Sophia una herramienta para la construcción y análisis de casos noticiosos en la enseñanza del periodismo

Luis Cárcamo-Ulloa
Instituto de Comunicación Social
Universidad Austral de Chile
Campus Teja, Valdivia, Chile
+56632221574
lcarcamo@uach.cl

Matthieu Vernier
Instituto de Informática
Universidad Austral de Chile
Campus Miraflores, Valdivia Chile
+56632221427
mvernier@inf.uach.cl

Eliana Scheihing García,
Instituto de Informática
Universidad Austral de Chile
Campus Miraflores, Valdivia Chile
+56632221427
escheihi@inf.uach.cl

Matías Aravena
Instituto de Informática
Universidad Austral de Chile
Campus Miraflores, Valdivia Chile
+56632221427
instituto@inf.uach.c

Javier Pérez
Instituto de Informática
Universidad Austral de Chile
Campus Miraflores,Valdivia Chile
+56632221427
instituto@inf.uach.cl

ABSTRACT

The development of the Sophia platform allows to support journalism education and social communication research. It collects in real time news articles published across Twitter by 290 chilean mass communications media. The set of media is heterogeneous in terms of a) political orientation, b) geographic location in Chile, and c) initial communication canal (radio, television, print/internet newspaper).

The "Social Networks and Communication Media" project has developped the online software Sophia which a) collects chilean news articles, b) stores all article headlines, c) provides access to the content of each article, d) offers the possibility to realize textual queries on a topic and e) to define "news cases" used to group together past data or future data about an event or a topic. It also allows f) to visualize data in the forms of histograms or pie charts, g) aggregating data according to each media or each media corporation, and h) to export required data in csv files.

RESUMEN

La plataforma Sophia es un desarrollo que permite apoyar la enseñanza del periodismo y la investigación en comunicación social. Almacena las noticias emitidas en Twitter por 290 medios de comunicación de masas chilenos. El conjunto de medios seguidos es un catastro heterogéneo en a) sus orientaciones políticas, b) geográficamente distribuidos por todo el territorio nacional y c) con orígenes mediales diversos (radios, televisiones, periódicos tradicionales y digitales).

El proyecto "Redes Sociales y Medios de Comunicación" ha desarrollado en una de sus líneas de investigación aplicada el software en línea Sophia. Se trata de una herramienta que a)colecta las noticias de 290 medios chilenos, b) almacena todos sus titulares, c) permite acceder al contenido de cada noticia en extenso, d) ofrece la posibilidad de realizar búsquedas sobre un tema e) a partir de las búsquedas se puede definir un "caso periodístico" de investigación y recolección de datos pasados (desde enero de 2017) y activar la colección a futuro, f) visualizar los datos colectados en histogramas y gráficos de torta, g) agrupar los datos por cada medio de comunicación o grupo mediáticos y h) exportar el listado de noticias de una caso en un archivo csv.

Categories and Subject Descriptors

K.3.1 [Computers and Education]:
Computers Uses in Education.

H.5.2. [Information interfaces and presentation]:
Datavisualización.

General Terms

Human Factors, Algorithms, Visualization, Web mining.

Keywords.

Enseñanza del Periodismo, Tratamiento automático del Lenguaje. DataVisualización, Análisis de prensa.

1. INTRODUCCIÓN

La enseñanza del periodismo y la investigación en comunicación social merecen una especial preocupación ante los cambios en el ecosistema mediático con la multiplicación de la información propiciada por internet y las redes sociales.

Si bien la diversificación de medios de comunicación alternativos encuentra en Internet nuevas herramientas para la libertad de expresión y el pluralismo, no es menos cierto que para profesionales de comunicación y analistas es difícil acceder de forma compilada y eficiente a lo que se puede decir sobre un tema. También, en forma cada vez más frecuente, ocupan espacios en redes sociales y prensa online las denominadas *fake news o noticias falsas*. Dicha situación es observada con desconfianza por las nuevas audiencias [5] y puede implicar cambios en el medio periodístico.

2. OBJETIVOS

General: Apoyar la enseñanza del periodismo y la investigación en comunicación social.

