densidad de la data, la herramienta de minado de Twitter y de WEKA

ofrecen un rico campo de análisis para el filósofo y el analista político que desee comprender cuáles son las preocupaciones, preferencias de acción y tendencias discursivas que pudieran caracterizar a los grupos políticos en punga en un momento dado de la historia de un país.

En un futuro trabajo, el clasificador construido puede ser explotado desde una aplicación de reconocimiento automático (machine learning) que permitiera a terceros establecer con alguna precisión si un tuitero pertenece a uno de los dos grupos en discusión. De igual manera, se espera en trabajos futuros explotar detalladamente el modelo de asociación construido y formalizar un estudio semántico con el uso de un dendograma jerárquico difuso.

REFERENCIAS:

Bifef, F. E. (2010). *"Sentiment Knowledge Discovery in Twitter Streaming Data".* University of Waikato. Nueva Zelanda: Hamilton

Gentry, J. (2015, julio 29). Package 'twitteR'. repositorio CRA

Gundecha, P. y. (2012). "Mining Social Media: A Brief Introduction". Revista Informs, disponible en http://dx.doi.org/10.1287/educ.1120.0105.

K. Sylwester, M. P. (2015). "Twitter Language Use Reflects Psychological, Differences between Democrats and Republicans" . Revista PLOS One

S. Moro, M. R. (Marzo, 2014). Using Data Mining for Bank direct marketing: An application of the CRISP-DM Methodology. Repositorium, http://repositorium.sdum.uminho.pt/bitstream/1822/14838/1/MoroCortezLaureano

Wikimedia. (2014, Marzo). A diagram showing the relationship between the different phases of CRISP-DM and illustrates the recursive nature of a data mining project. Retrieved from http://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png

Zhao, Y. (2015). R and Datamining: Examples and Case Studies. In Y. Zhao, "Text Mining" (p. capítulo 10). http://www.RDataMining.com.