Metodología para extracción de tópicos relevantes de la red social Twitter

Rubén Carvajal[a], Carmen Vaca[a], Charlie Medina[a], César Madrid[a]

[a] Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral, Km. 30.5 Vía Perimetral, Guayaquil, Ecuador
rubancar@espol.edu.ec, cvaca@espol.edu.ec, chagemed@espol.edu.ec, cmadrid@espol.edu.ec

Resumen. Existe un rápido incremento en la producción de información y datos de manera virtual, debido a sitios de microblogging como Twitter, red social que produce en promedio 6,000 tweets por segundo, y hasta 500 millones de tweets al día. Razón por la cual esta y muchas otras redes sociales presentan una sobrecarga de contenidos, dificultando a los usuarios la identificación de tópicos de información por la gran cantidad de tweets hablando de diferentes temas. Debido a esta incertidumbre que perjudica a los mismos usuarios que han creado el contenido, se propone un método que a través de la selección de perfiles de usuarios expertos en deportes y política, infiere cuáles son los tópicos más representativos que han ocurrido en un marco de tiempo de 1 día. Esto se calcula tomando en consideración la cantidad de veces que este tópico ha sido mencionado por los expertos en sus timelines. Este experimento incluyó un dataset extraído de Twitter, que contiene 5,815 tweets referentes a deportes y 4,648 tweets referentes a política. Todos los tweets fueron obtenidos de timelines de usuarios seleccionados por los investigadores, que fueron considerados como expertos en sus respectivos temas debido al contenido de sus tweets. Los resultados muestran que la selección efectiva de los usuarios junto con el índice de relevancia implementado para los tópicos puede ayudar a encontrar con mayor facilidad tópicos importantes tanto en tema deportivo, como político.

Palabras Clave: Redes sociales, microblogging, twitter, tópicos, LDA.

1 Introducción

Los servicios de microblogging, en especial Twitter, cumplen una doble función: son un micrófono para las masas [1] y una fuente de información que complementa a los medios de comunicación tradicionales. Este fenómeno ha generado una nueva ecología en el consumo de noticias, en la que las plataformas de 'Social Media' toman cada vez mayor protagonismo [2]. En Twitter específicamente, cada vez que el usuario ingresa a su cuenta, se encuentra con un 'feed' de posts que resulta de la combinación de todo el contenido producido por los usuarios a los cuales se está siguiendo. Estos usuarios a quienes se sigue, denominados 'followees', pueden ser amigos, figuras de relevancia local o regional, compañías, cadenas de noticias. A medida que el número de 'followees' aumenta, la cantidad de información que el usuario debe procesar aumenta también y se genera el problema de sobrecarga de información [3]. Estudios previos han analizado extensivamente el problema de la influencia o 'authorities users' en Twitter [4,5]. Es decir, la capacidad de un usuario de producir información valiosa

para sus seguidores donde el valor se mide por métricas tales como el número de re-tweets recibidos a un determinado tweet o la cantidad de debate generado por el mismo. Sin embargo, los consumidores de noticias que enfrentan el problema de sobrecarga de información muchas veces no tienen problemas en identificar a usuarios que constituyen 'authorities' en temas nacionales.

Los retos que enfrentan se relacionan más bien con la capacidad de obtener un resumen de su 'feed' de noticias donde sea posible ordenar por relevancia los posts de forma automática. En este trabajo, se introduce una metodología para convertir un 'feed' de tweets en un conjunto de tópicos con un puntaje asociado para identificar la relevancia del mismo. Los tópicos en el contexto de procesamiento de lenguaje natural son una lista de palabras que resume un evento o una serie de noticias. El proceso empieza con la selección de los perfiles de usuario expertos en cada una de las áreas que serían analizadas tomando como caso de estudio deportes y política. Se continúa con el filtrado y pre-procesamiento de los datos. La siguiente etapa es la extracción de tópicos usando un modelo probabilístico.

Finalmente como contribución de este trabajo, se diseñó una métrica cuantitativa que permite identificar los tópicos relevantes contenidos en un conjunto de tweets recibidos como entrada. El artículo está estructurado como sigue. En la Sección 2 se presenta la revisión de literatura. En la Sección 3 se explica la configuración de los experimentos en los que se procesa el 'feed' de tweets para la extracción de los tópicos. En la Sección 4 se discuten los resultados obtenidos para un dataset proveniente de tweets generados en Ecuador. Finalmente, en la Sección 5 se presentan las conclusiones y el trabajo futuro.