Específicos: a) Facilitar procesos de búsqueda de información noticiosa en medios de comunicación chilenos, b) Generar visualizaciones simples que permitan explicar flujos y volúmenes de información y c) Crear casos noticiosos que se constituyan en corpus de análisis exhaustivos o *background* para periodismo de investigación

3. INTERFAZ A USUARIO

En el presente apartado se presentan las principales interfaces o pantallas que permiten la interacción con Sophia.

La figura 1 muestra la interfaz general de búsqueda de información. Se llega a ella luego de una suscripción simple con Facebook o correo electrónico.

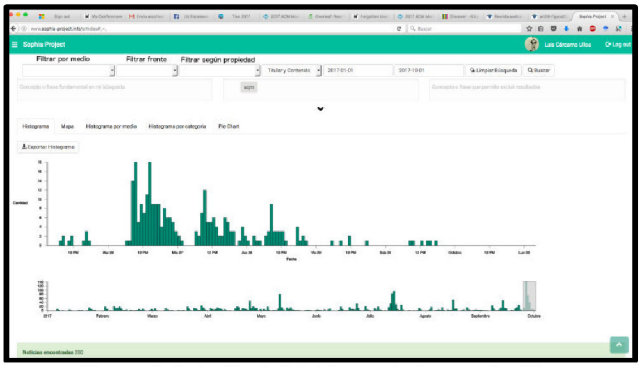

Figura 1. Interfaz inicial luego de la suscripción

La figura 2 presenta tres filtros que Sophia ofrece al usuario. **Filtrar por medio:** Permite seleccionar las noticias publicada por cada medio de prensa ingresado en la plataforma. **Filtrar por frente noticioso:** Selecciona una categoría temática asignada a la noticia (accidentes, deportes, ecología, economía, entretenimiento, judicial, política, salud, tecnología o educación y cultura). **Filtrar por propiedad o grupo de medios**: Permite obtener solo las publicaciones de un grupo de medios de prensa, ya que un medio de prensa puede pertenecer a un grupo que tiene asociado otros medios de prensa.

Figura 2. Detalle de los filtros

La figura 3 muestra las opciones para seleccionar la búsqueda en **una parte de la noticia:** Este filtro permite buscar el conjunto de palabras claves en el titular, el contenido o en toda la noticia. **Fecha de inicio**: Permite al usuario indicar desde qué fecha buscar publicaciones. **Fecha de término**: Similar al campo anterior, sólo que indica hasta qué fecha debe filtrar las publicaciones. **Limpiar búsqueda**: Permite a un usuario limpiar los campos ingresados. **Buscar:** Al hacer *click* en este botón, se envía la búsqueda y se recibe una respuesta con las publicaciones encontradas.

Figura 3. Detalles de una búsqued

La figura 4 corresponde a los campos de búsqueda. **Concepto o frase fundamental en mi búsqueda**: Corresponde a las palabras o frases que deben aparecer inexorablemente en la búsqueda del medio de prensa, se descartan todo el resto de publicaciones que no contengan las palabras o frases ingresadas. **Concepto o frase importante en mi búsqueda**: El conjunto de palabras o frases ingresadas en este campo, dará mayor puntaje de relevancia a las noticias que contengan estas en su contenido, estas palabras pueden o no estar en el cuerpo de la noticia. **Concepto o frase que permite excluir resultados:** Permite descartar todas las publicaciones que contengan las palabras o frases ingresadas en este campo.

Figura 4. Ingresar palabras claves

La figura 5 muestra como ante cualquier búsqueda se dibujan automáticamente dos histogramas. El primero da cuenta del ciclo reciente de noticias publicadas y el según de todo el periodo de búsqueda. Estas visualizaciones permiten apreciar el ciclo de un caso noticioso observando cuándo se habló más sobre nuestro tema.

