Caracterización de líderes políticos de Ecuador en Twitter usando aprendizaje de maquina no supervisado

Johnny Torres[1], Gabriela Baquerizo[2], Carmen Vaca
[1]Escuela Superior Politécnica del Litoral (ESPOL)
jomatorr@fiec.espol.edu.ec, cvaca@fiec.espol.edu.ec
[2]Universidad Casa Grande
gbaquerizo@casagrande.edu.ec

Resumen. El crecimiento exponencial del uso de las redes sociales permite a los usuarios comunicarse directamente con la audiencia y causar impacto. Líderes políticos hacen uso cada vez mas de dichas plataformas para interactuar con sus seguidores. En el presente proyecto se emplean técnicas de aprendizaje de maquina no supervisado para conocer cuáles son las temáticas que los lideres políticos colocan en la opinión pública a través de las redes sociales y, determinar cómo se relacionan los titulares de periódicos digitales en un determinado periodo de tiempo. Se ha obtenido datos de lıderes políticos en la plataforma Twitter para llevar a cabo experimentos que permitan aplicar técnicas de clustering de documentos para extraer los tópicos relevantes y, los resultados obtenidos se han evaluado con datos de las publicaciones digitales de diario El Universo y diario El Telégrafo en el periodo que comprende el estudio. De esta forma se ha podido identificar que la agenda mediática corresponde a los tópicos que se discuten en el espacio publico a través de las publicaciones que los lideres políticos generan en las redes sociales.

Palabras clave: Aprendizaje de maquina no supervisado, clustering, modelamiento de tópicos, twitter.

1 Introducción

En los últimos anos se han propuesto una serie de métodos para identificar usuarios denominados “influyentes”. Estos usuarios corresponden a individuos que generan cascadas de información que se vuelven virales [1], y se ha observado que medidas cuantitativas tales como el número de conexiones sociales no constituyen un predictor absoluto de influencia [2–4]. Estas conexiones sociales pueden corresponder al número de seguidores en Twitter o número de amigos en Facebook.

Al problema de la identificación de usuarios influyentes se agrega la dificultad para detectar el nivel de “pericia” de dichos usuarios en determinados tópicos. Pese a que un usuario puede expresar opiniones en diversos ámbitos, estudios previos han observado que los usuarios se vuelven expertos en no más de dos tópicos en promedio [5].

El presente trabajo se enfoca en identificar los tópicos más relevantes que los lıderes políticos ecuatorianos proponen a la opinión publica a través de la plataforma de Twitter en el periodo de 01 de junio a 23 de agosto de 2015 y, determinar cómo se relacionan con datos publicados en medios digitales de comunicación en un periodo especıfico de tiempo. Usando técnicas de aprendizaje de maquina no supervisado se extraen automáticamente los tópicos de dominio [6] de usuarios con audiencias representativas, donde el tamaño de la audiencia se mide por el número de seguidores en la plataforma en línea. Finalmente, para estos tópicos se utiliza una métrica de similar dad para

contrastar estos tópicos con el contenido publicado por medios digitales de corte generalista.

En la sección 2 de este estudio se destacan algunas publicaciones relacionadas al tema. Mientras que en el apartado 3 se explica el proceso de obtención de los datos que se usaron para los experimentos de Twitter y la fuente referencial. La identificación de métricas de volumen de actividad y atención se detallan en la sección 4, seguida de la 5 donde se describe el análisis de los tópicos relevantes extraídos de los datos. En la sección 6 se muestran los resultados obtenidos en los experimentos. Finalmente, en el bloque 7 se encuentran las conclusiones del estudio y se discute el trabajo futuro.

2 Trabajos relacionados

Twitter ha llegado a ser una de las plataformas de comunicación más usadas [7] a nivel mundial, y sus características particulares tales como mensajes limitados a 140 caracteres han permitido realizar numerosos estudios aplicando técnicas de aprendizaje de máquina para extraer información relevante. Weng et al. [8] proponen una técnica para medir la influencia de usuarios de Twitter alrededor de tópicos sensitivos, en el que se hace uso del modelo probabilístico LDA (Latent Dirichlet Allocation) [6] para descubrir los tópicos de interés de los usuarios usando los tweets que publican. En dicho estudio el objetivo es detectar automáticamente los tópicos de los cuales discuten los usuarios de manera general. El proceso consiste en agregar todos los tweets de un usuario en un solo documento previo a extraer los tópicos, es decir, cada documento del corpus que se usara en el proceso corresponde a la información publicada por un usuario.

En otro estudio realizado por Zhao et al. [9] se analiza el contenido publicado por usuarios de Twitter para determinar si puede ser considerado como una plataforma de noticias que cubre mayormente la misma información que los medios digitales, este autor realiza una comparación del contenido publicado en Twitter frente a lo publicado en un periódico digital generalista como el New York Times empleando modelamiento de tópicos no supervisado LDA. Bakshy et al. [10] destacan la importancia de disponer de una fuente referencial de datos para comparar los resultados obtenidos al aplicar modelos computacionales para determinar lıderes influyentes en Twitter.

Personas trabajando en el ámbito de periodismo necesitan mejores herramientas para explotar las ventajas ofrecidas por las plataformas de social media en las cuales el ciudadano puede producir información instantáneamente [11]. Sin embargo, esta ventaja de velocidad de producción de información en Twitter tiene como contraparte la poca credibilidad que pueden tener algunos posts hechos en dicha red social [12]. En este trabajo nosotros nos enfocamos en proveer una metodología para extraer información de Twitter que puede ser útil para los periodistas pues proviene de fuentes de información creíbles: usuarios que han sido previamente evaluados como lıderes de opinión.