2 Revisión de Literatura

Uno de los algoritmos tradicionales usados en la inferencia de tópicos es LDA [6], su probada eficiencia ha hecho de este método la base de muchos estudios en la detección de tópicos, la característica principal de este método es la fácil interpretación de su resultado, que son conjuntos de palabras con probabilidad de pertenecer a tópicos descubiertos. Hasta el momento la mayor parte del trabajo se centra en crear aplicaciones [7] dinámicas que recolectan información online y presentan tópicos a su salida, otra corriente de investigación es la formulación de métodos y pre-procesamiento especial de los datos [8] para mejorar los resultados y ejecución de LDA. Todos los trabajos descritos con anterioridad tienen en común el uso de documentos de una longitud mayor a 140 caracteres, lo cual es una restricción para nuestro estudio dado el uso de la plataforma Twitter, aunque existe atención prestada a esta temática [9], no hay un marco de trabajo definido que emplee métodos estándar como LDA en la detección de tópicos teniendo la restricción del tamaño del documento, para este problema Yangqiu Song et al. [10] propone mejoras a la salida de LDA, el cual devuelve tópicos relevantes asociados a temáticas que hayan sido especificadas por el investigador. Sin embargo en nuestro trabajo se busca una forma sencilla de identificar si un tópico en general es relevante con respecto a todos los que se obtienen a la salida de LDA, basado únicamente en los documentos (tweets) de entrada del proceso, para esto se siguieron lineamientos conocidos en el procesamiento

de texto y finalmente se introdujo un índice que nos ayuda a discernir la importancia de un tópico con respecto a todos los generados por LDA.

3 Metodología

3.1 Selección de usuarios

Para esta investigación se utilizó un dataset de 5,815 tweets referentes a deportes y 4,648 tweets referentes a política, los cuales fueron obtenidos de un grupo de usuarios de Twitter que sirvieron como grupo semilla. Se formó una lista de 50 usuarios para cada uno de los tópicos. Por juicio de experto se escogió de cada grupo 25 usuarios cuyos perfiles cumplan con requisitos como actividad constante, contenido enfocado en el área al que pertenecen, y calidad en la forma de escritura.

Entre los usuarios semilla escogidos para el tema de política, se encuentran varios politólogos, trabajadores del gobierno y del sector privado, figuras políticas importantes como excandidatos o actuales funcionarios públicos que desempeñan cargos de dirección en instituciones tales como alcaldías, ministerios y presidencia. Mientras que entre los usuarios semilla escogidos para el tema de deportes, se encuentran comentaristas deportivos, administradores y presidentes de clubes de fútbol, jugadores profesionales y usuarios relacionados a programas o revistas deportivas. Para cada uno de los usuarios seleccionados, se extrajo un conjunto de tweets para el periodo de tiempo comprendido en este estudio así como la fecha y hora del tweet.

3.2 Recolección y pre-procesamiento del dataset

La recolección de los tweets se la realizó haciendo uso de REST-API de Twitter junto con Python, para el inicio del proceso se definieron dos tópicos específicos (deportes y política) sobre los cuales probar nuestro método, se llegó a la conclusión que los tópicos mencionados con anterioridad son completamente excluyentes y abarcan gran cantidad de los datos producidos por los usuarios de twitter, acorde a lo mencionado por Kathy Lee et al. [11] se dice que la categoría de “Sports” abarca la mayor cantidad de tópicos dentro de la red social, seguida de las categorías “Other”, “Other news”,“Music”, “Tv &Movies”,“Tech” y ‘Politics” como las más relevantes, si bien la política no está en el top de esta lista, la hemos escogido dado el panorama nacional y los eventos de índole política que se presentan en el país, como demuestra la experiencia los eventos que incluyen expresiones populares de las personas siempre se ven reflejados en las redes sociales, esperamos de buena forma verificar hechos puntuales (tópicos) que aparezcan con el análisis de los tweets.