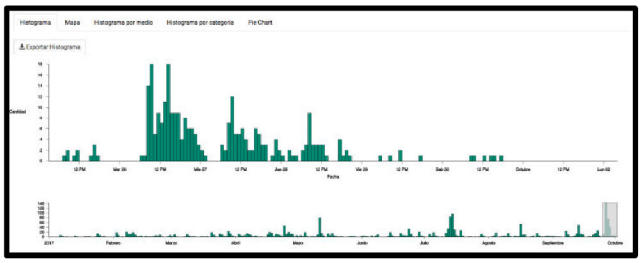

Figura 5. Histogramas

La figura 6 muestra cómo se presentan las noticias resultantes de la búsqueda. Cada una con la fecha de emisión, una fotografía (si en el medio original fue acompañada de una), la identificación del medio emisor y el titular de la noticia. Además, explicita un criterio de relevancia que se calcula en base a un algoritmo de similitud basado en el cálculo de Tf-Idf (del inglés *term frequency - inverse document frequency*) el cual es proveído por el motor de búsqueda implementado en el desarrollo de la plataforma.

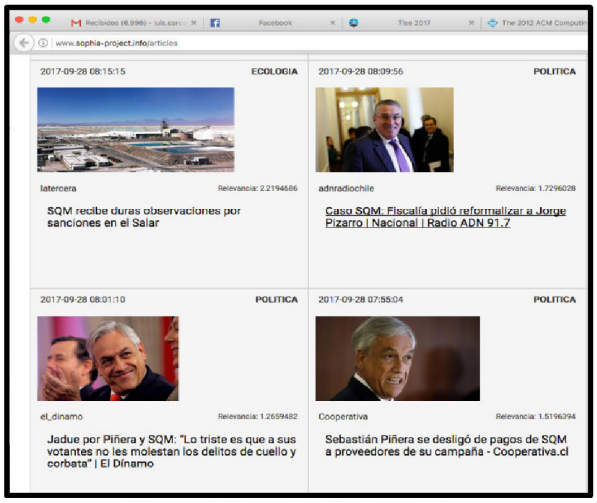

Figura 6. Presentación de noticias

La figura 7 corresponde a: **Noticias encontradas**: es decir el número total de noticias que coinciden con el criterio de búsqueda ingresado por el usuario. **Método de ordenamiento:** que permite ordenar el resultado de la búsqueda por *tiempo* (orden cronológico) o, también, permite ordenar por *relevancia*, del documento a partir de la búsqueda realizada. **Crear Caso Noticioso**: Permite a un usuario crear un *Caso Noticioso*, que guarda la búsqueda y tiene la opción de seguir colectando noticias (ver detalle más adelante) **Exportar resultado:** Permite a un usuario de exportar el resultado de la búsqueda (principalmente el mismo contenido presentado en la plataforma) a diferentes formatos, con el objetivo que estos datos puedan ser utilizados en otras plataformas o contextos.

Figura 7. Resultados y creación de caso noticioso

Funciones especiales para analistas

Las siguientes imágenes explican dos posibilidades creadas especialmente para usuarios analistas o que se encuentren desarrollando actividades de periodismo de investigación.
La figura 8 corresponde a la posibilidad de **Crear caso noticioso**, que consiste en guardar una búsqueda, atribuirle un nombre e iniciar una recolección automática de noticias relacionadas con esas palabras claves.

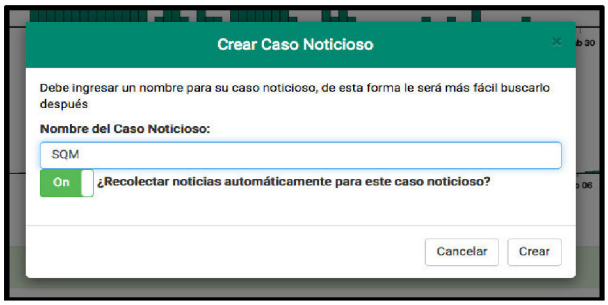

Figura 8. Cuadro de diálogo para creación de caso noticioso

La figura 9 se remite a la posibilidad de **Exportar datos** en un archivo CSV, JSON o TXT. Esta funcionalidad permite reutilizar los datos obtenidos para futuros análisis de contenido, discursivo o aplicar estrategias de lingüística de corpus.

Figura 9. Exportar datos

Visualizaciones de información

Una de las posibilidades naturales que ofrece Sophia es la visualización inmediata de los datos cuantitativos referidos al *número de informaciones que hablan de un tema de interés y la* composición de ese corpus.
La figura 10 muestra como para cualquier ejercicio de búsqueda Sophia ofrece la posibilidad de observar en un **pie chart** la distribución de noticias. Detallando los medios que conforman el 50% del corpus total colectado.

