Polaridad de las opiniones sobre un personaje público en el Ecuador

Polarity of opinions about a public person in Ecuador

Boris Herrera Flores

***Resumen*— La presente investigación es el estudio de las técnicas de minería de opiniones, enfocada a obtener información de un personaje público en el Ecuador, determinando indicios de polaridad a su gestión en forma positiva, negativa o neutra, resultado que le permitirá a dicho personaje público tomar decisiones sobre su accionar en función de una imagen de servicio a la comunidad. La extracción de las opiniones en redes sociales y técnicas basadas en Tecnologías del Lenguaje Humano posibilitó la interpretación de los datos polarizados precisando parámetros de relevancia a la opinión resultante enfocados a la toma de decisiones, procesamiento que se adapta a los nuevos formatos de comunicación logrando la interpretación y valoración de la opinión. Las redes sociales fue la plataforma para la captura de textos por medio de un API, que luego del procesamiento del lenguaje natural se obtuvieron resultados de indicios de la popularidad del personaje.**

Palabras clave: **Procesamiento del lenguaje natural, minería de opiniones, análisis de sentimientos, tecnologías del lenguaje humano, clasificación de opiniones.**

***Abstract*- The present investigation is the study of opinion mining techniques, focused on obtaining information from a public figure in Ecuador, determining signs of polarity for your management in a positive, negative or neutral way, a result that will allow said character public to make decisions about their actions based on an image of service to the community. The extraction of opinions in social networks and techniques based on Human Language Technologies enabled the interpretation of polarized data by specifying parameters of relevance to the resulting opinion focused on decision making, processing that adapts to the new communication formats achieving the interpretation and assessment of opinion. Social networks was the platform for the capture of texts by means of an API, which after the processing of the natural language obtained results of indications of the popularity of the character.**

Index Terms: **Natural language processing, opinion mining, sentiment analysis, human language technologies, rating of opinions**

Article history:
Received 18 September 2018
Accepted 24 September 2018

El autor pertenece a la Facultad de Ingeniería Ciencias Físicas y Matemática de la Universidad Central del Ecuador, Ciudadela Universitaria Av. América, Quito, Ecuador. (e-mail: bherrera@uce.edu.ec)

I. INTRODUCCIÓN

Actualmente la red social es la plataforma Web 2.0 para que los usuarios expresen sus opiniones, puntos de vista en forma inmediata con una participación frecuente usando un lenguaje informal, identifico la necesidad de procesar estos párrafos, oraciones y palabras precisando grados de relevancia a la informalidad, al definir polaridad positiva, neura o negativa, usando las Tecnologías del Lenguaje Humano.

Este estudio se enfoca a la comprensión de las expresiones subjetivas y el lenguaje informal usado por el Ecuatoriano al opinar sobre un personaje público a su accionar de servicio a la comunidad, donde las actuales tecnologías de lenguaje humano no son directamente aplicables y es necesario incorporar métodos y herramientas para el procesamiento del lenguaje natural por medio de la experimentación de diferentes técnicas y sistemas genéricos de análisis de sentimientos que extraigan primero la fuente de información para luego dar un tratamiento con métodos que determinen la polaridad de la opinión.

El objetivo es la mejora de los recursos, técnicas y herramientas que modelan el lenguaje subjetivo e informal que genera la red social Twitter, con el tratamiento del lenguaje emocional en un entorno de subsistemas inteligentes de procesamiento para la recuperación, tratamiento, compresión y descubrimiento de la información apto para la toma de decisiones.

II. BASES CONCEPTUALES

Twitter, es una aplicación web gratuita de microblogging que recibe un estimado de 313 millones de visitas al mes, que reúne las ventajas de los blogs, las redes sociales y la mensajería instantánea, esta nueva forma de comunicación permite a sus usuarios estar en contacto en tiempo real con personas de su interés a través de mensajes breves de texto a los que se denominan Tweets, por medio de una sencilla pregunta: ¿Qué estás haciendo? Los usuarios envían y reciben Tweets de otros usuarios a través de breves mensajes que no deben superar los 140 caracteres, vía web, teléfono móvil, mensajería instantánea o a través del correo electrónico; e incluso desde terceras aplicaciones.