Una vez definidos los usuarios expertos en ambos tópicos, procedemos a recolectar para cada usuario el máximo número de tweets que nos permite REST API de Twitter en una única consulta, se realiza la recolección haciendo una consulta por usuario y se obtienen 5,815 y 4,648 tweets para los usuarios vinculados al tópico deportes y política respectivamente. Tanto la figura 1 como la 2 muestran la agrupación temporal de los tweets extraídos del timeline de cada uno de los usuarios expertos para ambos tópicos,

en ambos se observa una concentración mayor de los tweets en el mes de agosto de 2015 (2015 08) como se esperaba, en el tópico deportes vemos datos que incluyen fechas desde el año 2012, se concluye la inclusión de algunos usuarios poco activos o algunas cuentas desactualizadas, ya que en los últimos tweets de su timeline presentan tweets viejos en el tiempo, un aspecto a tener en cuenta para una revisión posterior de nuestro trabajo, sin embargo se procedió a usar todos los tweets producidos en el año 2015 para ambos tópicos, que representan el 89,56% y 100% de los tweets recolectados para cada tópico, deportes y política respectivamente.

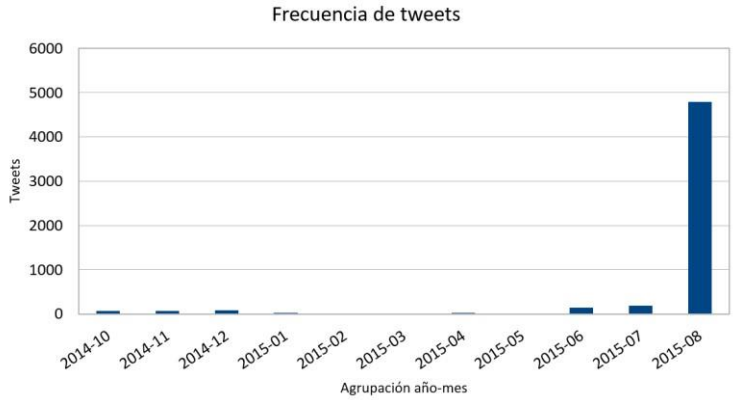


Fig. 1. Frecuencia de tweets de usuarios expertos en deportes, se observa que la mayor parte de los tweets se concentran en el mes 08 (Agosto) de 2015

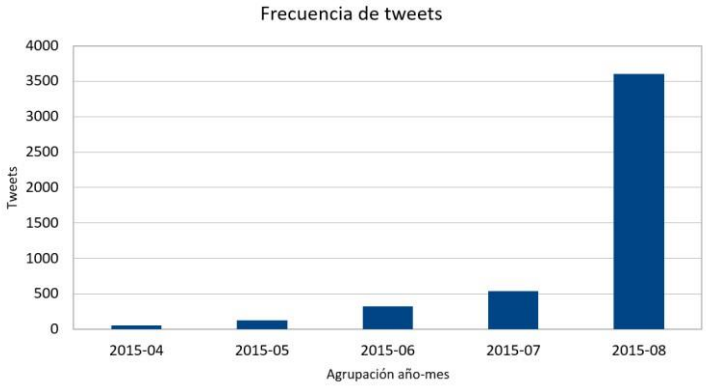


Fig. 2. Frecuencia de tweets de usuarios expertos en política, se observa que la mayor parte de los tweets se concentran en el mes 08 (Agosto) de 2015.

Finalmente se analiza el JSON y se extrae el tweet, del cual se eliminan saltos de línea, signos de puntuación y se corrigen secuencias de caracteres relevantes para nuestro estudio. Ej.: todas las cadenas con sucesiones de ‘ja’ o ‘ha’ son mapeadas a ‘jaja’, así ‘jajajaj’, ‘jajajajaja’ o ‘hahaha’ son traducidas a ‘jaja’.

Además se separa hashtags cortando por la letra mayúscula (estándar de escritura dentro de Twitter), así la cadena #vamosMiSeleccción será traducida a ‘vamos mi selección’, y en caso de no encontrar mayúsculas quedará igual. Este proceso se realiza basado en el trabajo de Mitchell et al. [12], en el cual se identifican palabras importantes a consideración de los investigadores y se corrigen o eliminan para obtener mayor exactitud en la ejecución de sus algoritmos. Se extrae además del tweet, el username y el timestamp de cada tweet y se procede a guardar todo en una base de datos, para su posterior análisis.