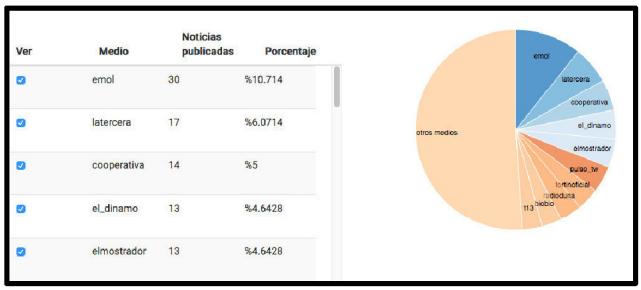

Figura 10. Visualización de torta

Una segunda visualización tiene que ver con una presentación de **histogramas** que representan las variaciones en un periodo de tiempo, pero además permite observar como los distintos medios componen el total de emisiones diarias sobre las palabras claves que componen el caso. El histograma además puede representarse en cantidad o porcentaje de emisiones.

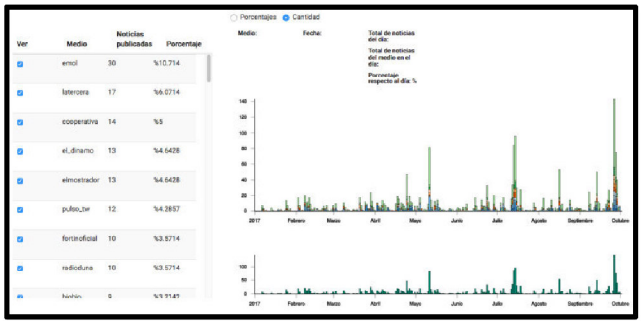

Figura 11. Visualización de histograma

4. VALOR AGREGADO PARA EL APRENDIZAJE

Desde la educación para el pensamiento crítico, Paul y Elder [7] remarcan que sin competencia alguna en la cultura de la información, los estudiantes no pueden convertirse en personas educadas porque no sabrán cuál información aceptar y cuál rechazar. Es el pensamiento crítico el que provee las herramientas para evaluar la información. Dicha competencia resulta especialmente sensible en la formación de profesionales del periodismo. Los profesionales de la comunicación social son un colectivo que coadyuvan a la formación de la opinión pública. Teun Van Dijk [12] releva la importancia de periodistas, profesores y políticos son agentes relevantes en la construcción de las opiniones de la ciudadanía.

Johnson y Morris [3,4], sintetizan que la ciudadanía crítica posee en realidad dos dinámicas: a) Busca la racionalidad científica para

analizar los datos del pasado; b) Recupera la subjetividad y valora la singularidad de cada individuo para construir activamente su propio pensamiento y sus acciones presentes. Barranquero-Carretero y Lema-Blanco [1] proponen para activar el pensamiento crítico incentivando la visibilización de medios de origen comunitario y/o sin fines de lucro. Este último sentido puede ser una pieza angular para la construcción de una ciudadanía crítica ya que la hegemonía de los grandes grupos mediales con las ideologías de los grupos de poder que los sustentan, reducen la pluralidad de voces y trasladan sus influencias también al ecosistema de medios digitales [9].

La idea de caso noticioso

Tanto desde la investigación periodística como desde la investigación analítica de la comunicación social, los fenómenos complejos de noticias tienden a denominarse "casos". Según Sunkel [10], el caso periodístico establece relaciones con alguna serie sociocultural que desborda al hecho único. Funciona como disparador de nuevos temas para su incorporación en la agenda pública. Desde la ciencia de datos un concepto relacionable es la detección de "eventos". Así es como existe TwitInfo [6] una herramienta desarrollada por investigadores del MIT con el fin de detectar, visualizar y explorar eventos. Dichos eventos son detectados mediante la actividad en Twitter y son visualizados en una línea de tiempo donde el eje X representa el tiempo y el eje Y el volumen de tweets. También, existe la herramienta Twitris [10] la cual es una aplicación desarrollada para analizar eventos y entregar distintas métricas a través de redes sociales, wikipedia, noticias u otros recursos disponibles en Internet. Otro esfuerzo interesante es el desarrollo chileno aurora twittera [2] que trabajó la diversidad y representación geográfica de contenidos chilenos en twitter. Finalmente, también resulta interesante revisar la noción de eventos espacio-temporal contextualizados propuesta recientemente [8].

