Minería de opiniones y visualización de datos aplicables a estudios de mercado

Tesina de grado

Alumnos:
Ignacio Saporiti,
Juan Agustín Tibaldo

Directores:
Dra. Claudia Pons,
Dr. Waldo Hasperué

Universidad Nacional de La Plata
Facultad de Informática

Indice

Indice de figuras

Objetivos

El objetivo principal es llevar adelante un trabajo de investigación sobre **análisis de opiniones**. Para esto, vamos a implementar un sistema informático que haga uso de distintas herramientas cognitivas disponibles actualmente en el mercado, y observar los resultados que son capaces de obtener. Documentaremos conceptos asociados con y usados por estas tecnologías. Sin embargo, no es objeto de este trabajo ir al fondo de la teoría en todos los conceptos; creemos que para algunos casos alcanzará con documentar su relevancia con la temática, y describirlos.

El trabajo se estructura de siguiente forma:

En el capítulo primero, se habla sobre redes y medios sociales. Nos preguntaremos qué son, cuáles son, para que se usan, en qué nos importa su análisis y qué información puede extraerse de ellos. Incluiremos una breve historia de las redes sociales y nos describiremos con cuáles son en la actualidad las redes sociales de mayor relevancia y los datos que podrían llegar a interesarnos obtener de ellas.

En el segundo capítulo trataremos el análisis de texto y la minería de opiniones. Hablaremos de técnicas relevantes, sus dificultades, sus algoritmos. También hablaremos de la macro técnica que la engloba: el procesamiento de lenguaje natural. De este haremos un análisis teórico.

En el tercer capítulo hablaremos del software desarrollado, una aplicación Web a la cual llamamos **Opinator**. Describiremos los objetivos y las motivaciones para llevar a delante tal desarrollo. Hablaremos de la arquitectura propuesta, las secciones que la componen, sus funciones y las tecnologías con las que está implementada.

En el cuarto capítulo hablaremos del ensayo realizado para validar el funcionamiento de la aplicación que hicimos, y comparar los resultados de las distintas herramientas externas que utiliza. Se darán detalles sobre el caso sobre el que se basó el experimento, y se harán observaciones sobre los resultados que se obtuvieron.

El quinto capítulo concluye el trabajo. En él, haremos un repaso general, haremos conclusiones y listaremos las posibles tareas, analíticas y técnicas, que se pueden realizar en el futuro.

En el final del trabajo se encuentra la bibliografía utilizada para escribir el contenido que se encuentra en este trabajo.

Capítulo 1. Introducción

1.1 Redes sociales ¿Qué son?

Red social: Una red social es un concepto sociológico, en primer lugar, que describe una estructura social compuesta por un conjunto de actores sociales, sean individuos u organizaciones, y sus interrelaciones. El concepto ha servido para describir con cierta precisión un tipo de software distribuido determinado (sobre el que trata gran parte de este trabajo), software capaz de emular esta capacidad de los actores sociales para relacionarse, pasando de este modo a ser un término que hoy en día más se lo relaciona con el mundo del software, que con el de la ciencia que le dió origen. Este software en cuestión es el denominado servicio de red social, sitio/web de red social, medio social (más habitual en la bibliografía en inglés) o popularmente red social. Podemos definir dos nuevos conceptos muy relacionados a red social.

Servicio de Red Social: un servicio de red social es un sitio web que permite, mediante el uso de los medios sociales y las tecnologías de internet y mobile, realizar todas las actividades que se desarrollan en una red social (sociología) en el ámbito del internet. Los servicios de redes sociales son los que nos importarán estudiar en este trabajo, y los que, gracias a su importancia y popularidad en todos los ámbitos de la vida, han tomado victoriosos el nombre del concepto teórico que los describe: **red social.** A partir de ahora y hasta concluir el trabajo cuando hablemos de redes sociales nos referiremos a servicios de redes sociales aclarando en todo caso de referirnos al término sociológico.

Medios sociales: en muchos artículos se habla indistintamente de medios sociales, redes sociales y servicios de redes sociales. Lo cierto es que estos conceptos tienen sus diferencias, las cuales en nuestro trabajo son relevantes a destacar. En sí un **medio social** es una herramienta que permite comunicarse de manera efectiva entre una o más personas, utilizando las tecnologías móviles y las que ofrece la web para transformar una comunicación en un diálogo. Se puede contrastar este tipo de medio con los medios tradicionales basados en papel o radio y tv en varias puntos como ser calidad, alcance, frecuencia de acceso, usabilidad, inmediatez, permanencia y tipo de interactividad.

A continuación se hace un contraste entre medios tradicionales y los medios sociales.

Puntos de contraste	Medios tradicionales	Medio sociales
Calidad	Buena calidad. En general el contenido es preparado por profesionales y cuidados por editores. Aun así no está libre de subjetividad ya que muchas veces el contenido está sujeto al punto de vista del medio.	Calidad variable : el contenido es generado por usuarios que no necesariamente están preparados en el tema. No es poco común que se *viralicen* contenido engañoso, falso, sesgado, etc. Aun así el contenido ha ido mejorando notablemente a lo largo de los años.
Alcance	Limitado a su medio.En general mayormente es una limitación geográfica. Por ejemplo, un periódico tiene una fuerte limitación geográfica, los periódicos se reparten en determinadas ciudades/pueblos y en ciertos casos se pueden pedir que sea enviado a direcciones particulares fuera de su alcance. Una cadena televisiva sólo tiene señal en tanto el distribuidor de televisión le convenga emitir tal señal. La radio tiene limitación del alcance de su señal.	Limitado a internet. No posee limitación geográfica salvo excepciones (contenido limitado a países como ser el caso de china que limita facebook o google, el caso de HBO que como muchos proveedores de entretenimiento sólo proveen contenidos a determinados países). En general se considera que sólo basta tener acceso a internet para acceder a todo su contenido, lo cual muchas veces hace que se lo considere un medio ilimitado dada las facilidades de acceso a internet. Siete mil millones de personas (el 95% de la

		población mundial) viven en áreas cubiertas por redes de internet móvil (2g)[13].
Frecuencia de acceso	Limitado a frecuencia del medio. Los periódicos, programas de radios y televisivos tienen frecuencia de salida (diaria, semanal, mensual, etc). En general tenemos un usuario que consume el medio en el momento dado por el medio.	Ilimitada, el contenido puede ser publicado, modificado, consumido, compartido en cualquier momento. El usuario puede consumir el contenido en cualquier momento. En general es raro que contenido sea dado de baja.
Permanencia	En general la permanencia es corta y el contenido inmutable. Por ejemplo, un periódico una vez publicado el contenido, este se consume y luego se descarta. El caso de la radio y televisión es considerado como un streaming de datos, los datos pueden ser grabados, pero su existencia es inmediata. Puede ser que el contenido sea repetido a lo largo del tiempo y de este modo se logra la permanencia en los medios tradicionales (una canción o un discurso repetido varias veces al día, propagandas, jingles)	El contenido permanece en el medio. Si bien va quedando desactualizado y tapado por nuevo contenido, no hay necesidad de borrarlo o reemplazarlo. Además todo contenido de un medio social puede ser almacenado. Aun así, es necesario que el contenido sea puesto nuevamente en el tope del stack de contenido si quiere tener permanencia en los usuarios. Muchas veces se utiliza la técnica de BUMP para darle permanencia a contenido en medios sociales muy volátiles.

Inmediatez	Muy relacionado a la frecuencia de acceso. La información sólo puede ser volcada cuando el medio se hace presente. Por ejemplo, los periódicos tienen que esperar hasta el día siguiente para hacer público un hecho que está sucediendo en un momento dado. Algún contenido en televisión y gran parte del contenido en la radio puede ser “en vivo” en cuyo caso se lo considera inmediato. Aun así, este contenido es premeditado, siendo que muchas veces no es capaz de captar contenido que surge de imprevisto.	El contenido puede ser volcado en el medio de forma inmediata. Es común que si se dan eventos o situaciones inesperadas, como ser accidentes, situaciones como la aparición de un famoso en la vía pública, problemas de tránsito, etc. sean los usuarios los que hagan público dicho evento de forma inmediata.
Interactividad	El formato clásico de los medios tradicionales es broadcast uno a muchos sin reciprocidad. Cuando un medio tradicional se expresa, no está esperando que el usuario comente, intervenga, etc. sobre su publicación. Aun así, ciertos medios tradicionales poseen cartas de lectores o posibilidad de comunicación con ellos por vía telefónica, no alcanzando por lejos la interactividad de los medios sociales. No hay diálogo entre consumidor y publicador.	Una de los grandes fuertes de los medios sociales es la posibilidad de establecer un diálogo entre el publicador y el consumidor. Ambos forman parte de un diálogo. Aun así el contenido publicado en el social media sigue siendo uno a muchos, pero los muchos pueden comunicarse con el publicador. Más acertadamente se puede decir que la gran diferencia entre ambos medios es la direccionalidad. Los medios tradicionales son vistos como una comunicación dirigida del

		emisor al receptor, en tanto en los medios sociales la comunicación es vista como unidireccional.
Direccionalidad	Del productor al consumidor.	El consumidor puede comunicarse con el productor, de esta forma se crea contenido que es bidireccional.