Cuadro 1: Lista consolidada de lıderes políticos

Nombre	Twitter	Seguidores
Rafael Correa	MashiRafael	2345358
José Serrano Salgado	ppsesa	761701
Jaime Nebot	jaimenebotsaadi	479682
Dalo Bucaram	daloes10	330982
Carlos Vera	CarlosVerareal	279078
Mauricio Rodas	MauricioRodasEC	242883
Guillermo Lasso	LassoGuillermo	203134
Viviana Bonilla	viviana_bonilla	193873
Jorge Glas	JorgeGlas	184643
Abdala Bucaram Ortiz	abdalabucaram	181262
Ricardo Patiño Aroca	RicardoPatinoEC	171265
Vinicio Alvarado	vinizeta	132820
Paul Carrasco C.	PaulCarrascoC	112313
Jimmy Jairala	jimmyjairala	103448
Fernando Cordero	fcorderoc	94042
Fernando Alvarado E	FAlvaradoE	90720
Susana González	6SusanaGonzalez	63618
Martha Roldós	martharoldos	45960
María Duarte	Maria_DuarteP	33878
Doménica Tabacchi	dometabacchi	30251
Lucio Gutiérrez	LucioGutierrez3	30132

3 Datos

Para este estudio se obtuvieron datos de lıderes de opinión en Ecuador y a partir de ese conjunto de usuarios, se extrajeron tweets disponibles de forma pública. La lista de lıderes políticos con mayor número de seguidores e influencia fue obtenida de dos fuentes independientes, el estudio realizado por la consultora Llorente & Cuenca [13] y los datos de informes estadísticos publicados por el sitio web Social Baker [14]. En la tabla 1 se muestra la lista consolidada de lıderes políticos.

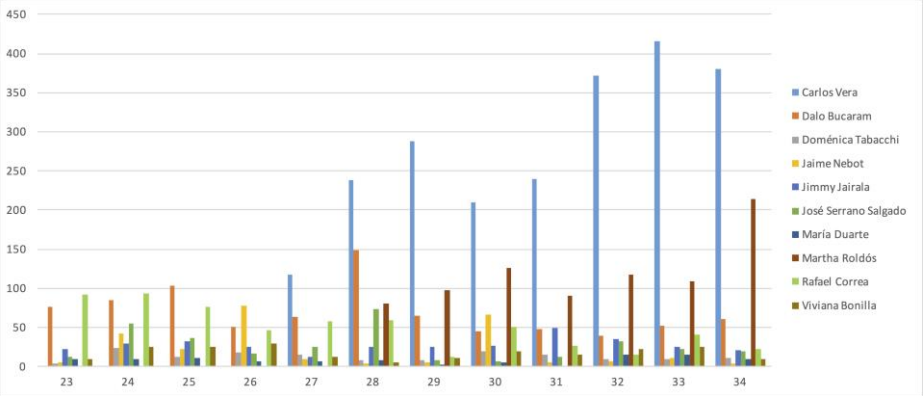
Una vez realizada la consolidación se obtuvo la información del perfil de cada usuario y los tweets publicados en su “timeline” durante el periodo de 01 de junio a 23 de agosto de 2015. Para extraer estos datos se empleó la interfaz de programación de aplicaciones (API) REST de Twitter [15]. Para cada usuario se obtuvo hasta un máximo de 3,000 tweets, que es el lımite impuesto por el API de Twitter para la extracción del timeline de los usuarios.

Los datos, que servirán como fuente referencial para evaluar la credibilidad de los tópicos extraídos de la plataforma de social media, son los titulares de las publicaciones más relevantes de la sección política de diario El Universo [16] y diario El Telégrafo [17] durante el periodo de 01 de junio a 23 de agosto de 2015. Estos datos se recolectaron manualmente por un periodista y se usan en los experimentos para realizar una comparación de los tópicos extraídos de las publicaciones en Twitter en relación con los temas tratados en los diarios digitales generalistas seleccionados para esta investigación.

4 Identificando lıderes influyentes

Una vez obtenidos los datos se utiliza la ecuación 1 para calcular el volumen de actividad de los lıderes políticos, esto consiste en el número de tweets publicados durante el periodo seleccionado para los experimentos.

Figura 1: Lıderes políticos con mayor volumen de actividad en Twitter.



$$Vac = sum(tweetCounts)_p \quad (1)$$

En la figura 1 se muestran los lıderes con mayor actividad en Twitter durante el periodo seleccionado, en el que destacan el periodista y político Carlos Vera y la economista y política Martha Roldos.

Como se ha mencionado en [3] el número de seguidores o publicaciones no son las métricas más adecuadas para determinar si un líder es influyente en Twitter, en su lugar deben utilizarse métricas que cuantifiquen el nivel de interacción que este mantiene con sus seguidores. En relación con esa premisa se puede establecer el volumen de atención que cada líder ha recibido por parte de su audiencia al contabilizar los retweets como se propone en la ecuación 2.

$$Vat = sum(retweets)_p \quad (2)$$

La figura 2 nos muestra que por volumen de atención solo el Presidente de la Republica Rafael Correa y el alcalde de Guayaquil Jaime Nebot se mantienen en el top 5. Nótese que han generado un volumen de atención alto en el número de retweets en ciertas fechas especıficas, más adelante se tratara de identificar los tópicos relevantes a través del uso de técnicas de aprendizaje de maquina no supervisado.