Una API de Twitter (siglas en inglés "Application Programming Interface"), es un conjunto de reglas y especificaciones que las aplicaciones pueden seguir para comunicarse entre ellas, sirviendo de interfaz entre programas diferentes de la misma manera que la interfaz de usuario facilita

la interacción humano-software. Twitter tiene tres tipos de APIs: REST API, Search API y Streaming API:

REST API, ofrece a los desarrolladores el acceso al core de los datos de Twitter. Todas las operaciones que se pueden hacer vía web son posibles realizarlas desde el API, soporta los formatos: xml, json, rss, atom.

Search API, suministra los tweets con una profundidad en el tiempo de 7 días que se ajustan a la consulta demandada. Es posible filtrar por, cliente utilizando, lenguaje y localización, no requiere autenticación y los tweets se obtienen en formato json o atom.

Streaming API, establece una conexión permanente por usuario con los servidores de Twitter y mediante una petición http se recibe un flujo continuo de tweets en formato json.

Con Search API y en el REST API existe una limitación de 150 peticiones a la hora por usuario o por IP si la llamada no estuvo autenticada.

El análisis de sentimientos, de textos en las redes sociales es el proceso que determina el tono emocional que hay detrás de una palabra determinada, si una frase contiene una opinión positiva, neutra, o negativa sobre un producto, marca, institución, organización, empresa, evento o persona, el objetivo es extraer aquellos términos semánticos que expresen un sentimiento en particular para conocer la opinión, las actitudes y las expectativas sobre un tema en concreto así como para analizar el comportamiento de los usuarios ante algún mensaje y, por tanto, determinar su impacto o poder anticipar su reacción.

La determinación de la polaridad, consistente en determinar cuándo una opinión es positiva, negativa o neutra con respecto a la entidad a la cual se está refiriendo desde dos enfoques diferentes dependiendo del tipo de método que se utiliza, como son:

- Método de clasificación supervisada: Los rasgos extraídos del texto y el método de aprendizaje determinan cuando este pertenece a la clase positiva o negativa.
- Método de clasificación no supervisada (clustering): Tienen en cuenta la presencia de palabras con orientaciones conocidas que son obtenidas de diccionarios o corpus como, por ejemplo, las palabras excelente o alegre, que son representativas de expresiones polares positivas.

Existen varias técnicas al momento de estimar un texto y su posible polaridad positiva, negativa o neutra. Al utilizar software en la minería de opiniones se encuentran dos funciones: classify polarity y score.sentiment

- classify polarity, Utiliza un marco probabilístico, tiene sus fundamentos en el Teorema de Bayes para calcular la probabilidad de una clase (positiva, negativa). Es una técnica de clasificación y predicción supervisada, entre sus ventajas está que su implementación es muy fácil y obtiene buenos resultados de clasificación en la mayoría de los casos:

p(CjF1; :::; Fn) (1)

Donde C representa un valor positivo (+) o negativo (-) y F1; F2; :::; Fn factores que representa las palabras de un tweet, donde se busca establecer la probabilidad a priori de que un tweet sea positivo o negativo.

- Score.sentiment, es una función muy simple de predicción no supervisada que asigna una puntuación simplemente contando el número de ocurrencias de positivos y negativos en un tweet comparando con un diccionario de palabras conocido como corpus.

score = sum(PalabrasPositivas) - sum(PalabrasNegativas) (2)

El éxito de esta función es tener la mayor cantidad de palabras positivas y negativas, existen algunos diccionarios ya cargados con estas palabras uno de ellos es MPQA, LIWC.

III. METODOLOGÍA

La metodología a utilizar es SEMMA creada por SAS Institute, que fue propuesta especialmente para trabajar con el software SAS Enterprise Miner. Si bien en la comunidad científica se conoce a SEMMA como una metodología, en el sitio de la empresa SAS se aclara que este no es el objetivo de la misma, sino más bien la propuesta de una organización lógica de las tareas más importantes del proceso de minería de datos. SEMMA establece un conjunto de cinco fases para llevar a cabo el proceso de minería:

Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado) y Assess (Evaluación).