Sophia no busca la detección automática de eventos, sino más bien que un usuario interesado defina un conjunto de palabras claves que sean capaces de colectar y organizar una serie de noticias o hechos individuales que constituyen un caso periodístico. La herramienta busca y organiza en la información emitida por un grupo heterogéneo de medios de comunicación chilenos. Dicha heterogeneidad es geográfica pues incluye a medios de todas las regiones del país y política pues se revisan medios de todo el arco político existente en Chile.

5. POBLACIÓN DESTINATARIA Y PRUEBA DE PLATAFORMA

La población destinataria está constituida principalmente por estudiantes de Periodismo o investigadores en comunicación social de Chile.

Desde el mes de junio a la fecha fue utilizada por 66 estudiantes de periodismo de la Universidad Austral de Chile que ejecutaron tareas de a) clasificación de noticias en frentes noticiosos y b) creación de casos noticiosos.

Estudiantes	Hombres	Mujeres	Total
1er año	16	15	31
2do año	15	20	35
Totales	31	35	66

Tabla 1. Distribución de los Estudiantes

Desde la tarea de clasificación de noticias se pudieron confirmar nueve categorías y ajustar la décima de Educación a Educación y Cultura. Desde la actividad de creación de casos periodísticos, los estudiantes agrupados en equipos de 4 personas crearon 10 casos noticiosos que sirvieron para ejecutar tareas de investigación en el tratamiento informativo de noticias sobre: inmigrantes, femicidios, hitos deportivos, entre otros.

La figura 12 es un ejemplo de análisis efectuado por estudiantes de segundo año de periodismo para un caso periodístico de su interés: "caso quemados". Se trata de un crimen perpetrado por agentes de la policía política de la dictadura (2 de julio de 1986) y que en lo últimos meses ha cobrado revuelo al ser puesto en duda, en la actualidad, por candidatos a diputados de la derecha política chilena.

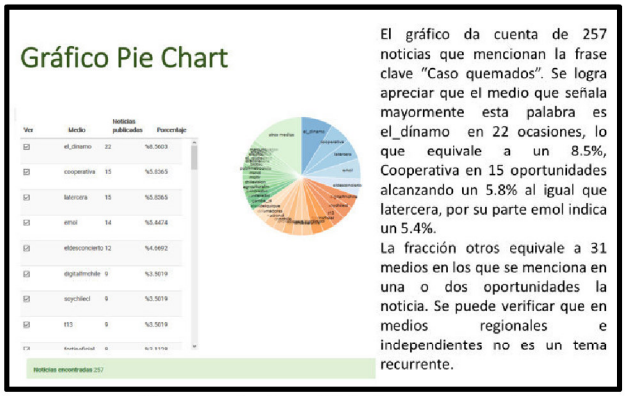

Figura 12. Ejemplo de análisis de un caso noticioso
(Autores: Victoria Carrillo y Diego Chaipul)

También se pudieron recoger sugerencias para mejorar funcionalidades de la plataforma:

> **-Guiar el uso de los histogramas**. Particularmente el selector temporal del segundo histograma no resulta del todo intuitivo.
>
> **-Resaltar el botón buscar**. Cambiar color para resaltar.
>
> **-Destacar el menú de utilidades generales.** Pasa desapercibido para los usuarios y de ello dependerá la comprensión y uso de las funciones de analista.
>
> **-Diferenciar sentido en campos de búsqueda.** Los estudiantes no comprenden la diferencia entre "Concepto o Frase fundamental en la búsqueda" y "Concepto o Frase importante en la búsqueda"

Las pruebas de pilotaje de la plataforma continuarán durante el segundo semestre lectivo (agosto –diciembre de 2017)

6. SUGERENCIAS METODOLÓGICAS DE USO

El proyecto "Redes Sociales y Medios de Comunicación Modelo de análisis basado en minería de datos para la comprensión del ecosistema informativo chileno en internet y la educomunicación ciudadana en la red" ha desarrollado en una de sus líneas la plataforma web Sophia. Se trata de una herramienta que a)colecta las noticias de 290 medios chilenos, b) almacena todos sus titulares, c) permite acceder al contenido de cada noticia in

extenso, d) ofrece la posibilidad de realizar búsquedas sobre un tema e) a partir de las búsquedas se pueden definir "casos periodísticos" de investigación y recolección de datos pasados y activar la colección a futuro f) visualizar los datos colectados en histogramas y pie charts y g) agrupar los datos por cada medio de comunicación o grupo mediáticos.