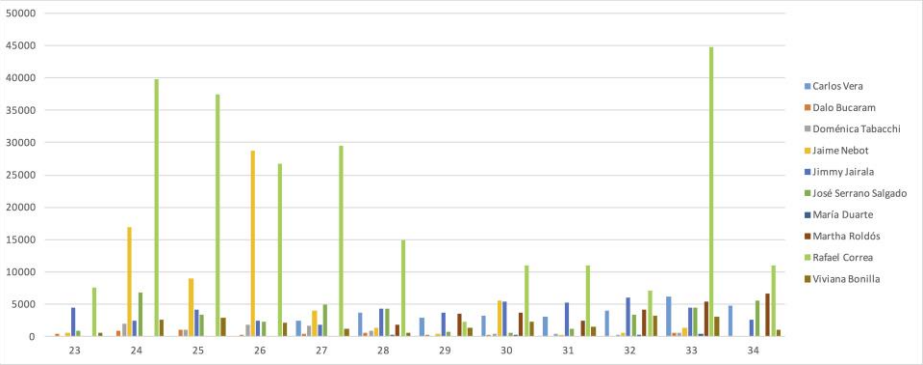
5 Obteniendo tópicos relevantes

El problema de determinar tópicos en un documento se puede resolver aplicando técnicas de aprendizaje de máquina, específicamente aprendizaje no supervisado, ya que se trata de encontrar estructuras en datos no etiquetados, es decir no hay ejemplos dados al algoritmo que permita evaluar si hay error o ventaja en una potencial solución.

Algunos de los enfoques de aprendizaje no supervisado para extracción de tópicos comprenden modelos de clustering tales como K-means, Modelos de Mixture,

Clustering Jerárquico; o modelos de variables latentes tales como Dirichlet Allocation (LDA),

Figura 2: Lıderes políticos con mayor volumen de atención en Twitter.



Análisis Semántico (LSA), Factorización de Matriz No Negativa (NMF). Este artıculo se centrará en encontrar tópicos relevantes empleando los modelos de variable latente, específicamente LDA [18, 19], el cual es un modelo popular para extraer tópicos en documentos.

El modelo de datos para los experimentos comprende tweets publicados por los lıderes políticos que son parte de una colección de documentos. En la ecuación 3 se describe cada tweet, que es un pequeño documento especıfico que se compone de un numero de identificación o ID (´ *tweetID*), el usuario que lo público (´*createdBy*), la fecha, la hora de publicación (´ *createdAt*) y, el texto del tweet publicado (*tweet*).

$$D = d_i; \quad \text{done} \quad d_i = (tweetID_i, createdBy_i, createdAt_i, tweet_i) \tag{3}$$

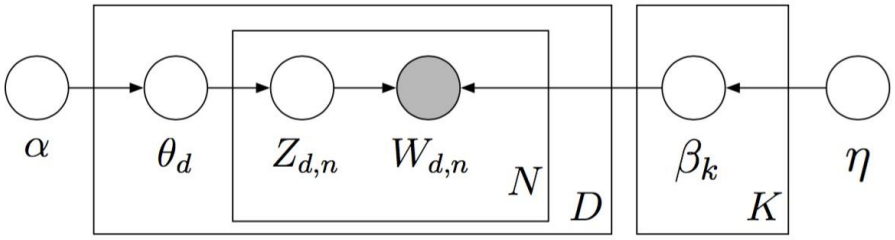
El primer paso para extraer la información relevante de estos documentos consiste en una tarea de limpieza del texto de los tweets, que corrige acentos y elimina símbolos de puntuación, además de transformar la codificación del texto de cada tweet a tipo ASCII. Después se realiza una tarea de procesamiento de lenguaje natural (PLN) para eliminar palabras con limitado significado semántico y mantener términos relevantes como sustantivos o adjetivos del contenido de cada tweet. Por último, cada palabra clave es transformada a la raíz del vocablo mediante un análisis lingüístico morfológico.

El siguiente proceso consiste en construir un diccionario de términos en el que se asigna una identificación numérica a cada expresión, dicha identificación se usa en una matriz de términos donde se asigna un peso a la palabra extraıda de cada tweet. El valor asignado a cada vocablo está basado en la frecuencia con la que aparece dicho termino dentro del documento. El contenido del documento *i* se organiza en un conjunto de tuplas *(k,w)* de *n* palabras claves y pesos asociados como se define en la ecuación 4.

$$tweet_i = (k_{ij}, w_{ij}); \quad donde \quad j = [1...n]; \tag{4}$$

Como representación de la colección de documentos se obtiene una matriz donde las columnas muestran los pesos asignados a cada palabra clave y las filas se asocian a los documentos. Los documentos están ordenados como series de tiempo de manera

Figura 3: Representación grafica del modelo LDA.



que los tweets más antiguos están primeros y los más recientes son las columnas finales. La tabla que representa los *m* documentos como se define en la ecuación 5.

$$DT = [tweet_i, d_i]; \quad i = [1...m]; \tag{5}$$

Una vez obtenida la matriz de documentos, en los experimentos se aplica el modelo Latent Dirichlet Allocation (LDA) para la extracción de tópicos. Un tópico es una distribución de probabilidad sobre una colección de palabras, es decir una relación estadística entre un grupo de variables aleatorias observadas y latentes (o desconocidas) que nos permiten definir un modelo probabilístico para generar tópicos [6, 18]. Consecuentemente, si tomamos el conjunto de palabras con mayor valoración en dicha distribución obtendremos un resumen automático de los contenidos más importantes en la colección de documentos que nos ayudan a responder la pregunta ¿ De qué se habla en los tweets publicados?.