Aun así hoy en día la mayoría de los medios tradicionales han encontrado su espacio en los medios sociales. Diarios, canales de televisión, revistas, programas de radio tienen sus cuentas de Twitter o Facebook. Esto no quita que la comparación de arriba se siga cumpliendo, ya que la entidad en el medio social que representa al medio tradicional sigue siendo una entidad de medio social, entonces se le aplican las cualidades de su tipo expresadas arriba. Durante todo nuestro trabajo nos estaremos refiriendo a estos servicios online como redes sociales y en algunas ocasiones como medios sociales de forma indistinta.

Otra definición de red social y su importancia : Las redes sociales, como ser blogs, microblogs, foros de discusión y sitios para compartir multimedia , están siendo utilizados de una forma creciente para comunicar noticias de última hora, participar en eventos, estar conectado en cualquier momento, en cualquier lado. Según el índice Alexa, 5 del top 10 de aplicaciones web con más tráfico son redes sociales. Para darnos una idea de su importancia, el rol socio cultural que tienen ha llevado a la librería del congreso de Estados Unidos a archivar todos los tweets que alguna vez se hayan generado y que estén a disposición pública hasta la actualidad en un archivo propio[14]. Las redes sociales proveen información valiosa sobre la interacción humana y sobre el comportamiento colectivo, atrayendo la atención de disciplinas que incluyen la sociología, negocios, psicología, política, informática, economía y otros ciencias que estudian las sociedades y las culturas.

A continuación presentamos la definición que da wikipedia de red social (traducida del inglés)

“Los medios sociales son medios para la interacción social que usan técnicas de comunicación de alta escalabilidad y accesibilidad. Es el uso de tecnologías web y mobile lo que transforman a la comunicación en un diálogo interactivo. ”

Moturu define medios sociales como “el uso de las herramientas electrónicas y de internet con el propósito de compartir y discutir información y experiencias en otros seres humanos de una forma más eficiente”[1]

Una definición más formal podría ser la propuesta por **Obar y Wildman [15]**

1) Son aplicaciones de internet que actualmente están basadas en la web 2.0

2) El contenido generado por el usuario es la base es las redes sociales.

3) Grupos e individuos crean perfiles de usuarios (o grupos) para una aplicación y este es mantenido por la red social.

4) El servicios de red social facilita el desarrollo de redes sociales online conectando perfiles con otros perfiles o grupos.

Evolución histórica : Las redes sociales surgen en conjunto con internet. En una primera etapa sus funcionalidades estaba centradas en servicios de chat, boletines de publicación electrónica y foros.

Al volverse más complejos estos servicios agregaron funcionalidades de perfil y listas de amigos. Hoy en día todas las redes sociales incluyen de alguna combinación de estos servicios antes mencionados : perfil, lista de amigos, chat, boletín de publicaciones y foro.

1.2 ¿Cuáles son?

Hoy día el número de redes sociales activas no es reducido en su cantidad ni en su diversificación. Esto se debe en gran medida a que una red social puede crearse a partir de cualquier elemento en común entre las personas. Por ejemplo facebook busca conectar virtualmente a amigos y conocidos en la realidad, pero no se limita sólo a eso, ya que también permite contactarnos con figuras importantes, como políticos (ver figura 1), artistas (ver figura 2

y 3), empresarios o entes abstractos como ser productos, iniciativas, empresas y medios de comunicación entre otros.

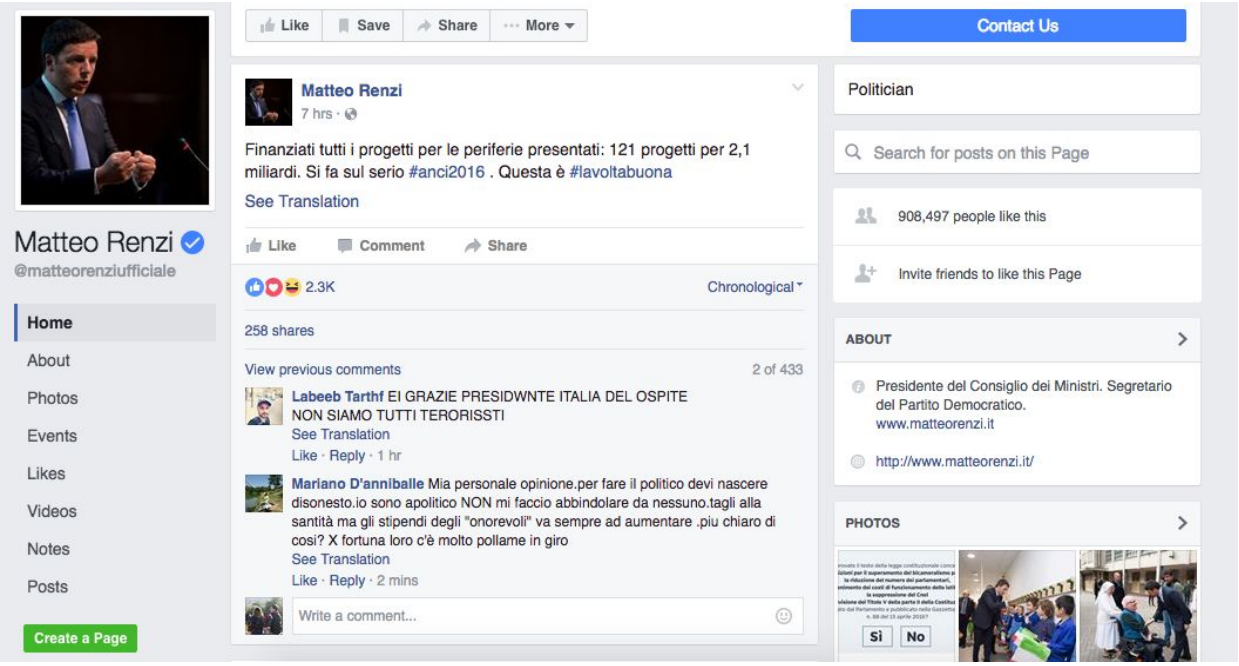

Figura 1. Cuenta oficial en Facebook del ex Primer Ministro de Italia, Matteo Renzi.

Figura 2. Cuenta oficial en Twitter de The Rolling Stones.

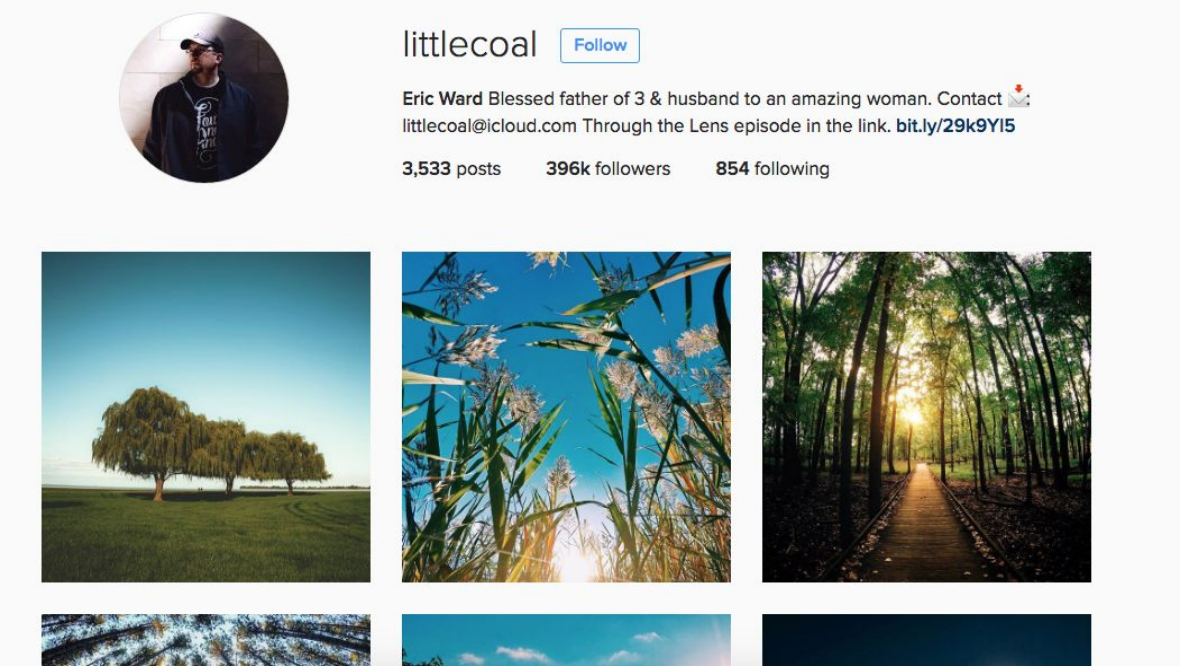

Figura 3. Las redes sociales permiten a los artistas darse a conocer en todo el mundo. Cuenta de Instagram de Eric Ward, fotógrafo de paisajes.

Algo parecido pasa con Twitter, salvo que en este caso está orientado a seguir a gente que nos influencie y nos inspire, su forma simple de mensajes hace que el contenido sea más directo, sencillo y al grano. Instagram por otro lado lo que busca es compartir fotos de momentos, situaciones o lugares de una forma amigable, posee filtros de nivel profesional, predefinidos que permiten mejorar fotos en segundos (ver figura 4).