3.3 Simplificación de texto

La tarea de simplificación del texto se la llevó a cabo con el uso de dos librerías: NLTK Steven Bird et al. [13] para la eliminación de stop words y el software provisto por el grupo de investigación CliPS De Smedt T.& Daelemans [14] que incluye métodos para detección de las formas del lenguaje (pronombres, sustantivos, verbos, etc.) dentro de una oración y permite además la simplificación de sustantivos a singulares y verbos a su forma más básica; lo cual es importante ya que palabras como ‘jugando’ o ‘jugado’ se ven reducidas al verbo ‘jugar’ o sustantivos como ‘hinchas’ a ‘hincha’, evitado procesamiento posterior y mejorando significativamente el trabajo de los algoritmos como
LD
A

3.4 Descripción de parámetros para el proceso de LDA

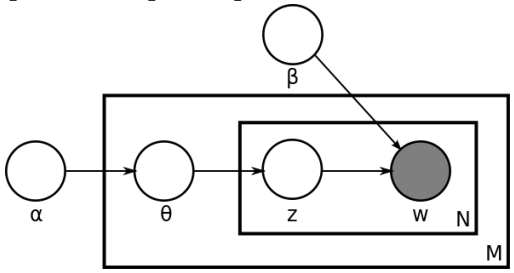


Fig. 3. Representación del modelo LDA, *M* representa el *número de documentos*, y *N* representa el *número de palabras por documento.*

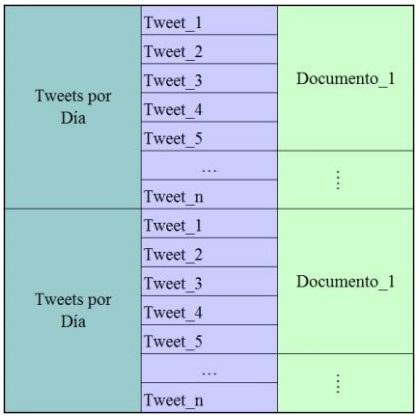
En la (figura 3) M representa el número de documentos de la colección y N el número de palabras en cada uno de los documentos, α es el parámetro Dirichelt de la distribución de tópicos por documento y β es el parámetro Dirichelt de la distribución de palabras por tópico, en nuestro trabajo se utilizan en todas las ejecuciones los

valores de 0.1 y 0.01 para α y β respectivamente, la teoría acerca de LDA [6] nos dice que para valores pequeños de α se restringe el contenido de los documentos, lo que quiere decir que cada documento (tweet) es una mezcla de unos pocos tópicos o inclusive uno solo, así mismo valores pequeños de β indica que cada uno de los tópicos es una mezcla de unas pocas palabras, lo cual es un ajuste óptimo al trabajar con documentos de no más de 140 caracteres y con mucha probabilidad de que cada documento (tweet) trate únicamente de un solo tópico.

3.5 Detección de tópicos

3.5.1 Preparación de documentos

La entrada del algoritmo LDA requiere la matriz bag of words, se procedió a construir una matriz diferente por día. El objetivo a alcanzar al escoger un día como unidad temporal es detectar tópicos específicos y eventos relevantes. Como se demuestra en el trabajo de Becker et al. [15], la asociación temporal juega un rol predominante a la hora de explorar tópicos a partir de información publicada en Social Media. Para este trabajo los documentos (tweets) se agruparon en dos configuraciones diferentes (figura 4) para posterior comparación de los resultados, primero se tomó cada tweet como un documento individual y segundo se realizaron agrupaciones por usuario haciendo que cinco tweets consecutivos en el timeline de un mismo usuario sea un solo documento, más adelante se explicarán los resultados de ambas ejecuciones.



. **Fig. 4.** Niveles de agrupación con que se probó el algoritmo LDA

3.6 Relevancia de tópicos

El índice de relevancia para cada tópico que se formula a continuación es determinado por los mismos datos usados por LDA para generarlos, la fórmula tiene en cuenta el

aporte de cada una de las palabras al tópico al que pertenece y el peso de dicho tópico con respecto a los documentos que se usaron para generarlo.

$$\text{índice}_{\text{relevancia}} = P \sum_{w=1}^{k} N(w)p(w) . \quad \textbf{(1)}$$

N = número de documentos de la colección en los que aparece la palabra w. p = fracción de documentos de la colección en los que aparece la palabra w, tomando como base el total de documentos en los que aparece al menos una palabra del tópico.