El modelo Latent Dirichlet Allocation (LDA) propuesto por Blei et al. [6] usa un tipo de variable oculta para obtener tópicos de un conjunto de documentos. En LDA los datos observados son las palabras de cada documento mientras que las variables ocultas representan la estructura tópica latente, es decir los tópicos y como lo muestra cada documento.

El proceso de generación de tópicos para cada documento está compuesto de dos fases:

1. Se escoge aleatoriamente una distribución sobre los tópicos

Para cada palabra en el documento

a) Se escoge aleatoriamente un tópico de la distribución sobre los tópicos en el paso 1.

b) Se escoge aleatoriamente una palabra de la distribución correspondiente sobre el vocabulario.

La figura 3 muestra una representación grafica del modelo LDA, en que los nodos denotan variables aleatorias y son etiquetados de acuerdo a su rol en el proceso generativo. Los nodos "ocultos", distribución de tópicos y asignaciones están no sombreados. Los nodos observados -las palabras en los documentos- están sombreadas. Los rectángulos denotan replicación. *N* denota colección de palabras dentro del documento; mientras que *D* denota la colección de documentos.

En los experimentos una colección de tweets, representan los documentos $D=d_1, d_2,... d_n$ y un numero de topicos $T=t_1,t_2,... t_n$. Podemos extraer los tópicos *T* de tal manera que un documento D_i puede ser visto por su distribución de tópicos, por ejemplo: $Pr(D_1, t_1) = 0.40$, nos indica la probabilidad de que tal documento pertenezca a un determinado tópico.

Para realizar la comparación de tópicos extraídos de los tweets de los lıderes políticos y las publicaciones de noticias en periódicos digitales se ha utilizado la métrica de similitud de coseno. Los tópicos extraídos de cada tweet son comparados con respecto a los tópicos extraídos de las noticias del periodo de estudio. Considerando el vector de tópicos de cada tweet $A= \{a_1,a_2,... a_n\}$ y el vector de tópicos de cada noticia $B=b_1,b_2,... b_n$, el coseno de dos vectores puede ser definido usando el producto punto Euclideano como se muestra en la ecuación 6.

$$A \cdot B = \|A\|\|B\| \cos(\theta) \tag{6}$$

Dado los dos vectores de tópicos, *A* y *B*, la similitud de coseno, cos(θ) esta representada usando el producto punto y la magnitud como se muestra en la ecuación 7.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{7}$$

Para el cálculo de la similitud total entre los tweets de los lıderes políticos y las publicaciones de noticias, se toma para cada líder el valor más alto en que los topicos de un tweet se asemejan a una noticia en un perıodo. Luego se obtiene el promedio de la similitud de todas las noticias en ese perıodo.

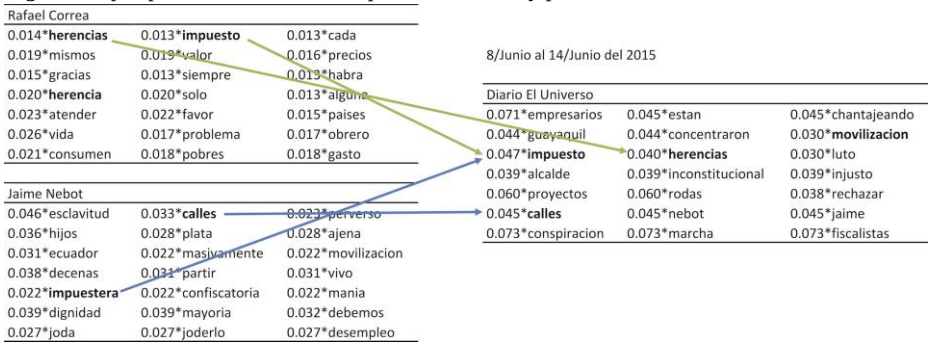
6 Resultados

En los experimentos se han utilizado tweets publicados por cada uno de los lıderes políticos durante el periodo de 01 de junio a 23 de agosto de 2015, una recopilación de 64,492 tweets en total. Se han empleado 678 titulares extraídos de las principales

noticias de la sección política del sitio web de diario El Universo y 1,015 de diario El Telégrafo, publicados en el periodo correspondiente al estudio. Para el análisis y extracción de tópicos se ha utilizado como herramienta las librerías desarrollados por Radim et al. [20].

Un ejemplo de extracción de tópicos de las noticias se puede apreciar en la figura 4, donde cada fila representa un tópico mientras que el valor asociado a cada palabra muestra la contribución de dicha palabra respecto al tópico. Se ha realizado un análisis basado en diferentes niveles de agregación de tweets. El primer nivel de agregación de tweets constituye el periodo de tiempo en intervalos de una semana y, el segundo nivel de agregación corresponde a los usuarios de Twitter o publicaciones de diario El Universo y diario El Telégrafo

Figura 4: Ejemplo de extracción de tópicos de Tweets y publicaciones de Noticias.