Para el esquema general al detectar la polaridad de opiniones, se debe tomar en cuenta las palabras considerando su sentido correcto en un contexto determinado:

- Identificar Tokens: En esta primera etapa se realiza un pre-procesamiento del texto de las opiniones y devuelve los términos que aportan información útil, o definido como limpieza de los datos como por ejemplo duplicados generados por trolls
- Desambiguar lexicalmente cada token: En la segunda etapa se parte de cada término que aporta información útil, este le da forma, lematiza y desambiguan lexicalmente.
- Obtener todas las acepciones de cada palabra: Esta etapa parte del listado de términos lematizados y desambiguados y devuelve todas las acepciones del término en el idioma en que se esté realizando el análisis.
- Clasificar cada token en positivo y negativo: Una vez obtenidas todas las acepciones de cada término, en esta etapa se propone determinar la polaridad del término de acuerdo a los algoritmos que se empleen, finalmente la polaridad de la opinión se determina por el número de palabras positivas y negativas que contiene.
- Evaluar la opinión: Si el número de palabras positivas que contiene una frase es mayor se considera la frase como positiva y en caso contrario es negativa.

Los índices de relevancia identificados en twitter son:

- Ratio seguidores/seguidos: Los usuarios que tienen un ratio de seguidores/seguidos cercano al 1 suelen contar

con el efecto followback: muchas cuentas les siguen para ganar seguidores fácilmente. Para encontrar su relevancia usaremos una métrica basándonos en el ratio de seguidores/seguidos, pero esta será solo una parte, la parte fundamental para encontrar la relevancia se encuentra en el número de retweets que ha tenido el tweet asignándole

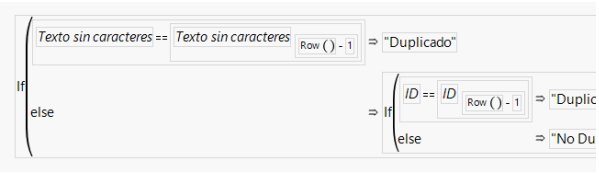

Figura 1. Regla de JMP para limpieza de datos

un factor de multiplicidad de 6.5 al valor original, así mismo al campo favoritos que indica cuantas veces el tweet ha sido marcado como favorito, se le asignó un factor de multiplicidad de 3.5.

- Retweets/nº seguidores: Los retweets son un indicador universal de que lo que dices es interesante o digno de atención. Cuantos más retweets tengas, más interesante o relevante es lo que dices o quien eres. El número de tweets es un factor importante, ya que no es lo mismo conseguir 100 retweets con un solo tweet que hacerlo a lo largo de un mes con 300.

El software para la extracción de datos utilizada es #TAGS, una aplicación que funciona con Google Docs, que permite entre otras cosas conocer los tweets y las conversaciones que ha tenido un usuario determinado ofreciendo características adicionales del tweet como ubicación geográfica, id usuario, seguidores, favorito, retweets, etc.

En el estudio realizado aplicando la metodología indicada anteriormente, el personaje público obtuvo 71.845 tweets en el periodo.

IV. RESULTADOS EXPERIMENTALES

En estudio realizado de 71.845 tweets usando el software de JMP de SAS se aplicó la limpieza a los datos, eliminando links en el texto, tweets vacíos, duplicados aplicando la fórmula de la Figura N°1

La polaridad se ensayó con la siguiente definición de un corpus básico de polaridad en la Tabla I.

Se determina si es positivo con el valor de 1, neutro con el valor de 0 y negativo con el valor de -1.

Las siguientes son las métricas utilizadas para el estudio:

Número de seguidores, A esta métrica la hemos definido de tal manera que si nuestro personaje público tiene más de 200 seguidores en cada uno de sus tweets es aceptado por parte de los twitteros en la red social.

Figura 2. Regla de JMP de Número de seguidores

Aceptabilidad, esta métrica nos da a conocer que tan aceptado es el mensaje compartido por nuestro personaje por los twitteros en la red social.