P = fracción de documentos de la colección que incluyen palabras del tópico, esta vez tomando como base el total de documentos usados en el análisis. Como se observa, tenemos dos partes importantes en nuestro índice, lo primero

$$\sum_{w=1}^{k} N(w)p(w) . \quad \textbf{(2)}$$

la sumatoria de la frecuencia ponderada de cada una de las palabras que conforma el tópico, lo que nos da una idea del peso de un tópico determinado por sus propias palabras, y lo segundo, el valor P nos da una idea del peso de dicho tópico con respecto a todos los documentos incluidos en la colección, finalmente el producto de ambos valores nos da como resultado un índice de relevancia del tópico, el cual podemos usar para decidir de forma sencilla los tópicos más relevantes a la salida de un algoritmo como LDA, a continuación exponemos un ejemplo de cómo se calcula el índice:

En el cuadro 1 se muestra una entrada de 5 documentos para LDA y uno de los tópicos obtenido por el mismo incluye las palabras ‘barcelona’ y ‘river’, procedemos a calcular el índice de relevancia de la siguiente forma:

En la proxima sección presentamos los resultados obtenidos de la extracción de tópicos.

Tabla 1. Ejemplo de documentos para LDA

Tweets
barcelona ganó a **river** en el monumental
emelec venció a liga en casa blanca
barcelona y emelec superaron con facilidad a sus rivales
barcelona superó a **river** por la mínima diferencia

$$N(\text{barcelona}) = 3,\ p(\text{barcelona}) = \frac{3}{3}. \quad \textbf{(2)}$$

$$N(\text{river}) = 2,\ p(\text{river}) = \frac{2}{3}. \quad \textbf{(3)}$$

$$P = \frac{3}{5}. \quad \textbf{(4)}$$

$$\text{índice}_{\text{relevancia}} = \frac{3}{5} \times \left[3 \times \frac{3}{3} + 2 \times \frac{2}{3}\right] = 2{,}6\ . \quad \textbf{(5)}$$

4. Resultados

4.1 Evaluación de la configuración de documentos

Como se explicó con anterioridad, se realizó dos configuraciones para los documentos de entrada a LDA, la primera haciendo que cada tweet sea un documento y segundo haciendo que cada documento sea la unión de cinco tweets consecutivos en el timeline de un mismo usuario, como resultado de la ejecución del algoritmo se visualizan resultados muy parecidos, aunque en la ejecución de LDA con la configuración de tweets agrupados se visualizan resultados más difusos para ciertos tópicos, incluyendo palabras que podrían quedar afuera del tópico fácilmente, por ejemplo para el día 19 de agosto de 2015, los tópicos más relevantes para ambas configuraciones de acuerdo a nuestro índice son:

Configuración un documento-un tweet: guaira sudamericana católica copa partido.

Configuración un documento-cinco tweets: sudamericana copa guaira católica primera.

Como se observa la palabra primera en la segunda configuración no representa mayor significado para el tópico, a pesar de esto una ventaja de la agrupación por tweets consecutivos es la disminución en el impacto de los tweets repetitivos tipo spam, ya que al ser agrupados en un mismo documento el impacto de esas palabras para un tópico disminuye. Finalmente se muestran en el cuadro 2 los resultados de la ejecución de

LDA para 11 días consecutivos con la configuración de un tweet-un documento usando el tópico deportes, cabe mencionar que se ejecutó el algoritmo LDA para los días en los que los documentos superaban un valor límite de 50 debido a que por debajo de este valor los tópicos resultantes de LDA no son coherentes. La agrupación por tópicos finalmente no mejoró de mayor forma los resultados en la ejecución, en este trabajo se presentan los resultados (Cuadro2) solo para la ejecución de LDA con la configuración un tweet-un documento, los resultados de la ejecución con la configuración de tweets agrupados pueden ser consultados utilizando el URL incluido al final del presente trabajo.