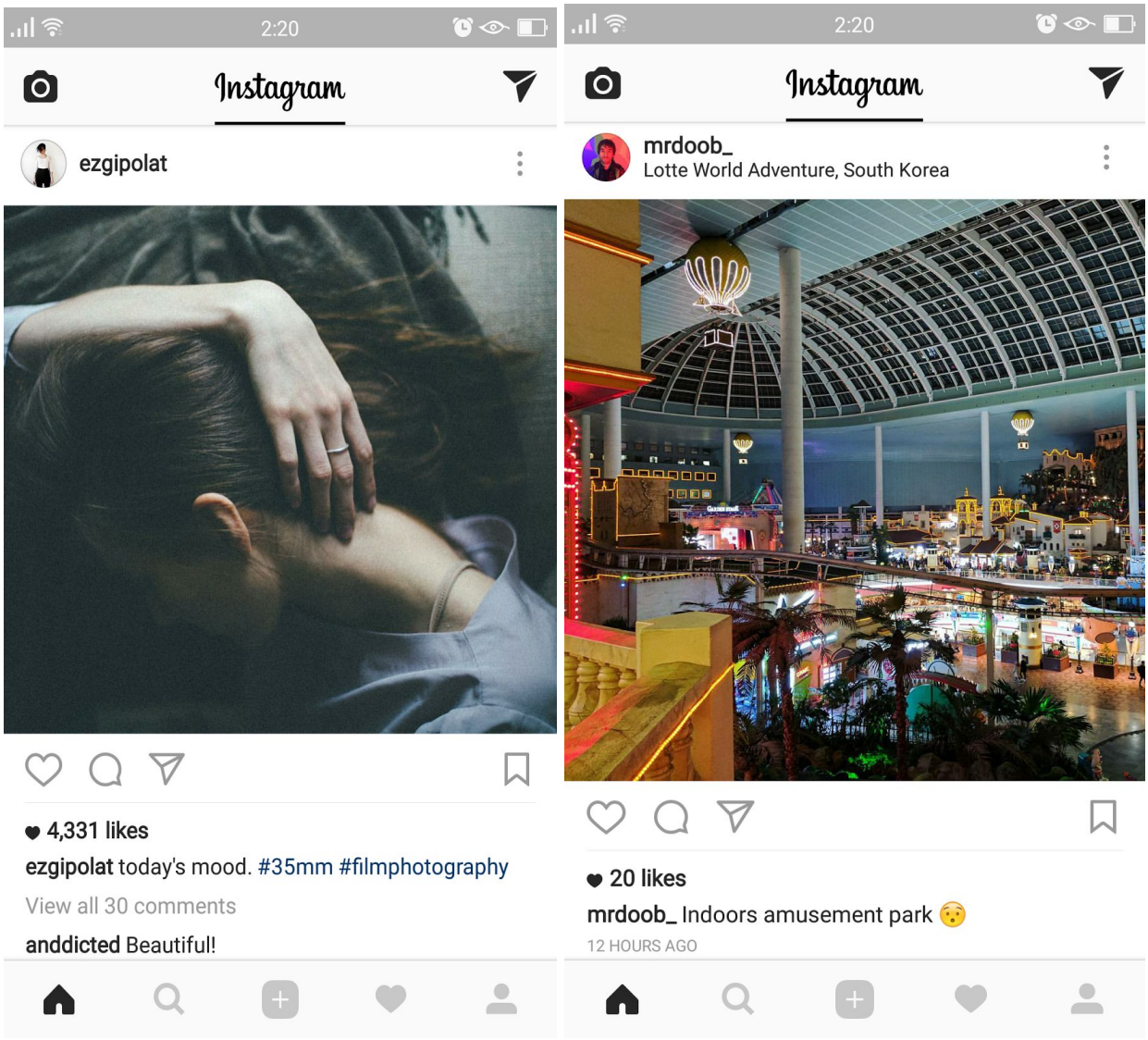

Figura 4. Ejemplos de filtros de mejoras de fotografías en Instagram.

Por otro lado tenemos redes sociales como ser upwork o linkedin, donde el foco es más profesional, entonces en estos sitios el vocabulario es distinto así también como las conecciones que hacemos. En estos últimos nos enfocamos en buscar conecciones que necesitemos en el ámbito laboral, siendo que ayudan a encontrar gente capaz de suplir un puesto que actualmente necesita una compañía para resolver sus objectivos.

A continuación agregamos una lista con las redes sociales más relevantes a nivel mundial facilitada por Alexa Internet

Name	Description/focus	Date launched	Registered users	Registration	Global Alexa[1] page ranking
Google+	General	28 Junio 2011	1,600,000,000	Abierto a mayores de 13 años.	NA
Facebook	General: photos, videos, blogs, apps.	Febrero 2004	1,280,000,000	Abierto a mayores de 13 años.	2
Twitter	General. Micro-blogging, RSS, updates	15 Julio 2006	645,750,000	Abierto a usuarios de cualquier edad.	8
Qzone	General. En chico simplificado;da servicios a usuarios de china continental		480,000,000	Abierto a usuarios de cualquier edad.	NA
Sina Weibo	Sitio de microblogging social en China continental.	14 Agosto 2009	300,000,000	Abierto a usuarios de cualquier edad.	28
Instagram	Sitios para compartir videos y fotos.	Octubre 2010	300,000,000	Abierto a mayores de 13 años.	41
Habbo	General para teens. Mundialmente alcanza 31 comunidades. Chatear y perfiles de usuario.	Agosto 2000	268,000,000	Abierto a mayores de 13 años.	15,255

VK	General, incluyendo subida, reproducción y búsqueda de música y videos. De gran popularidad en rusia y países de la ex unión soviética.	Septiembre 2006	249,409,900	Abierto a usuarios de cualquier edad.	38
Tumblr	Plataforma de Microblogging y redes sociales. Compartir imágenes.	Febrero 2007	226,950,000	Abierto a usuarios de cualquier edad.	41
LinkedIn	Redes profesionales y de negocio.	Mayo 2003	200,000,000	Abierto a mayores de 18 años.	12
Renren	Sitio muy popular en china. Fue conocido como 校内 (Xiaonei) hasta Agosto del 2009.	Diciembre 2005	160,000,000	Abierto a usuarios de cualquier edad.	95

1.3 ¿Para qué se usan?

Su uso no está limitado a una función específica y cada red social puede tener varias funciones relacionadas al ámbito general que busca abarcar. En general podríamos decir, sin embargo, que todas las redes sociales buscar poner en contacto personas o instituciones entre ellas. Las mismas son usadas por los usuarios para relacionarse entre ellos, expresar ideas, compartir información o contenido multimedia. Si bien esto es verdad, también hay un creciente uso por parte de empresas u organizaciones para promocionarse, siendo que las redes sociales muchas veces otorgan herramientas pagas que facilitan el posicionamiento del contenido de estas empresas u organizaciones en la red social[16]. Es notable el uso de las redes sociales también como herramienta para información y organización política siendo el caso de su uso en las primaveras árabes lo que marcó la tendencia de tomar a las redes sociales como un instrumento importante en la política [17] . Otros caso notables es el uso de

las redes sociales en la campaña de Obama en estado unidos [18] . En nuestro país, durante las elecciones del 2016 también se hizo gran énfasis en los medios de comunicación para promover la candidatura del PRO a la presidencia [19]. Durante la campaña el candidato Mauricio Macri estuvo presente en redes sociales como Instagram (red en la que ocupa puesto número 9 de líderes más seguidos a nivel mundial)[21], Snapchat, una red social para teens, facebook, twitter entre otras.

1.4 ¿Qué información se puede extraer de ellas?

Las redes sociales pueden ser accedidas de muchas formas, por ejemplo, teniendo acceso a APIs REST que proporcionan hacer consultas como es el caso de Facebook, Twitter, Linkedin. Accediendo directamente al contenido publicado manualmente o con alguna herramienta de data scraping sobre el contenido html en sí. Usando herramientas o plugins especializados que consumen las APIs facilitando aún más nuestra tarea.

A continuación presentamos información que un conjunto de las redes más importantes están guardando [22]:

- Facebook junta 63 piezas de datos distintos para su API, más que ninguna otra red social. Facebook actualmente dispone de mucho contenido que es compartido, este contenido puede ser accedido mediante esta API. Para darnos una idea del volumen de datos que maneja, el botón “like” de facebook es apretado 2.7 billones de veces por día.
- Google+ ayuda a google a contextualizar la cantidad de datos de búsqueda que sus bases de datos atesora, ofreciendo pistas en lo que la gente podría llegar a buscar teniendo en cuenta una información inicial. Actualmente la información trabaja en el otro sentido, permitiendo que lo que tiene mucho alcance en Google+ también sea relevante en Google.
- Un 22 por ciento de usuarios de Linkedin tienen entre 500 y 1000 contactos de primer grado (conectados directamente en el grafo de amigos), y un 19 por ciento

tienen entre 300 y 500. Esta información es clave para entender la nueva forma de reclutar personal y mantenerlos en la empresa.

- En uno de sus picos más altos, Twitter estaba procesando 143,199 tweets por segundo globalmente. Estos tweets proveen una ventana en tiempo real hacia noticias e información que le importa a las personas. 52 por ciento de los usuarios de Twitter en Estados Unidos consumen noticias en esta aplicación (el porcentaje es aún mayor que los que lo hacen en facebook) de acuerdo con datos de Pew Research[20].
- Cientos de miles de imágenes de productos son “pinned” a páginas personas en Pinterest todos los días. Estas representan una ventana única a millones de futuros compradores. Por ejemplo, más del 17 % de todos los pinboards son categorizados bajo la keyword “Home”, mientras que un 12 % cae bajo “Style/fashion”. Sobre esto, un 80 % de los pins en Pinterest son repins, por lo que imágenes de productos tiene

Las redes sociales tuvieron un rápido crecimiento en los últimos años. Estas redes cuentan con usuarios, los cuales pueden interactuar e interconectarse entre sí en forma de contactos, o miembros de grupos, comunidades, entre otros. Dentro de estos medios, los usuarios además pueden generar contenido, expresarse, y emitir expresiones sobre temas variados. Antes de su existencia, las personas se enteraban de lo que ocurría en el mundo a través de una forma unidireccional de comunicación, que era aquella que brindaban los medios informativos tradicionales, como la radio, la televisión, y el diario. Para el consumidor, había pocas o nulas opciones para comunicarse.