La tabla 2 muestra los tópicos obtenidos para los lıderes con mayor volumen de atención. Se observa que las palabras que mas contribuyen a los tópicos (ecuador, derecho, rebelión, pobres, Correa, Guayaquil, país) corresponden a los temas de coyuntura política y social discutida por la opinión publica en los principales diarios del país durante el periodo de estudio.

Cuadro 2: Tópicos mas relevantes por volumen de Atención.

Lider	Semana	Retweets	Topicos
MashiRafael	6/8/15	42849	0.013*ecuador + 0.012*derecho + 0.012*rebelion 0.019*solo + 0.015*bienes + 0.015*gasto 0.011*clase + 0.011*media + 0.009*favor 0.022*vida + 0.019*gracias + 0.015*pobres 0.016*habra + 0.013*menos + 0.010*basilica
jaimenebotsaadi	6/22/15	29886	0.020*debia + 0.018*correa + 0.016*reducido 0.032*nueva + 0.019*guayaquil + 0.019*dueno 0.024*reves + 0.024*quiere + 0.020*cree 0.026*pais + 0.025*queremos + 0.019*simplemente 0.022*cada + 0.022*incendie + 0.022*legitimo

En el cuadro 3 se puede observar que los tópicos que tuitean varios lıderes políticos tiene un promedio de 31.77% considerando retweets (similitud 1) y 21.92% usando solo contenido creado originalmente por cada usuario (similitud 2). En ambos casos se observa que algunos líderes superan el 50% de la similitud con los temas que se

publican diario El Universo. Entre los lıderes políticos que sobresalen por volumen actividad están Jaime Nebot, Martha Roldos, María Duarte, Dalo Bucaram, Carlos Vera y Rafael Correa.

La figura 5 muestra que los tópicos que mas se destacan son los de la semana del 6 de julio, en las publicaciones tomadas de diario El Universo y los mensajes publicados en las redes sociales, en este caso en la plataforma Twitter, son: Guayaquil, marcha, gobierno, tiranía, revolución, dialogo, democrática, país, corrupción, constitución, poder, paro nacional, oposición, oligarquía; temas que son colocados en la opinión pública por

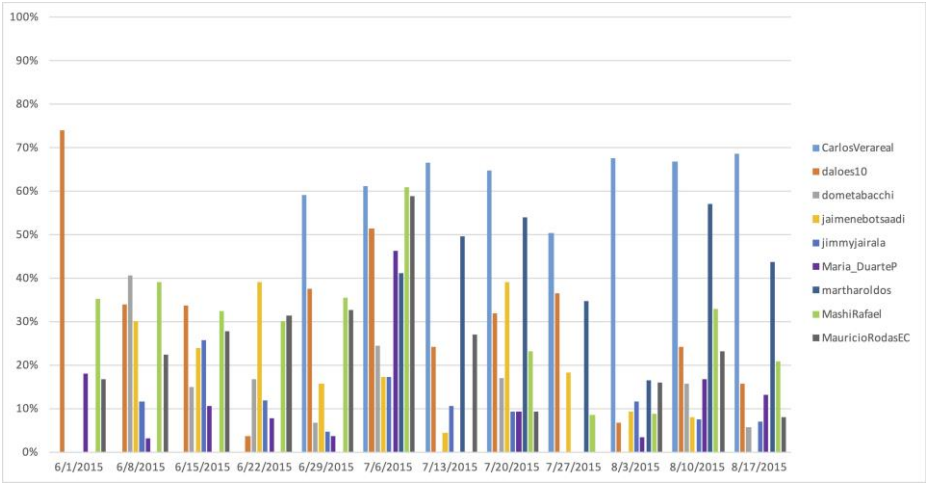
Cuadro 3: Similitud de tópicos de tweets y noticias de diario El Universo.

Similitud 1	Líderes									
Semana	CarlosVerareal	daloes10	dometabacchi	jaimenebotsaadi	jimmyjairala	Maria_DuarteP	martharoldos	MashiRafael	MauricioRodasEC	Total
6/1/15	0.00%	76.17%	0.00%	0.00%	16.67%	49.04%	0.00%	35.16%	16.67%	21.52%
6/8/15	0.00%	52.59%	36.04%	30.06%	45.84%	15.78%	0.00%	44.86%	23.49%	27.63%
6/15/15	0.00%	45.67%	27.11%	25.55%	35.54%	33.75%	0.00%	29.74%	25.31%	24.74%
6/22/15	0.00%	30.49%	35.30%	41.88%	14.35%	36.23%	0.00%	23.53%	31.55%	23.70%
6/29/15	53.70%	41.28%	39.54%	13.34%	19.31%	44.19%	0.00%	37.56%	27.90%	30.76%
7/6/15	67.63%	64.02%	20.17%	10.00%	35.67%	30.79%	72.85%	51.11%	38.44%	43.41%
7/13/15	71.21%	49.88%	18.37%	14.67%	22.93%	21.71%	75.78%	12.37%	23.31%	34.47%
7/20/15	75.89%	31.92%	32.73%	40.73%	22.13%	29.07%	77.72%	40.77%	9.42%	40.04%
7/27/15	67.02%	36.56%	20.18%	18.31%	27.37%	18.99%	77.34%	0.00%	0.00%	29.53%
8/3/15	76.73%	11.44%	7.67%	9.19%	31.70%	21.68%	79.80%	9.23%	15.34%	29.20%
8/10/15	77.47%	50.92%	40.10%	3.61%	37.69%	35.41%	72.40%	29.25%	19.25%	40.68%
8/17/15	80.10%	41.49%	27.73%	0.00%	28.80%	13.96%	88.30%	28.52%	11.06%	35.55%
Total	**47.48%**	**44.37%**	**25.41%**	**17.28%**	**28.17%**	**29.21%**	**45.35%**	**28.51%**	**20.14%**	**31.77%**