4.2 Clasificación por tópicos

La ejecución de los algoritmos para identificación de tópicos como LDA requiere como parte de sus parámetros de entrada el número de tópicos, por lo que hemos ejecutado LDA con este parámetro establecido en 5, 7 y 10, se procedió a comparar los resultados y se decidió que 7 es un número ideal dada la agrupación por días, establecer los tópicos a 10 nos da como resultado muchos tópicos irrelevantes y por otro lado establecerlo a 5 nos produce tópicos difusos, ya que se observan claramente combinaciones entre tópicos. A continuación, exponemos los resultados del análisis con el respectivo índice de relevancia para cada tópico y estableciendo a 7 el número máximo de tópicos para LDA. El cuadro 2 muestra el resultado del proceso de ejecución de LDA, junto con el índice de relevancia para cada tópico, cabe mencionar que los tópicos (tres por día) son los tres clasificados con mayor índice de los siete que se obtienen en la salida de LDA, se visualizan titulares repetidos entre algunos de los tópicos porque refieren al mismo suceso, así también tenemos algunos tópicos en color azul, los cuales reflejan eventos que ocurrieron en esos días pero que en la página de diario El Universo (tabla 3) no reflejan un titular acorde al tópico obtenido.

Tabla 2. Tópicos generados por LDA para la configuración un tweet-un documento.

Tópicos por día	Índice	Titular
Día: 2015-08-19		
guaira sudamericana católica copa partido	54.63	T1
cambio ser deber futbol salir	5.25	
tiempo minuto guaira segundo ucatólica	34.22	T1
Día: 2015-08-18		
bsc barcelona almada poder ir	97.18	T2
video nicaragua hoy madrid partido	15.26	
united manchester video champion bruja	16.02	T3
Día: 2015-08-17		
emelec dcuenca min campeonatonacional ingresar	14.90	T4
vía ir jugador lui caravana	139.63	
seguir gracias futbol participar alentar	133.98	
Día: 2015-08-16		
ser futbol deber alemán ingresar	16.50	
vivo aquí síguelo partido riv	17.50	
river barcelona ec campeonatonacional aucas	63.06	T5
Día: 2015-08-15		
ecuador ir bien iniciar to2015	12.96	T6
copa sornoza pilsener 4 vivo	10.10	
ser deber futbol runa idv	17.63	T7
Día: 2015-08-14		
ir 4 españa vía poder	34.09	
to2015 ecuatoriano 5 cotopaxi m	251.30	T8
copa sudamericana emelec gol primera	15.23	T9
Día: 2015-08-13		

emelec huanuco leon copasudamericana sudamericana	52.84	T10
partido sudamericana emelec cse capwell	57.01	T10
gol partido equipo copa jugar	20.60	
Día: 2015-08-12		
sudamericana copa ir river campeon	25.54	T11
ldu partido sudamericana loja 12	15.06	T12
copa final santa liga loja	9.54	T12
Día: 2015-08-11		
barcelona sc noticia campeón fin	21.09	T13
sevilla barcelona super uefa cup	39.52	T13
fcbarcelona año ir ganar sport	29.20	T13
Día: 2015-08-09		
nacional barcelona falta ta blanco	42.98	T14
nacional barcelona partido preciar perlaza	21.96	T14
gol aucas juego minuto ldu	5.81	
Día: 2015-08-08		
minuto ver gracias visita 8	3.71	
chito gol ecuador victoria ir	16.77	T15
gol quito river minuto ecuador	8.24	T16

Tabla 3. Titulares de diario El Universo

Titular	Titular en diario El Universo
T1	Universidad Católica cayó 1-0 ante Deportivo La Guaira y está fuera de la Sudamericana
T2	El técnico Guillermo Almada aclaró que no se va de Barcelona
T3	Manchester United gana 3-1 al FC Brujas por la fase previa Champions League
T4	Emelec se llevó tres puntos de Cuenca y sigue líder del Campeonato Ecuatoriano
T5	Barcelona perdió 0-1 ante River Ecuador en el Monumental
T6	Juegos parapanamericanos Toronto 2015 (to2015)
T7	Independiente recupera terreno al vencer 2-4 a Mushuc Runa
T8	Participación de ecuador en juegos panamericanos Toronto 2015
T9	Participación de Emelec en copa sudamericana
T10	León de Huánuco cayó 3-1 ante Emelec de local por Sudamericana
T11	River Plate se impuso 3-0 al Gamba Osaka y es campeón de la Copa Suruga Bank
T12	Liga de Loja empató 0-0 con Independiente de Santa Fe por la Copa Sudamericana
T13	FC Barcelona venció 5-4 al Sevilla y conquistó la Supercopa de Europa
T14	Barcelona cayó 1-0 ante El Nacional y perdió el liderato del campeonato
T15	El ecuatoriano Marlon 'Chito' Vera logra su primera victoria en la UFC
T16	River Ecuador aplastó 3-1 a Deportivo Quito en el estadio Chucho Benítez