Estos medios brindan la oportunidad de que cualquier persona pueda comunicar, por lo que la podemos considerar un canal bidireccional de la información. Junto con estas oportunidades, surgen desafíos desde el punto de vista análitico, estadístico, y hasta comercial. Dada la enorme cantidad de sus usuarios, el contenido generado es muy valioso para saber interpretar el humor de una sociedad, las opiniones sobre un producto, sobre un candidato a elecciones, etcétera.

Se pueden categorizar a los medios sociales a partir del contenido que se comparte a través de ellos. Sitios como Facebook e Instagram, se centran el intercambio de fotos y otros; Youtube

es, la plataforma más conocida para compartir videos; y Twitter es el servicio más conocido de micro blogging, presente en distintos formatos como Web y aplicación móvil.

Si bien cada servicio tiene sus ventajas, todos se asemejan en cuanto a que el usuario puede expresarse escribiendo en comentarios. Es por eso que la mayoría de datos presentes en lo que se comparte en las redes sociales viene en forma de texto, de manera no estructurada. Otros inconvenientes respecto el contenido en las redes es que el vocabulario presente puede no ser estándar, dificultando su interpretación, y además hay veces en que las opiniones son dadas sobre contenido en otras redes, o recursos accesibles en la Web, por lo que es importante hacer un análisis completo de contenido y de enlaces. En casi todas las plataformas se ha aprovechado el concepto de etiqueta, para facilitar el agrupamiento de contenidos dentro de contextos, comunidades, o situaciones en común.

Los datos textuales que se comparten en las redes sociales nos permiten analizar una cantidad y variedad de información que antes no se hubiese podido realizar, siquiera con entrevistas y otros métodos manuales. Sin embargo, estos datos tienen algunas características que dificultan su estudio, como por ejemplo estar escritos en un dialecto por fuera de lo tradicional.

Las facetas más distintivas del texto en medios sociales son:

- **Sensibilidad temporal:** los usuarios publican contenido regularmente en la semana, y en los casos de redes de microblogging, hasta varias veces en un mismo día, por lo que los datos generados son un claro reflejo del momento en que fueron emitidos.
  Algunos de los comentarios pueden contener opinión sobre eventos de espectáculo y deportivos, productos lanzados recientemente, o campañas políticas, por lo que, con el análisis adecuado, sería posible interpretar la opinión o el humor de sectores de la sociedad, realizando una adecuada caracterización de la misma.

- **Homogeneización de opiniones:** los escritos y las opiniones generadas por los usuarios no son tan independientes ni idénticamente distribuidos como lo eran antes de las redes sociales. Esto quiere decir que los usuarios se influencian unos a otros, condicionando lo que se va a comunicar. Por ejemplo, si un usuario con muchos seguidores o amigos en una red, hace un comentario negativo respecto a un producto,

es posible que algunos de esos contactos adopten esa información para moldear su propia percepción de ese producto.

- **Corta longitud:** esto sucede especialmente en las redes de microblogging, que ponen límites de caracteres en los escritos que un usuario puede realizar por vez. El caso paradigmático es el de Twitter, aunque hay otros casos como Picasa. Por esto, para estudiar los comentarios en este tipo de redes, es necesario contar con métodos de análisis de textos que funcionen adecuadamente para variantes de corta longitud.
  La corta longitud permite que los usuarios participen en las redes de manera ágil, pero desde el punto de vista del procesamiento es contraproducente, ya que se cuentan con unas pocas palabras de donde se debe extraer la información y el contexto.
  Para solventar dicha escasez, se usan métodos tradicionales

- **Frases no estructuradas:** el contenido se caracteriza por su variedad en cuanto a la calidad. Esto depende de varios factores. En parte, del tipo de red social, ya que las respuestas de un sitio de microblogging no van a ser las mismas que las comentadas en un foro, por ejemplo. Por otra parte, depende también del grado de conocimiento del usuario; habrá usuarios con experiencia sobre lo que comentan, como también habrá usuarios que no se destacan por la calidad de sus respuestas. Además, en algunos casos los comentarios pueden ser abusivos, dada la posibilidad a los usuarios de comentar desde el anonimato, en algunas redes.
  Es por esto que a la tarea de análisis de texto se le suma la complejidad de separar los comentarios de alta calidad de los de baja calidad.
  Además, se suma que los usuarios puedan componer mensajes utilizando nuevos términos, acrónimos y abreviaciones, que no terminan siendo palabras tradicionales. Estas, si bien son maneras prácticas e intuitivas de comunicación entre los usuarios, son términos para los cuales no es sencillo realizar el análisis semántico a través de minería de texto.
  Por último, dada la arbitrariedad de los comentarios, es de suma importancia tener en cuenta el contexto en el que se realizan, ya que pueden haber datos que solamente generan “ruido” en el análisis.

- **Tener abundante información:** en los medios de redes sociales, no solo se comparte texto, sino también imágenes, videos, emojis y otros, como por ejemplo los tags de Twitter, que comienzan escritos con el símbolo #, seguido de una cadena de texto arbitraria. Pareciera que esta información no es relevante para realizar un análisis profundo, pero, si se tienen en cuenta los metadatos, etiquetas, y enlaces asociados a este contenido, se puede enriquecer el la información obtenida de un comentario con estos contenidos. Por esto, es que los algoritmos que se encuentran en desarrollo en la actualidad, tienen en cuenta esto, llegando a resultados interesantes. Por ejemplo, hay investigaciones hechas en sitios de microblogging en donde se toman los enlaces de los comentarios sobre un tema en común, para detectar si el tema es un rumor o un hecho creíble. Otro caso para mencionar, es uno en que se toma información sobre los enlaces entre usuarios en una red, para identificar comunidades en común.

Las opiniones son centrales a casi todas las actividades dentro de la sociedad, dado a la influencia que estas tienen en nuestro comportamiento. Cuando necesitamos tomar una decisión, solemos recurrir a la opinión de otros. Esto ocurre en el ámbito de las empresas y las organizaciones, que para mantenerse relevantes, deben estar en contacto frecuente con sus clientes, y así conocer sus opiniones sobre sus productos o servicios. A su vez, los consumidores consultan las opiniones de otras personas sobre productos o servicios que quieren adquirir.
Como se dijo en el capítulo anterior, con el crecimiento de la Web 2.0, arribaron distintos tipos de sitios y espacios en donde los usuarios comenzaron a compartir su opiniones, como en el caso de los foros de discusión, blogs, redes sociales, entre otros. De esta manera, una persona ya no depende de que un conocido cercano conozca sobre algo que quiere adquirir; puedes consultarlo en la Web. Para una organización o empresa, a su vez, ya no son necesarias las encuestas, ya que la información de opinión está presente y de manera pública y abundante en los distintos medios mencionados.

Esta abundancia de información tiene su faceta desafiante. Un lector promedio va a tener dificultad en identificar las opiniones en una cantidad relevante de comentarios, y en extraer las opiones de ellas. También se verá con dificultades para agrupar sentimientos, para generar un informe compacto y resumido sobre los sentimientos de los consumidores. Es por esto que,

para sacar realmente provecho de esta fuente de información, es necesario automatizar el proceso de análisis de sentimientos.

Capítulo 2. Análisis del texto y minería de opiniones

2.1 Introducción

Para poder acceder a la información y el conocimiento disponible entre tanto contenido textual, son necesarios algoritmos avanzados que sean capaces de descubrir patrones de manera dinámica y escalable. De esto último se encarga la *Minería o Análisis de Texto*, que a lo largo de los años ha tomado prestado conceptos de *Machine Learning, Minería de Datos, Natural Language Processing e Information Retrieval*.

De Wikipedia tenemos la siguiente definición:

"El Análisis o Minería de Texto comprende un conjunto de técnicas lingüísticas, estadísticas y de Machine Learning que modelan y estructuran la información contenida en fuentes textuales, para realizar análisis de datos e investigación"

La minería va más allá del intento de obtener información, sino que también busca asimilar para permitir la toma de decisiones.

Uno de los grandes desafíos dentro de la minería de texto, es que los datos deben tratarse de manera semántica, más que como una colección de palabras. Esto quiere decir que, de la detección de entidades tales como, por ejemplo, personas y organizaciones, más las relaciones entre ellas, se puede obtener información del análisis de las relaciones entre ellos, presentes en el texto. Para esto, se requieren algoritmos de procesamiento de lenguaje natural. Los métodos actuales todavía no están lo suficientemente preparados para funcionar adecuadamente en textos escritos de manera libre, como se espera encontrar en comentarios y entradas generados en las redes sociales.

2.1.1 Métodos y técnicas

Se pasan a detallar formas de aplicar Análisis de texto para poder extraer información de los contenidos de los usuarios en las redes sociales.