Los destinatarios finales son estudiantes de periodismo o investigadores en comunicación social. Sin embargo, como explicaremos en las conclusiones y proyecciones también presta utilidad en la generación de Datasets para estudiantes de Ingeniería Informática

7. NIVELES DE USO

SOPHIA tiene tres nivele de uso:

1. **Suscriptor básico**: puede hacer búsquedas de noticias según palabras claves y visualizaciones de las frecuencias de aparición de las palabras claves en histogramas y gráficos de torta.

2. **Analista o estudiante de Comunicación**: puede hacer todo lo anterior pero además puede crear casos noticiosos y programa su seguimiento futuro colectando noticias a partir de un conjunto de palabras claves.

3. **Administrador: incorpora funciones** tales como incorporar nuevos medios y asignar cambio de privilegios a los usuarios (paso de suscriptor a analista).

8. ENTREGA O ACCESO A LA PLATAFORMA SOPHIA

Url Sophia: http://www.sophia-project.info

En dicha URL se puede registrar como usuario con facebook y solicitar condición de analista.
Para efectos de la presente presentación hemos creado un perfil de analista ad-hoc:
username: tise2017
password: tise2017

Url general del proyecto: www.migracionescomunicativas.cl

9. ASPECTOS TÉCNICOS DE LA PLATAFORMA SOPHIA

La arquitectura de la plataforma Sophia corresponde a una arquitectura orientada a microservicios. La plataforma está compuesta por diferentes aplicaciones que trabajan de forma independiente, en donde cada una cumple un rol específico. En la arquitectura, las aplicaciones se comunican a través de una **Rest API**.
En la figura 13, se puede ver los elementos que componen la arquitectura, en donde por un lado la aplicación Sophia Collector se encarga de recopilar los tweets publicados por los diferentes medios de prensa, los cuales son almacenados en una base de datos orientada a documentos (MongoDB).

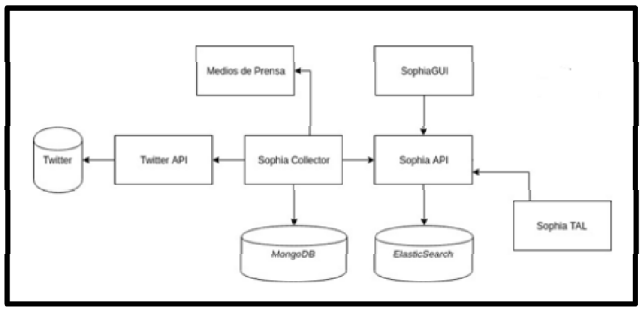

Figura 13 Arquitectura de la plataforma Sophia

El colector descarga el contenido (y otros metadatos) del sitio web del medio de prensa. Almacenando esta información en el motor de búsqueda ElasticSearch a través de SophiaAPI.

Por otra parte, SophiaGUI corresponde a la aplicación web a través de la cual los usuarios pueden acceder a los datos almacenados (realizar búsquedas, importar datos, crear casos noticiosos, etc), esta fue desarrollada principalmente utilizando Django, AngularJS, D3.js, entre otros.

El servicio SophiaTAL (de tratamiento automático del lenguaje), se encarga principalmente de analizar el contenido de los medios de prensa almacenado (noticias principalmente), y utilizando herramientas de inteligencia artificial y NLP, las cuales permiten identificar la categoría temática de la noticia, como también la identificación de las palabras claves de la misma.

10. CONCLUSIONES, UTILIDAD Y PROYECCIONES

La plataforma Sophia se presenta como un producto de software que permite apoyar la formación de estudiantes de periodismo, al proveer una herramienta para realizar búsquedas a partir del contenido publicado por estos medios de prensa en redes sociales. A partir de visualizaciones de datos que permite facilitar la comprensión de ciclos de información, proveyendo la opción de crear casos noticiosos los cuales pueden ser seguidos y analizados a lo largo del tiempo.