5. Conclusiones y trabajo futuro

Esta investigación propone un método sencillo de implementar y que pueda ser incluido en cualquier aplicación que requiera una exploración de tópicos. Una vez escogidos los usuarios semilla y extraídos los tópicos se observó que utilizando un índice, sencillo de calcular, es posible determinar los tópicos más relevantes a partir de aquellos extraídos usando LDA. Esto se confirma dado que la mayor parte de tópicos resultantes pudieron ser asociados a titulares de diario El Universo producidos

para los mismos días en los que se ejecutó el método. La facilidad en el cálculo de este índice y su probada eficiencia se evidencia tanto para los tópicos resultantes en el tema de deportes (cuadro2) como en el tema de política (link al final de la página), la discriminación de los tópicos no resultantes se observa coherente dado que se eliminan tópicos con palabras que juntas resultan en temas demasiado ambiguos. Finalmente, dos claves para que este proceso mejore es la selección cuidadosa de los usuarios expertos en cada uno de los temas que necesitemos analizar y el procesamiento cuidadoso, como eliminación de stopwords, simplificación del lenguaje como reducción de verbos a su forma básica y singularización de sustantivos, procesos que se aplicaron en el presente trabajo pero que pueden ser aún mejorados. Se espera que con la mejora en la detección de spam en la etapa inicial del proceso se puedan obtener excelentes resultados con una mayor precisión en los términos que conforman el tópico, además de esto se puede usar el índice de relevancia de cada uno de los tópicos y medir su cambio con respecto a franjas horarias e identificar la franja temporal en la cual un tópico alcanzó una mayor relevancia, y dado que el método acepta como entrada usuarios de la red social twitter se puede establecer y probar su eficiencia en escenarios más diversos y no solo dirigido a revelar tópicos como política y deportes. Todo el recurso adicional del proyecto puede ser consultados en el siguiente link: https://goo.gl/yEovhc.

Referencias

1. Dhiraj Murthy et al. Twitter: Microphone for the masses? Media Culture and Society,33(5):779, 2011.
2. Nic Newman, William H Dutton, Grant Blank: Social media in the changing ecology of news: The fourth and fifth estates in britain. InternationalJournalofInternetScience,7(1):6–22, 2012.
3. Avery E Holton Hsiang Iris Chyi: News and the overloaded consumer:Factors influencing information overload among news consumers. Cyberpsychology, Behavior, and Social Networking, 15(11):619–624, 2012.
4. Eytan Bakshy, Jake M Hofman, Winter A Mason, Duncan J Watts: Identifying influencers on twitter. In Fourth ACM International Conference on Web Seach and Data Mining (WSDM), 2011.
5. Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 45–54. ACM, 2011.
6. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.
7. Loulwah AlSumait, Daniel Barbar´a, and Carlotta Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, pages 3–12. IEEE, 2008.

8. Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 569– 577. ACM, 2008.
9. Liangjie Hong, Brian D Davison: Empirical study of topic modeling in twitter. In Proceedings of the First Workshop on Social Media Analytics, pages 80–88. ACM, 2010.
10. Yangqiu Song, Shimei Pan, Shixia Liu, Michelle X Zhou, and Weihong Qian. Topic and keyword re-ranking for lda-based topic modeling. In Proceedings of the 18th ACM conference on Information and knowledge management, pages 1757–1760. ACM, 2009.
11. Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, Alok Choudhary: Twitter trending topic classification. In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pages 251– 258. IEEE, 2011.
12. Mitchell, L, Frank, M, Harris, K, Dodds, P, Danforth, C: The Geography of Happiness: Connecting Twitter Sentiment and Expressions, Demographics, and Objective Characteristics of Place. (2013).
13. Steven Bird, Ewan Klein, Edward Loper: Natural language processing withPython. .O'Reilly Media, Inc.", 2009.
14. Tom De Smedt and Walter Daelemans. Pattern for python. The Journal of Machine Learning Research, 13(1):2063–2067, 2012.
15. Hila Becker, Mor Naaman, Luis Gravano. Learning similarity metrics for event identification in social media. In Proceedings of the third ACM international conference on Web search and data mining, pages 291–300. ACM, 2010.