La técnica de *Ocurrencia de eventos*, se realiza analizando comentarios, recopilando texto, enlaces y otras formas de información, que traten sobre una temática de la actualidad. Pudiendo hacer ese estudio en tiempo real, se podría seguir la evolución de la noticia, por lo que se puede considerar que la actividad dentro de una red social puede funcionar como sensor de lo que pasa en la realidad. Esto hace que aplicar los métodos correctos para su estudio sea tan importante. Algunos métodos se centran en la manera en que “explota” el interés sobre una noticia, y cómo se difunde entre los usuarios. Otros, se encargan de analizar los tags en fotos o videos, más los metadatos que pueda haber sobre tiempo y geolocalización, con el fin de reconocer eventos asociados. De esta manera, se podría agrupar contenido referente a noticias, eventos y sucesos.

Otra alternativa de aplicación, es la de analizar sitios de consulta, en donde usuarios pueden exponer sus dudas, para que otros participantes expertos en el tema puedan responder y ayudar de manera colaborativa. Estos servicios son de gran utilidad para reunir personas que buscan resolver problemáticas, interactuar con “expertos” sobre un tema en particular, o satisfacer la curiosidad. Ejemplos de estos son *StackOverflow*, en donde hacen preguntas y respuestas respecto a desarrollo de software, y *Yahoo! Respuestas*, que es un sitio donde se pueden realizar preguntas sobre cualquier tema.

Dada la gran base de datos de pregunta-respuesta que se ha construido a lo largo del tiempo, es muy probable que, cuando un nuevo usuario de estos sitios necesite hacer una consulta, no haga falta que realice la pregunta, porque esta ya haya sido realizada y resuelta por otros antes. Es importante, entonces, que ese nuevo usuario pueda encontrar esas consultas resueltas, por lo que estos servicios, y los buscadores Web, deben ser capaces de interpretar una consulta, para responder con las mejores respuestas, y de esta manera satisfacer las necesidades de la consulta.

Una forma de implementar un método de búsqueda de las mejores respuestas es analizando su calidad. Para esto, Harper [8] analiza las preguntas para clasificarlas entre informacionales y conversacionales. La primera abarca las consultas que buscan obtener información y satisfacer una duda, por ejemplo. *¿Java soporta herencia múltiple?*. La segunda categoría incluye aquellas consultas que buscan iniciar un debate o charla alrededor de un tópico, e.g. *¿Qué lenguaje me recomiendan para aprender programación orientada a objetos, y por qué?*. Los autores creen en que las preguntas conversacionales no tienen tanta calidad para responder a dudas como sus pares informacionales, por lo que implementaron algoritmos de machine learning capaces de clasificar las preguntas, y obtener las aquellas que mejor respondan dado el contexto.

Agichtein y sus colegas [7], en cambio, interpretaron la calidad de los pares pregunta-respuesta, según datos estadísticos sobre su uso, y datos respecto a la interacción entre usuarios que preguntan y los que responden. También contaron con un algoritmo encargado de clasificar y seleccionar las mejores respuestas en función de este concepto de calidad.

Un método adicional es el de etiquetado social, que consiste en que los usuarios de Internet agrega etiquetas a los contenidos que ellos mismos generan. Las etiquetas sirven entonces para organizar, agrupar y buscar contenidos y recursos relacionados a esa etiqueta. Es una tarea que los usuarios adoptan proactivamente para hacer que sus contenidos sean más fáciles de localizar por los demás.

La cantidad de datos etiquetados por los propios usuarios es tal que se la está considerando cada vez más para usarlos en combinación con análisis de texto.

Se reconoce que, en las redes sociales, los motores de búsqueda tienen algunas falencias para que un usuario pueda encontrar los contenidos que le resulten de interés a partir de etiquetas. Esto es en parte, porque una etiqueta es texto no estructurado, y cada usuario puede crear una etiqueta cualquiera, sin seguir convención alguna. Entonces, para un mismo contenido, puede haber una gran variedad de etiquetas. Por otro lado, los buscadores no obtienen los mejores resultados, porque que se basan en buscar palabras clave, y no hacen un análisis de la relación semántica entre etiquetas.

Actualmente, el estudio de etiquetado social se divide en dos categorías. La primera busca mejorar la calidad de las recomendaciones de etiquetas para una búsqueda, mientras que la

segunda busca usar a las etiquetas como soporte para distintos usos. Sigurbjornsson y Van [10] realizaron un estudio sobre formas de asistir a los usuarios durante la creación de etiquetas para los contenidos compartidos en Flickr. Para esto, usan estrategias para detectar las mejores etiquetas relacionadas al contenido, para recomendarlas al usuario.

2.1.2 Dificultades actuales

Los escritos presentes en las redes sociales son cortos y no estructurados, y la mayoría de los métodos que se usan para extraer información tratan a esos escritos como bolsas de palabras, obviando hacer un estudio del contexto en que se escribió, y la forma en que se relaciona con otros escritos.
Esto se debe al tratamiento que se le da a los textos en la mayoría de los métodos actuales, en donde se los trata como una “bolsa” de palabras sueltas, y en donde resulta difícil relacionar términos o analizar el contexto en que esas palabras fueron escritas.

2.1.3 Algunas soluciones

Para lograr un mejor análisis, es importante agregar información, es decir, metadatos que nos permitan ampliar lo que sabemos sobre estos datos textuales. Un modelo exitoso es el de agregar de manera comunitaria, esto es, que cada usuario pueda brindar datos extra, en vez de que uno solo o unos pocos realicen esa tarea. Por ejemplo, para el caso de las redes sociales, dado un texto, el usuario puede relacionarlo con otros contenidos a través de etiquetas, como también puede vincularlo a otros usuarios a través de lo que coloquialmente se conoce como “menciones”. Estos son datos que surgen voluntariamente del usuario, aunque puede haber agregación involuntaria. Desde la red, se pueden analizar conceptos como la hora y localización de la publicación, cuántos y qué usuarios la comentan y/o marcan como favorita (o como “me gusta” en Facebook, Instagram).

Otro ejemplo agregación comunitaria, es el de Wikipedia, que si bien no es una red social, abarca a una colección de usuarios que editan, amplían y generan entradas de texto al estilo

de una enciclopedia, y relacionan a estas entradas con enlaces, citas, contenidos audiovisuales, etc. En Wikipedia, no hay expertos ni dueños del contenido . Esto quiere decir que, a priori, cualquiera de la comunidad puede editar, y entre usuarios se regulan para buscar que los textos sean fieles a los hechos y estén completos y bien redactados.
Cada artículo de Wikipedia trata sobre un tema específico. Si el nombre del tema se repite en distintos conceptos, se desambigua en distintas entradas. Cada artículo pertenece a una categoría, y cada uno puede estar relacionado con otros, lo que los enriquece, ya que contando con esta agregación se puede realizar un estudio más completo.
Existe una iniciativa de Somnath et al [9], en donde se busca completar o ampliar a textos cortos, relacionándolos con entradas de wikipedia, para la cual se obtuvieron buenos resultados.

Actualmente se están utilizando métodos combinados de análisis de texto y de enlaces. Tradicionalmente, el análisis de texto compara la similitud entre dos escritos basándose en la similitud de atributos. En cambio, cuando el contenido contiene enlaces que lo relaciona con otros contenidos, la comparación entre documentos pasa a basarse en la conectividad entre los mismos. Por ejemplo, puede compararse cuántos caminos posibles hay entre los autores de ambos documentos, para detectar qué los une o qué tienen en común. Esto serviría, no sólo para caracterizar a los autores, y así sacar un perfil de ellos, sino que también ayudaría a entender el contexto de cada documento, que, como se dijo anteriormente, tiene un papel fundamental en esta rama.
Dado lo mencionado, los enlaces se convierten en un aliado que nos permite ampliar la capacidad de entender semánticamente un escrito, cosa que no era tan sencilla tomando sólo el contenido de los mismos.
A la hora de analizar textos con enlaces, se deben tener en cuenta algunas características:

- Las conexiones entre usuarios en una red son multidimensionales, es decir, que como en la realidad, pueden haber distintos tipos de relaciones personales, si bien a nivel conexión del sistema es indistinto. Por ejemplo, la conexión personal entre una persona con su pareja, no va a ser la misma conexión que con un profesor de la facultad, dentro de la misma red social. Sin embargo, los enlaces posiblemente sean los mismos. Existen algunos medios que dan la posibilidad de diferenciar las relaciones personales, como la pareja, la familia, etc.

- Existen dificultades para implementar la estructura que mejor modele a las conexiones. Por ejemplo, si se usa una matriz de adyacencias, el contenido será escaso, y la misma será altamente dimensional. Además, los procesos para obtener contenidos recorriendo la estructura a priori se asumen costosos, dada las características mencionadas.
- Las redes tienen carácter dinámico, a diferencia de los escritos tradicionales cuyo contenido permanecía igual, hasta que se hace alguna reedición con agregados. El dinamismo de las redes implica que el accionar de los usuarios cambie: de un momento a otro pueden ingresar nuevos participantes, mientras que algunos pueden dejar de usar el medio. Dado este panorama, es muy importante hacer un análisis periódico, para no quedarse con una "fotografía" desactualizada de los contenidos, como también se debe prestar atención al modo en que se integra la información actualizada a la ya existente, para mantener el sentido ni perder conocimiento.