TABLA I. Corpus básico de polaridad

No.	Positivo	Negativo
1	Mejor	Ratas
2	Felicitaciones	Ladrón
3	Felicidades	Mentira
4	Compañeros	Malo
5	Bien	Peor
6	Bienvenido	Rechazo
7	Confianza	Nunca
8	Confiable	Vago
9	Respaldo	Lame botas
10	Bueno	Tirano
11	Solidario	Corrupto
12	Hermoso	Borrego
13	Lindo	Chavista
14	Éxito	Traición
15	Ético	Traidor
16	Aprobado	Inepto
17	Adelante	Decepción
18	Ganador	Perdedor

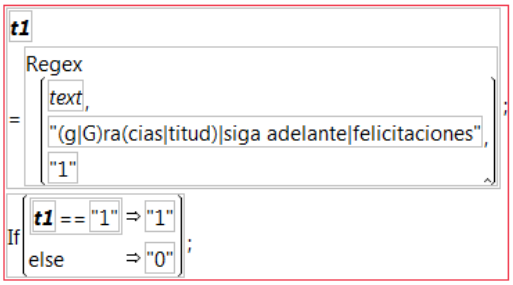

Figura 3. Regla de JMP de Aceptabilidad

Ratio de iteraciones /fans, esta fórmula nos permite determinar cuan relevante es una cuenta en relación al número de fans.

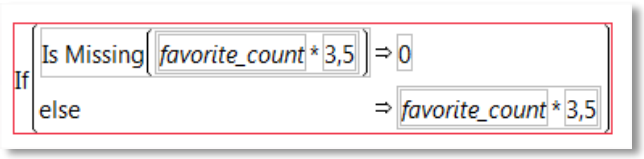

Figura 4. Regla de JMP de Valoración de favoritos

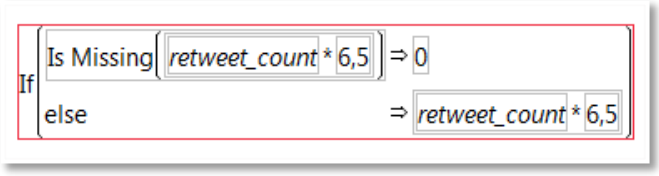

Figura 5. Regla de JMP de Valoración de retweets

Figura 6. Regla de JMP de Ratio seguidores/seguidos

Figura 7. Criterio de relevancia y resultados

En lo referente a la polaridad del ejercicio se pude identificar que:

- 3% Negativo
- 92% Neutro
- 4% Positivo

Las opiniones sobre el personaje público son en su mayoría neutras, definiendo que su accionar tampoco en negativo pero que tampoco da mucho que celebrar con opiniones positivas.

Al aplicar el score.sentiment, que es la función de predicción no supervisada se obtiene una puntación positiva en los tweets, esto es comparando con un diccionario de palabras corpus.

Al establecer las métricas de relevancia a las polaridades positiva y neutral, se puede identificar lo siguiente:

- 14% Nada relevante
- 84% Poco relevante
- 0% Algo relevante
- 2% Normalmente relevante

Por medio del conteo de la métrica de relevancia, la información del cuadro indica que las opiniones sobre el personaje público fueron de poca relevancia, por los bajos niveles de retweets, y de poca marcación en favoritos.

V. CONCLUSIONES Y TRABAJOS FUTUROS

- Existen retos en el procesamiento del lenguaje natural que deben ser resueltos en la detección automática de la polaridad de un texto, esto es el corpus en español que entienda el léxico ecuatoriano o latinoamericano, este trabajo de investigación lo detecta e identifica una línea de estudio que puede ser trabajada en la comunidad de investigadores.