Similitud 2	Líderes									
Semana	CarlosVerareal	daloes10	dometabacchi	jaimenebotsaadi	jimmyjairala	Maria_DuarteP	martharoldos	MashiRafael	MauricioRodasEC	Grand Total
6/1/15	0.00%	73.87%	0.00%	0.00%	0.00%	18.03%	0.00%	35.16%	16.67%	15.97%
6/8/15	0.00%	33.95%	40.52%	30.06%	11.54%	3.17%	0.00%	39.01%	22.36%	20.07%
6/15/15	0.00%	33.59%	14.93%	23.77%	25.58%	10.58%	0.00%	32.46%	27.84%	18.75%
6/22/15	0.00%	3.65%	16.74%	38.98%	11.83%	7.61%	0.00%	30.07%	31.30%	15.58%
6/29/15	59.01%	37.54%	6.81%	15.68%	4.77%	3.50%	0.00%	35.35%	32.65%	21.70%
7/6/15	61.10%	51.43%	24.28%	17.32%	17.32%	46.29%	41.01%	60.83%	58.67%	42.03%
7/13/15	66.43%	24.07%	0.00%	4.27%	10.67%	0.00%	49.42%	0.00%	27.00%	20.21%
7/20/15	64.74%	31.73%	16.86%	39.01%	9.35%	9.32%	53.93%	23.08%	9.37%	28.60%
7/27/15	50.20%	36.56%	0.00%	18.31%	0.00%	0.00%	34.60%	8.45%	0.00%	16.46%
8/3/15	67.40%	6.63%	0.00%	9.19%	11.52%	3.28%	16.43%	8.75%	15.91%	15.46%
8/10/15	66.79%	24.03%	15.57%	7.88%	7.55%	16.65%	57.04%	32.98%	23.08%	27.95%
8/17/15	68.61%	15.75%	5.56%	0.00%	6.97%	13.21%	43.57%	20.68%	7.98%	20.26%
Grand Total	**42.02%**	**31.07%**	**11.77%**	**17.04%**	**9.76%**	**10.97%**	**24.67%**	**27.23%**	**22.74%**	**21.92%**

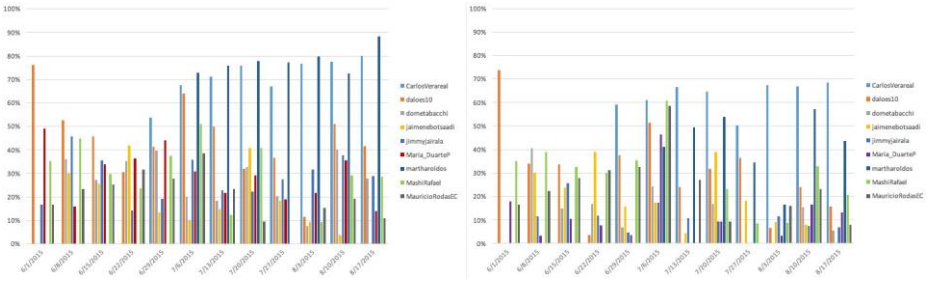
los lıderes políticos con mayor volumen de actividad y que tienen más del 50% de similitud con las publicaciones que se realizan en diario El Universo durante esa misma semana.

Figura 5: Comparación de tópicos con noticias de diario El Universo sin usar retweets.



Con esto inferimos que había una alta actividad en redes sociales frente a la coyuntura política que atravesaba el país durante esos días previos al paro nacional convocado para el 13 de agosto como mecanismo de protesta contra las medidas planteadas en el proyecto de reforma de los pagos por herencias y plusvalía propuesto por el ejecutivo.

Figura 6: Similitud con y sin retweets en relación con las noticias de diario El Universo.



Se realizó una comparación de similitud usando los retweets que no fueron creados originalmente por los usuarios, es decir contenido que fue creado por otros usuarios y los lideres simplemente realizaron retweet. En la gráfica derecha de la figura 6 muestra que cuando no se consideran los retweets la similitud baja considerablemente para la mayoría de líderes, en relación a la gráfica de la izquierda que sı considera los retweets, es decir contenido creado por otros usuarios de Twitter.