Por otra parte, la plataforma es utilizada actualmente en contextos educativos en la Universidad Austral de Chile. Específicamente por alumnos de la carrera de periodismo los cuales realizan búsquedas en la plataforma, exportan y analizan el contenido.

Sin embargo, también alumnos de la carrera de Ingeniería Civil en Informática de la misma casa de estudio, hacen uso de la plataforma en el curso de Tratamiento Automático del Lenguaje, al utilizar la plataforma para la generación de datasets para ser utilizados durante dicha asignatura.

A futuro se espera integrar en la plataforma nuevas visualizaciones de datos, que permitan comparar los medios de prensa, por ejemplo, a partir del volumen de publicaciones por cada categoría temática, palabras claves más utilizadas, etc. Como también poder visualizar la variación de las palabras claves utilizadas por los medios a lo largo del tiempo.

Con respecto a los casos noticiosos, se desea implementar sistemas que permitan notificar al usuario cuando existan cambios significativos en el volumen de publicaciones de dicho caso, con el objetivo de poder informar cuando un caso noticioso se "reactiva".

11. AGRADECIMIENDOS

Este artículo se realizó en el marco de un estudio subvencionado por el Fondecyt n° 1150545 "Redes Sociales y Medios de Comunicación: Modelo de análisis basado en minería de datos para la comprensión del ecosistema informativo chileno en internet y la educomunicación ciudadana en la red". Comisión Nacional de Investigación Científica y Tecnológica (CONICYT). Ministerio de Educación de Chile y contó con el apoyo de la Dirección de Investigación y Desarrollo de la Universidad Austral de Chile.

12. REFERENCIAS

[1] Barranquero-Carretero, A., Lema-Blanco, I. (2016). La juventud española y los medios del Tercer Sector de la Comunicación. Madrid: Centro Reina Sofía sobre Adolescencia y Juventud /FAD. (goo.gl/EGRfJ3).

[2] Graells-Garrido, E., Lalmas, M., & Baeza-Yates, R. (2016, March). Encouraging diversity-and representation-awareness in geographically centralized content. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (pp. 7-18). ACM.

[3] Johnson, L., & Morris, P. (2010). Towards a framework for critical citizenship education. The Curriculum Journal, 21(1). https://doi.org/10.1080/09585170903560444

[4] Johnson, L., & Morris, P. (2012). Critical citizenship education in England and France: A comparative analysis. Comparative Education, 48(3), 283-301. https://doi.org/10.1080/03050068.2011.588885

[5] Marchi, R y otros. (2012). With Facebook, blogs, and fake news, teens reject journalistic "objectivity". *Journal of Communication Inquiry*, *36*(3), 246-262

[6] Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011, May). Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 227-236). ACM.

[7] Paul, R., & Elder, L. (2005). Estándares de competencia para el pensamiento crítico. Estándares, principios, desempeño, indicadores y resultados con una rúbrica maestra en el pensamiento crítico. Dillon Beach: Fundación para el Pensamiento Crítico. (http://goo.gl/UMVRP1).

[8] Peña-Araya, V,,Quezada, M., Poblete, B. & Parra, D.(2017) Gaining historical and international relations insights from social media: spatio-temporal real-world news analysis using Twitter. *EPJ Data Science*. 6:25. https://doi.org/10.1140/epjds/s13688-017-0122-8

[9] Sáez-Trumper, D. (2011). La información en Internet: Breve estado del arte para discutir el poder de los usuarios v/s los medios tradicionales de comunicación en la red. *Rev. austral cienc. soc*, (20), 71-79.

[10] Sheth, A., Jadhav, A., Kapanipathi, P., Lu, C., Purohit, H., Smith, G. A., & Wang, W. (2014). Twitris: A system for collective social intelligence. In *Encyclopedia of Social Network Analysis and Mining* (pp. 2240-2253). Springer New York.

[11] Sunkel, G. (2005). La construcción narrativa del escándalo político en la prensa chilena. *Signo y Pensamiento*, *24*(47).

[12] Van Dijk, T. A. (2009). *Discurso y poder*. Editorial Gedisa.