2.2 Procesamiento del Lenguaje Natural

Las personas se comunican de diferentes maneras: hablando y escuchando, con gestos, o a través de texto, es decir, palabras escritas o impresas en una superficie plana o pantalla de un dispositivo electrónico, para que sean leídas por un destinatario.

El Procesamiento del Lenguaje Natural (NLP, por sus siglas en Inglés) es un conjunto de procesos informáticos que se encargan de analizar e interpretar textos. Tiene su origen en disciplinas variadas, principalmente del área lingüística, las ciencias de la computación, más pequeños aportes de la psicología y la lógica.

Tradicionalmente, el NLP ha sido visto como un proceso compuesto por un número de pasos, contrastando las diferentes teorías entre la sintaxis, semántica y pragmática. Analizar al texto primero en términos de sintaxis provee de un orden y estructura que hacen posible el análisis en términos de semántica; y luego el paso de análisis pragmático se intenta explicar la elección de determinadas formas de realizar el enunciado en función de los factores contextuales. Sin embargo, esta visión tripartita compuesta por sintaxis, semántica y pragmática, solo sirve como un punto de partida si se quiere analizar texto en lenguaje natural real, como es el caso del que se podría extraer en redes sociales, en donde lo contextual juega un papel crucial, y en donde se pueden presentar desafíos difíciles de sortear, como por ejemplo, sarcasmo y *slang* (argot).

Por eso, es necesaria una descomposición más fina en cuanto a los pasos que componen a NLP.

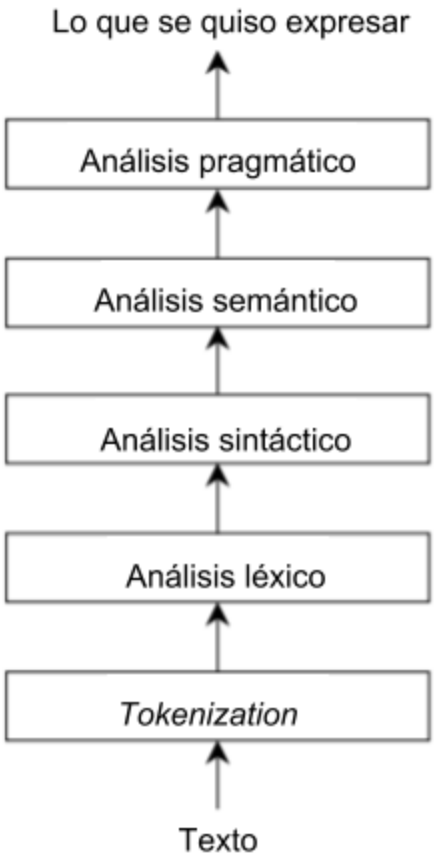

Figura 5. Pasos sugeridos para obtener un proceso de NLP satisfactorio. [12]

A los tres pasos básicos, Chapman y Hall [13] sugieren agregar dos más al principio: **tokenization** o tokenizing, y **análisis léxico** (ver Figura 5).

2.2.1 Tokenizing (Pre Procesamiento de texto)

Tokenizing es una de las operaciones más básicas que pueden aplicarse a un texto, y consiste en dividir una secuencia de caracteres en palabras, signos de puntuación, números y otros elementos. Con esta técnica, no se puede conseguir mucha información sobre la estructura

sintáctica, y mucho menos puede conocerse sobre la semántica de esa secuencia. Sin embargo, se puede extraer información tal como por ejemplo, todos los nombres propios presentes, ya que bastaría con buscar todas las palabras que comienzan con mayúscula. De todas formas, no se podría, por ejemplo identificar si los nombres corresponden a personas, organizaciones o países; peor aún, no puede distinguirse entre una palabra que es nombre propio y una al comienzo de una oración, que también empieza con una letra mayúscula.

A partir del tokenizing puede realizarse una **búsqueda de patrones**, que es una forma alternativa de detectar nombres u otro tipo de palabras. Para el caso de los nombres, el algoritmo debe buscar las partes del texto que suelen ir acompañadas de los mismos, como por ejemplo:

- Títulos, precedentes a una palabra que comiencen en mayúscula, como *Dr., Sr., Sra, Lic., Ing.*, etc.
- Palabras que comiencen en mayúscula, seguidas de un verbo que aplique a cosas que pueda hacer una persona: *María dijo*, *Juan corre*, etc.

Partes del discurso

Un proceso que va un paso más lejos, es el de asociar cada token con su categoría gramatical, mejor conocida como su **parte del discurso**, las cuales pueden ser: **sustantivo**, **nombre propio**, **verbo**, **adjetivo**, **adverbio**, **pronombre**, **preposición**, **conjunción** y **determinante**. Se tiene un algoritmo clasificador, que detecta a qué categoría pertenece cada palabra de entrada, y le agrega la etiqueta correspondiente. Estos clasificadores son preparados con un texto de entrenamiento, de manera que cuenten con un conjunto de reglas que les permitan identificar las palabras. Para palabras desconocidas, suelen omitirse y guardadas para que sean identificadas manualmente. En esta clasificación pueden encontrarse dificultades, como por ejemplo palabras que pueden pertenecer a categorías distintas, y solo se pueden desambiguar haciendo un análisis más profundo, teniendo en cuenta el contexto en el que aparecen los términos o *token*.

Estructura de las palabras

Las frases son estructuras compuestas por palabras, pero estas últimas también tienen estructura propia. En muchos idiomas, esta estructura puede cambiar, como en el Inglés y el Español, en donde los verbos varían dependiendo del tiempo en que son usados. Por ejemplo, *shake*, el término inglés en presente para "temblar", en tiempo pasado se escribe *shook.* Un ejemplo del Español, es la palabra *traducir*, que cambia a *tradujo* si se quiere expresar en tiempo pretérito.

2.2.2 Análisis léxico

La unidad básica en extracto de lenguaje natural es la palabra. El análisis léxico se encarga de estudiar la palabra, concentrándose en su estructura, y teniendo enfoques distintos para obtener mejores resultados. Un enfoque consiste en concebir a la palabra simplemente como una cadena de símbolos, como por ejemplo *"escribir"*, mientras que en otra estrategia se piensa en la palabra como una etiqueta (formalmente llamada *lema*) que está asociada a un conjunto de cadenas, como por ejemplo *"escribir"* puede ser el lema del conjunto {*"escribir", "escribiendo", "escrito" ...*}.
Esta última concepción es conocida como Lematización y es muy útil, ya que permite conocer las distintas variantes morfológicas de una palabra. Se conoce como morfología de una palabra al grado de distintas formas que puede adoptar, por ejemplo, en distintos tiempos discursivos, o con prefijos y sufijos.

Se puede pensar en la relación de una palabra morfológicamente compleja con su lema como una parte del análisis léxico, conocida como *parsing*, mientras que la otra parte del proceso es el paso inverso, es decir, aquel en donde se relaciona un lema con una palabra cualquiera del conjunto de variantes morfológicas del lema. Este último proceso se llama *generación morfológica*, o simplemente, *generación*.

La Lematización se usa, por ejemplo, en técnicas de *Information Retrieval*, para crear una lista de términos clave a partir de un texto cualquiera: primero se procesa el texto en búsqueda de palabras morfológicamente complejas; luego estas se descomponen en dos: por un lado, las palabras que son equivalentes a su forma canónica (iguales a su lema), conocidas como *stems*, y por el otro, añadiduras, que indican tiempos verbales, persona, o cantidad. Se

descartan los añadidos, y de esa manera se obtiene una lista de palabras clave en su forma canónica, a pesar de que las palabras originales del texto tuvieran mayor complejidad morfológica.

Sin embargo, en lenguajes con gran riqueza morfológica, en vez de una lista o conjunto de transformaciones concretas, es más económico usar reglas que sirvan para identificar las transformaciones que esa palabra canónica pueda sufrir. Otro caso de ventaja del uso de reglas por sobre un conjunto de palabras, es que con esta última si en el análisis de texto se encuentra una transformación de un lema que no se encuentra en ese conjunto, el proceso no podrá identificarla, mientras que la estrategia de reglas sí tiene el potencial de hacerlo.

Todo proceso de análisis léxico se encuentra con algunas dificultades. En los párrafos anteriores vimos que el análisis léxico puede usarse para *parsing* y para *generación morfológica*. Idealmente, el mecanismo usado para hacer *parsing* debería ser el mismo que el usado para la realizar la *generación* de variantes, para mantener un proceso claro. Sin embargo, en la práctica se ve que esto no es así, ya que los mecanismos usados para uno no son lo suficientemente eficientes para el otro, por lo que se implementan mecanismos distintos.

Otras dos dificultades se presentan con palabras morfológicamente complejas. La primera viene de la noción de que éstas son estructuras compuestas por un término invariable más añadiduras que indican cantidad, persona o tiempo, es muy idealista. En algunos lenguajes, la cantidad de combinaciones de transformaciones que un término puede tener, hace que sea muy difícil construir y mantener una lista para un lema. Por otro lado, en lenguajes como el Inglés, no alcanza con concebir a las transformaciones como `TRANSFORMACIÓN = TÉRMINO CANONICO + AÑADIDURA`. Por ejemplo, esta lógica aplica para la palabra ingresa en pasado simple *looked*, que puede representarse como `looked = look + {Pasado}`; pero no aplica para *saw,* un término en el mismo tiempo verbal. En el último caso, más que una añadidura, se necesita alternar la palabra.
La segunda dificultad que se presenta, es que en el contexto de una añadidura particular, no se garantiza que el *stem* sea equivalente a la forma canónica o lema. Pasando por el idioma Español, el plural de *lápiz*, *lápices*, no podría obtenerse con la lógica `LEMA + es`, ya que aplicarla nos daría por resultado "*lápices*". El algoritmo de análisis debe contar con la

información necesaria para saber qué *lápices* forma parte de las variantes de la forma canónica *lápiz.*

2.2.3 Análisis sintáctico

El análisis sintáctico se encarga de estudiar cadenas de palabras, es decir, oraciones. Su objetivo es detectar su estructura, y para eso se basa en gramáticas formales. Esta tarea no es un fin en sí misma, sino un paso intermedio necesario para etapas posteriores, específicamente, la etapa de análisis semántico, en donde se estudiará el significado del texto. El resultado obtenido en este subproceso de NLP, es típicamente una estructura jerárquica.