Cuadro 4: Similitud de tópicos de tweets con noticias de Diario El Telégrafo

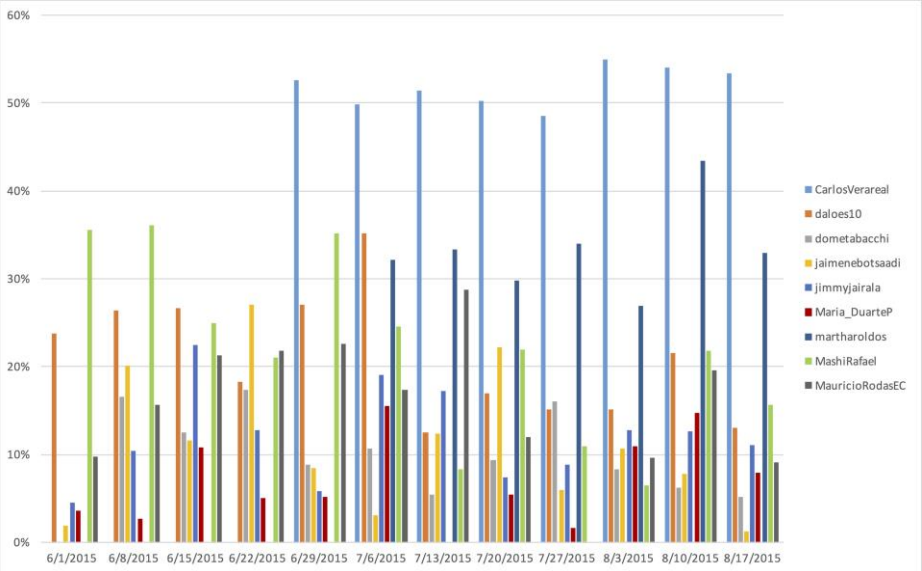
Similitud 1	Líderes									
Semanas	CarlosVerareal	daloes10	dometabacchi	jaimenebotsaadi	jimmyjairala	Maria_DuarteP	martharoldos	MashiRafael	MauricioRodasEC	Total
6/1/15	0.00%	30.26%	0.00%	1.75%	16.52%	27.60%	0.00%	31.41%	9.87%	13.05%
6/8/15	0.00%	34.36%	18.55%	21.46%	30.41%	20.80%	0.00%	36.49%	15.20%	19.70%
6/15/15	0.00%	33.29%	17.31%	13.27%	29.96%	28.84%	0.00%	30.74%	24.83%	19.80%
6/22/15	0.00%	32.64%	24.89%	31.94%	21.51%	30.57%	0.00%	23.94%	25.53%	21.22%
6/29/15	40.59%	35.95%	28.67%	10.11%	24.11%	32.43%	0.00%	43.27%	22.94%	26.45%
7/6/15	56.97%	45.33%	3.13%	3.13%	25.85%	9.36%	43.76%	27.65%	19.42%	26.06%
7/13/15	64.14%	25.38%	15.01%	5.31%	29.79%	22.77%	65.66%	8.38%	27.94%	29.37%
7/20/15	67.28%	24.96%	10.25%	22.99%	9.46%	20.95%	59.67%	21.94%	12.03%	27.73%
7/27/15	49.94%	20.98%	14.72%	6.11%	16.95%	16.38%	58.66%	15.85%	0.00%	22.18%
8/3/15	61.89%	19.79%	9.33%	3.63%	38.34%	23.40%	59.62%	5.84%	9.72%	25.73%
8/10/15	68.85%	36.48%	31.88%	7.87%	30.64%	28.45%	64.32%	28.39%	18.77%	35.07%
8/17/15	60.74%	33.61%	14.24%	0.00%	28.48%	11.11%	73.66%	20.35%	13.90%	28.45%
Total	**39.20%**	**31.09%**	**15.66%**	**10.63%**	**25.17%**	**22.72%**	**35.45%**	**24.52%**	**16.68%**	**24.57%**

Similitud 2	Líderes									
Semanas	CarlosVerareal	daloes10	dometabacchi	jaimenebotsaadi	jimmyjairala	Maria_DuarteP	martharoldos	MashiRafael	MauricioRodasEC	Total
6/1/15	0.00%	23.79%	0.00%	1.88%	4.49%	3.61%	0.00%	35.59%	9.78%	8.79%
6/8/15	0.00%	26.42%	16.62%	20.11%	10.44%	2.66%	0.00%	36.16%	15.68%	14.23%
6/15/15	0.00%	26.70%	12.50%	11.55%	22.42%	10.79%	0.00%	24.99%	21.26%	14.47%
6/22/15	0.00%	18.22%	17.31%	27.04%	12.77%	5.08%	0.00%	21.03%	21.83%	13.70%
6/29/15	52.66%	27.01%	8.80%	8.42%	5.88%	5.16%	0.00%	35.15%	22.56%	18.40%
7/6/15	49.89%	35.17%	10.65%	3.13%	19.07%	15.58%	32.19%	24.52%	17.32%	23.06%
7/13/15	51.37%	12.56%	5.49%	12.40%	17.25%	0.00%	33.40%	8.38%	28.72%	18.84%
7/20/15	50.23%	16.93%	9.42%	22.27%	7.38%	5.49%	29.87%	21.92%	12.03%	19.50%
7/27/15	48.50%	15.13%	16.12%	6.04%	8.80%	1.61%	34.03%	10.97%	0.00%	15.69%
8/3/15	54.91%	15.19%	8.34%	10.72%	12.75%	10.97%	26.94%	6.49%	9.71%	17.34%
8/10/15	54.02%	21.58%	6.17%	7.87%	12.68%	14.77%	43.46%	21.79%	19.56%	22.43%
8/17/15	53.46%	13.01%	5.25%	1.32%	11.10%	7.98%	33.01%	15.65%	9.13%	16.66%
Total	**34.59%**	**20.98%**	**9.72%**	**11.06%**	**12.08%**	**6.97%**	**19.41%**	**21.89%**	**15.63%**	**16.93%**

En el cuadro 4 se puede visualizar la relación entre los topicos de tweets y noticias de diario El Telégrafo con un promedio general de 24.57% considerando los retweets creados por terceros y un promedio de 16.93% para los tweets creados originalmente por cada líder. Los líderes políticos que sobresalen con más de un 30% en ciertas semanas por similitud con diario El Telégrafo son: Carlos Vera, Dalo Bucaram, Martha Roldos, Rafael Correa.