Valoración de favoritos	Valoración de Retweets	Ratio Seguidores/Seguidos	Relevancia
0	0	0,952681388	0
0	0	0,952681388	0
0	6,5	11,757731959	76,425257732
84	149,5	124,77484183	29134,925568
0	3425,5	0,5596868885	1917,2074364
24,5	39	19,728401192	1252,7534757
0	0	111,77002967	0
28	208	17,157360406	4049,1370558
0	0	17,756722151	0
0	26	17,698209719	460,15345269
0	0	13,534482759	0
73,5	143	124,77484183	27013,753256
0	0	6,8653846154	0
0	962	0,023255814	22,372093023
0	6,5	2,1091674765	13,709588597
3,5	13	2,1091674765	34,801263362
0	253,5	2,1080467229	534,38984426
0	6,5	2,1080467229	13,702303699
0	0	2,1102106969	0
24,5	52	4,6952736318	359,18843284
0	97,5	4,6514143095	453,51289517
0	188,5	4,6514143095	876,79159734
38,5	84,5	4,6952736318	577,51865672
0	97,5	4,6514143095	453,51289517
7	13	4,6514143095	93,02828619
35	58,5	4,6952736318	439,00808458
0	110,5	4,6514143095	513,9812812
0	0	4,6514143095	0
10,5	39	2,1091674765	104,40379009
0	182	2,1080467229	383,66450357
0	260	2,1080467229	548,00214706

Figura 8. Regla de JMP de Relevancia y resultados

- La definición de polaridad de sentimientos dentro de este estudio de opinión de un personaje público determina indicios o tendencias al tipo de impacto, no tiene margen de error como una encuesta tradicional pero si provee un resultado que ayuda a reforzar la toma de decisiones.
- La contribución del usuario en el lenguaje informal es determinante en una temporalidad cual puede variar en opinión, dichos textos deben ser preprocesados para obtener una data que aporte a un resultado relevante.
- Los niveles de re-tweets y marcación en favoritos determinan la relevancia de una opinión, métrica utilizada para el estudio de una opinión en twitter, pueden ser también identificadas otras como ubicación geográfica, Número de seguidores, etc.
- El trabajo futuro es trabajar en otros dominios como el educativo que permitiría identificar el manejo del contenido por el Docente, el seguimiento a las unidades o temas planteados y la relación interpersonal Docente Estudiante.

VI. REFERENCIAS

[1] Borras-Morell, Jose Enrique, "Data Mining for Pulsing the Emotion on the Web" Collection: Methods in Molecular Biology Volume: 1246 pp: 123-130, 2015.

[2] Chen, Xin; Vorvoreanu, Mihaela; Madhavan, Krishna, "Mining Social Media Data for Understanding Students' Learning Experiences" IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES Volume: 7 Number: 3 pp: 246-259, 2014.

[3] Khan, Farhan Hassan; Bashir, Saba; Qamar, Usman, "TOM: Twitter opinion mining framework using hybrid classification scheme" DECISION SUPPORT SYSTEMS Volume: 57 pp: 245-257, 2014.

[4] Akaichi, Jalel; Dhouioui, Zeineb; Lopez-Huertas Perez, Maria Jose "Text Mining Facebook Status Updates for Sentiment Classification" Conference: 17th International Conference System Theory, Control and Computing (ICSTCC) Ubication: Sinaia, ROMANIA pp: 640-645, 2013.

[5] Colace, F.; De Santo, M.; Greco, L, "SAFE: A Sentiment Analysis Framework for E-Learning" International Journal of Emerging Technologies in Learning Volume: 9 Number: 6 pp: 37-41, 2014.

[6] Colace, Francesco; Casaburi, Luca; De Santo, Massimo, "Sentiment detection in social networks and in collaborative learning environments" COMPUTERS IN HUMAN BEHAVIOR Volume: 51 Number especial: SI pp: 1061-1067, 2015.

[7] Baldominos Gomez, Alejandro; Luis Mingueza, Nerea; Cristina Garcia del Pozo, Ma, "OpinAIS: An Artificial Immune System-based Framework for Opinion Mining" INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE Volume: 3 Number: 3 pp: 25-34, 2015.

Boris Herrera Flores, estudió su grado en la Universidad Central del Ecuador obteniendo el título de Ingeniero en Informática en la Universidad Central del Ecuador, después realizó los estudios de Maestría en la misma Universidad graduándose como Magister en Gestión Informática Empresarial. Actualmente se desempeña como Docente en la Facultad de Ingeniería Ciencias Físicas y Matemática, Carrera de Ingeniería Informática y cursa el Doctorado de Informática en la Universidad de Alicante