Cabe tener en cuenta que existen diferencias entre los algoritmos de análisis sintácticos de NLP y los de lenguajes de programación, cuyo funcionamiento es relativamente familiar para los estudiantes de informática.
Una de las diferencias más importantes está en la capacidad que tiene cada sintaxis para generar texto. Las sintaxis de lenguajes de programación están pensadas y diseñadas para poder ser analizadas en tiempo lineal, en función de la longitud del programa de entrada. Es por esto que para estos fines se utilizan clases restringidas de gramáticas libres de contexto. Para el caso de análisis sintáctico de texto de los lenguajes naturales, se requieren más recursos, dado que no existen tantas restricciones para las producciones posibles.

Una segunda diferencia radica en la ambigüedad del lenguaje natural. Teniendo en cuenta la oración:

Guarda la ropa en el canasto en el lavadero.

En este ejemplo, se nos presentan dos posibles interpretaciones:

- Que se debe guardar [la ropa en el canasto] en el lavadero, o
- Que se debe guardar la ropa [en el canasto en el lavadero]

Además, si se le siguen agregando frases siguiendo la misma línea de ambigüedad, la cantidad de análisis distintos crece exponencialmente. El problema en este caso radica en que sólo un análisis será correcto, dependiendo del contexto en el que la frase es utilizada. Si del análisis sintáctico se retornan todas las posibles estructuras para una frase ambigua, el uso de recursos (tiempo y memoria) del proceso se verá altamente perjudicado, teniendo en cuenta de todas las estructuras posibles, solo una vale para las etapas posteriores. Ante esta problemática, se dedica bastante esfuerzo en implementar soluciones eficientes de desambiguación.

Una tercera diferencia entre el análisis sintáctico para NLP y el utilizado para los lenguajes de programación, está la dificultad para identificar y extraer todas las posibles construcciones presentes en el lenguaje natural. Esta dificultad puede darse en parte por errores del autor en la producción del texto, como también porque la gramática no cubre todos los casos que se pueden presentar. En el caso de un lenguaje de programación, la especificación de la sintaxis siempre es completa, de manera de que que todas las posibles construcciones para ese lenguaje son identificables.
En el caso del lenguaje natural, es muy difícil entender si un error de análisis fue por un error del autor en su producción, o porque la gramática no cubrió un caso particular, que tranquilamente podría ser parte de una construcción válida para el autor.

Gramáticas

Pinker [11] considera que los textos deben ser analizados como estructuras jerárquicas en vez de secuencias “planas” organizadas en patrones. Además, distingue dos técnicas formales para modelar el conocimiento gramatical: las **gramáticas regulares**, y las **gramáticas libres de contexto.**

Una gramática define un lenguaje, describiendo cómo se puede generar una cadena a partir del mismo.
Una gramática formal es una cuádrupla *G = (N, T, P, S)* donde

- *N* es el conjunto finito de símbolos no terminales,
- *T* es el conjunto finito de símbolos terminales,
- *P* es el conjunto finito de reglas de producción,
- *S* es el símbolo inicial, donde $S \in (N \cup T)$.

Cada regla de producción en *P* tiene la forma:

$\alpha \rightarrow \beta$, $\alpha = \varphi A \rho$, $\beta = \varphi \omega \rho$

$\varphi, \omega, \rho \in (N \cup T)^*$

A es *S* o $A \in N$

Las **gramáticas regulares**, también conocidas como *de tipo 3*, generan lenguajes regulares, los cuales son reconocidos por un autómata finito. Se caracterizan por ser las gramáticas más restrictivas.

Las gramáticas regulares pueden clasificarse entre:

- Lineales por la derecha, si en el lado derecho de todas las producciones el símbolo no terminal aparece a la derecha de un símbolo terminal:

$$A \rightarrow aB \quad \text{ó} \quad A \rightarrow a \qquad \begin{cases} A \in N \cup \{S\} \\ B \in N \\ a \in T \end{cases}$$

- Lineales por la izquierda, si en el lado derecho de todas las producciones el símbolo no terminal aparece a la izquierda del símbolo terminal.

$$A \rightarrow Ba \quad \text{ó} \quad A \rightarrow a \qquad \begin{cases} A \in N \cup \{S\} \\ B \in N \\ a \in T \end{cases}$$

Las **gramáticas sensibles al contexto**, o *de tipo 1*, son las que generan lenguajes sensibles al contexto. A su vez, los lenguajes sensibles al contexto son reconocidos por autómatas linealmente acotados.

Las reglas de producción de las gramáticas sensibles al contexto se distinguen por ser de la forma:

$$\gamma A\beta \rightarrow \gamma\omega\beta \quad \begin{cases} A \in N \cup \{S\} \\ \gamma, \beta \in (N \cup T)^* \\ \omega \in (N \cup T)^* - \{\varepsilon\} \end{cases}$$

Las gramáticas sensibles al contexto fueron presentadas originalmente por Noam Chomsky, y desde ese entonces, tuvieron una gran relevancia a la hora de desarrollar formalismos que expliquen la sintaxis de un lenguaje.

Árboles de sintáxis

Como se mencionó anteriormente, el resultado típico de un análisis sintáctico suele ser una estructura jerárquica, llamada *árbol de sintaxis*, que representa las distintas aplicaciones de reglas gramaticales realizadas en la frase analizada. Esto quiere decir que cada nodo del árbol se descompone a partir de aplicar dichas reglas (ver Figura 6).

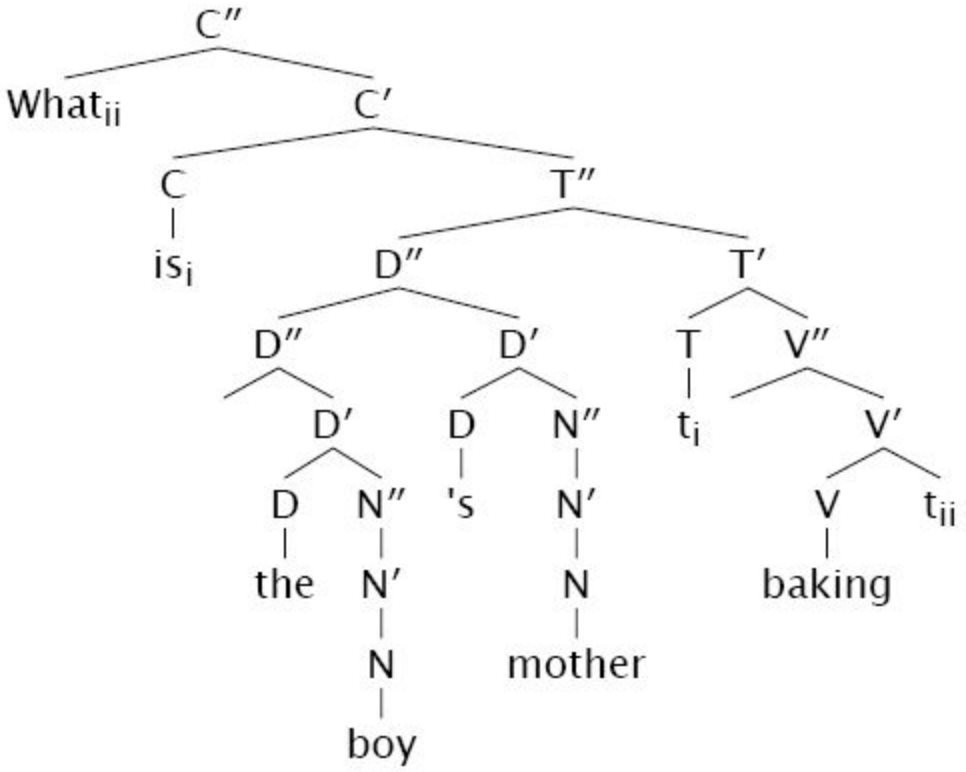

Figura 6. Árbol de sintaxis para la oración en Inglés "*What is the boy's mother baking".*

Se puede observar que los nodos hoja en este árbol contienen las palabras que conforman la frase, y los nodos no terminales contienen etiquetas que representan el tipo de palabra que se estaba analizando.

Características deseables

Robustez: es la habilidad de lidiar con entradas de texto que no cumplen con lo que normalmente se espera. Por ejemplo, en el caso de análisis sintáctico basado en el uso de gramáticas, las entradas "esperadas" son aquellas cadenas que puede generar una gramática G. Como se vió anteriormente, uno de los motivos de este obstáculo es que a veces una gramática puede no cubrir todo un lenguaje natural, o bien el texto de entrada puede tener errores de autor.