En la figura 7, se observa que en la semana del 06 de julio existe una mayor similitud de los tópicos tomados de diario El Telégrafo y los mensajes publicados en las redes sociales con un 23.06% en este caso en la plataforma Twitter y, aparecen temas como: marcha, dialogo, tiranía, sabatinas, presidente, Correa, proyecto, revolución, ciudadana, protesta, democracia, derecho, Guayas, patria, impuesto, guayaquileño.

Figura 7: Similitud de tópicos de tweets con noticias de Diario El Telégrafo



En este caso inferimos que la actividad en redes sociales va en sintonía con la coyuntura política que atravesaba el país durante esos días, porque al igual que los medios digitales los tweets hacen referencia a las protestas frente al proyecto de ley de reforma tributaria planteado por el ejecutivo.

7 Conclusiones

Los resultados obtenidos en nuestros experimentos nos permiten determinar que existe una relación entre los topicos que expresan los usuarios influyentes en el ámbito político de Ecuador y los temas que se tratan en los periódicos digitales universo.com y eltelegrafo.com.ec.

Es decir, se traslada a la plataforma de social media Twitter en nuestro caso los temas que discute la opinión pública y se refleja la coyuntura nacional por la que atraviesa el país. Se pudo observar que los tópicos de ciertos líderes políticos van en la línea de las publicaciones que realiza diario el Universo y diario El Telégrafo. Cabe señalar que el primero es un medio de comunicación privado y el segundo se presenta como un periódico público, es decir, que manejan dos líneas editoriales distintas. Esto permitió contrastar la información en función de los tópicos que manejan los discursos que presentan ambos diarios frente a los discursos que proponen los líderes políticos en las redes sociales.

Para validar la hipótesis planteada al inicio de este artıculo hemos medido la similitud entre documentos comparando cada uno de los tópicos extraıdos de los tweets contra el conjunto de tópicos extraıdos de las noticias publicadas por estos diarios digitales. Los resultados muestran un porcentaje promedio mayor al 50% para ambos diarios, lo que indica un grado de coherencia positiva entre Twitter y medios tradicionales de noticias en Ecuador. Es decir, que la hipótesis planteada inicialmente

es correcta, la extracción de tópicos relevantes de plataformas de social media es coherente con los tópicos publicados en los medios de comunicación en Ecuador.

Se podría decir que a través del uso de una plataforma de microblogging es posible obtener un punteo de los eventos de carácter político ocurridos en Ecuador durante un periodo determinado. Los resultados obtenidos también muestran que no es equivalente el volumen de actividad de un usuario frente a la atención que este recibe. Y reflejan que si es posible aplicar técnicas de aprendizaje de maquina no supervisado en este tipo de plataformas digitales para extraer tópicos de relevancia que se publican en medios de comunicación del país.

En un trabajo futuro se evaluará otras técnicas de extracción de tópicos de documentos con el objetivo de determinar el rendimiento y la eficacia en comparación con el modelo LDA. Adicionalmente se pretende complementar el análisis de los líderes políticos mediante la caracterización la audiencia [3] de manera que se pueda establecer un índice de influencia de cada líder considerando la interacción con su audiencia.

Referencias

[1] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 45–54.

[2] S. Asur, B. A. Huberman, G. Szabo, and C. Wang, "Trends in social media: Persistence and decay," *Available at SSRN 1755748*, 2011.

[3] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy." *ICWSM*, vol. 10, no. 10-17, p. 30, 2010.

[4] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Machine learning and knowledge discovery in databases*. Springer, 2011, pp. 18–33.

[5] J. Kulshrestha, M. B. Zafar, L. E. Noboa, K. P. Gummadi, and S. Ghosh, "Characterizing information diets of social media users," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[7] (2013) What happens in an internet minute?

[8] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 261–270.

[9] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval*. Springer, 2011, pp. 338–349.

[10] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Identifying influencers on twitter," in *Fourth ACM International Conference on Web Seach and Data Mining (WSDM)*, 2011.

[11] A. Hermida, "Twittering the news: The emergence of ambient journalism," *Journalism Practice*, vol. 4, no. 3, pp. 297–308, 2010.

[12] M. Schmierbach and A. Oeldorf-Hirsch, "A little bird told me, so i didn't believe it: Twitter, credibility, and issue perceptions," *Communication Quarterly*, vol. 60, no. 3, pp. 317–337, 2012.

[13] C. L. . Cuenca. (2015) Mapa del poder. [Online]. Available: http://www.mapadepoderecuador.com/

[14] Social baker estadısiticas y marketing en redes sociales. [Online]. Available: http://www.socialbakers.com/

[15] Twitter rest api para desarrolladores. [Online]. Available: https://dev.twitter.com/rest/public

[16] Dario el universo. [Online]. Available: http://www.eluniverso.com/

[17] Dario el telegrafo. [Online]. Available: http://www.telegrafo.com.ec/

[18] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[19] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 306–315.

[20] R. Rehuvek and P. Sojka, "Software framework for topic modelling with large corpora," 2010.