La robustez encara este problema buscando retornar siempre resultados, sin importar los motivos por los que la entrada no pertenece a L(G), es decir, el lenguaje generado por la gramática G. Se basa en el principio de que pequeñas desviaciones de la entrada esperada sólo llevan a pequeños obstáculos en el resultado, de manera que, aplicándola, un algoritmo de análisis sintáctico puede retornar siempre un resultado, con leve degradación en el resultado, en vez de simplemente detenerse ante un error, lo cual no sería de ninguna utilidad. Aunque el resultado pueda tener leves desviaciones e impurezas, siempre será más útil que un resultado vacío.

Sin embargo, se debe tener en cuenta que si el número de errores es grande, el resultado se degrada considerablemente, por lo que es importante también buscar controlar y mitigar la ocurrencia de los errores.

Suelen considerarse dos estrategias de análisis: *deep parsing* y *shallow parsing*. Cada una usa representaciones distintas para lograr el objetivo. En el caso de *deep parsing,* la estructura contiene información para resolver dependencias de "larga distancia", para resolver y entender las relaciones entre las palabras de manera directa. En cambio, *shallow parsing* usa representaciones más simples, y cumple su tarea asignando etiquetas a las distintas palabras que se le presentan, para luego identificar patrones y reglas en ellas. La información almacenada en *deep parsing* es mayor que la que requiere la variante *shallow.*

Otra manera de comparar a ambas estrategias es a través del grado de completitud del *parsing.* Se dice que el *deep parsing* realiza una análisis completo, mientras que *shallow parsing* realiza un análisis parcial. En análisis parcial, algunos elementos quedan sin procesar, y se delegan para ser vistos en pasos futuros.

Desambiguación

Anteriormente, vimos como pueden haber situaciones en donde una oración puede tener múltiples interpretaciones válidas. Vimos que con un algoritmo robusto, se puede obtener al menos un resultado. Ahora bien, con la Desambiguación, lo que se busca obtener es solamente un resultado.

Es cierto que, de haber varios resultados obtenidos en un análisis sintáctico para una misma oración, sólo se puede desambiguar con total certeza aplicando un estudio contextual, que no forma parte de esta etapa de NLP, sino más bien de pasos posteriores. Sin embargo, durante el análisis sintáctico se puede reducir la información obtenida, y de esa manera achicar también la cantidad de posibilidades. De esta manera, de N resultados iniciales, pueden pasar dos cosas: podemos quedarnos con los M mejores resultados (M<N) o bien, puede quedar un sólo resultado, que luego puede ser pasado a las etapas siguientes del proceso, para un análisis más profundo que abarca su significado.

Una forma simple de desambiguar en esta etapa, es aplicando dominios; si bien de un análisis sintáctico pueden obtenerse múltiples interpretaciones para una misma frase, si ponemos a esas interpretaciones en el contexto de un dominio en particular, muchas de ellas carecerán de sentido. Si a priori se conoce el dominio del texto que se está procesando, podrían filtrarse muchas interpretaciones, y de esa manera llegar a un resultado más óptimo.

Un enfoque simple para aplicar desambiguación fué descrito por Burton [23], y consiste en desarrollar una gramática semántica por cada dominio. Una alternativa más sofisticada consiste en configurar la gramática para cada dominio nuevo. Este enfoque fue presentado como *gramática especializada* por Grishman [24]. El método nace de la conclusión de que, en un determinado dominio, se suelen combinar algunas reglas gramaticales, más que otras. Identificando estas combinaciones, y agrupandolas por dominio, se obtiene una gramática con una mayor cantidad de reglas, y por lo tanto mayores posibilidades de reducir o directamente eliminar la ambigüedad.

2.2.4 Análisis semántico

En el campo de la lingüística, el análisis semántico se basa en analizar el significado de las palabras, las oraciones, y sus contextos. Existen dos categorías dentro del análisis semántico.

Por un lado se tiene la semántica léxica, que invierte sus esfuerzos en estudiar el significado de las palabras y las combinaciones que se realizan con ellas. La otra categoría se llama semántica combinacional o composicional, que se centra en las cadenas de palabras de longitud indefinida, es decir, frases y oraciones.

Se reconocen dos desafíos a superar o evitar en este campo: la **circularidad** y el **retroceso infinito**.

La **circularidad** puede explicarse con la presencia de definiciones circulares en los diccionarios. En una definición circular, se supone una comprensión anterior del término que es definido. Por ejemplo, si buscamos en el diccionario la definición de la palabra *nogal*, encontramos:

***nogal**: Planta arbórea de tronco robusto del que salen gruesas ramas, con copa grande y redondeada, hojas puntiagudas y aromáticas, con flores blanquecinas, cuyo fruto es la **nuez**.*

A su vez, si buscamos la definición de nuez, encontramos:

***nuez**: "Fruto del **nogal**, de forma ovoide, con el epicarpio fino y liso, de color verde, con el mesocarpio correoso y caedizo, el endocarpio duro, rugoso y de color pardo, dividido en dos mitades que encierran la semilla comestible y oleaginosa".*

Vemos como nuez se define en términos del nogal, y visceversa.

A su vez, el retroceso infinito se da cuando se quiere explicar un escrito en un lenguaje L con un metalenguaje M_1, y para explicar a M_1 se requiere del soporte de M_2, pudiéndose así extender (teóricamente) la secuencia de metalenguajes de manera infinita.

La mayoría de autores coinciden en que la solución más óptima es encarar el análisis con un conjunto bien definido de elementos primitivos. Lamentablemente, no se ha llegado a un consenso en cuanto a la naturaleza de esos elementos. Se han propuesto distintos enfoques, muchos antagonistas entre ellos: se ha propuesto tanto que los elementos primitivos sean específicos de un lenguaje, como también se ha deslizado la idea de usar elementos universales a todos los lenguajes. También se impulsaron las ideas de que dichos elementos fueran procedentes del lenguaje natural, o sino inventados por el analista mismo.

Una de las principales misiones del análisis semántico dentro de NLP es la de resolver ambigüedad. Vimos que el análisis sintáctico puede especializarse en dominios, para reducir la cantidad de interpretaciones posibles, y de esa manera aminorar el grado de ambigüedad. Sin embargo, las capacidades del análisis sintáctico tienen un límite, y puede no llegar a encontrar la única interpretación entre varias que vale para el proceso de NLP.

Ambigüedades como factores de error

Las producciones escritas por humanos pueden tener múltiples interpretaciones, debido a tres tipos de ambigüedad. A continuación describimos cada una de ellas.

Ambigüedad léxica

La ambigüedad léxica surge de los distintos significados de las palabras. Esto se da sobre todo por cualidades que pueden tener las palabras: la homonimia y la polisemia. La **homonimia** se da cuando dos o más palabras tienen la misma forma, sea la misma pronunciación oral o la misma escritura. Por ejemplo, hay homonimia en la palabra *nada*, ya que significa tanto “ninguna cosa” y también es una forma del verbo que significa la acción de trasladarse en el agua.

La **polisemia**, en cambio, se produce cuando una misma palabra tiene varias acepciones o significados. Por ejemplo, la palabra *cresta* es polisémica, ya que acepta las siguientes definiciones:

Cresta:

- Parte del cuerpo de algunos animales que crece generalmente sobre la cabeza.
- Cumbre de una ola.

De ambas cualidades, la que genera mayor problemática de ambigüedad es, generalmente, la polisemia, debido a las diferentes interpretaciones de una palabra, y la facilidad para fallar en detectar la definición correcta.

Ambigüedad de alcance

La ambigüedad de alcance [25] es la que ocurre cuando dos cuantificadores o expresiones similares pueden tener alcance una sobre otra de diferentes maneras.
Un ejemplo de ambigüedad de alcance puede demostrarse con la frase:

"Todo hombre ama a una mujer."

El significado más prominente, es que, por cada hombre hay una mujer, y que es posible que cada hombre ame a una mujer diferente. Sin embargo, la frase tiene otro significado, que dice que todos los hombres aman a una misma mujer en particular. Este significado se podría desambiguar fácilmente si la frase terminara aclarando de qué mujer se está hablando: *"Todo hombre ama a una mujer, llamada Lucía."*
Puede darse otro ejemplo con una broma [26], compuesta por dos frases: una que hace uso de la ambigüedad, y otra que aclara el significado original de la frase:

"En Argentina, una mujer da a luz cada 48 segundos. Debe estar exhausta."

La primera interpretación que una persona le daría a la primer frase, es que una mujer diferente da a luz cada 48 segundos, pero analizando detenidamente, la frase también podría significar el caso sumamente improbable de que existe una mujer en Argentina que da a luz 1800 veces por día.

Ambigüedad referencial

La ambigüedad referencial se da cuando una palabra o frase puede referenciar a dos o más propiedades o cosas. A veces, teniendo en cuenta el contexto, es suficiente para saber a cuál de todas las cosas se refiere, pero esto no siempre aplica.

Por ejemplo, considerando la siguiente frase:

“Pavarotti fue una gran estrella de Ópera.”,

podemos encontrar ambigüedad en la palabra *gran*, ya que podría referenciar al hecho de ser corpulento, o al hecho de ser famoso.

Enfoques para la representación semántica

Existe una considerable variedad de estrategias para representar el resultado de un análisis semántico. A continuación nombramos dos que nos llamaron la atención, dado que son relativamente recientes, y porque tienen una visión innovadora sobre cómo entender el significado de manera programática.

Uno de ellos es **El metalenguaje semántico natural** (NSM por sus siglas en Inglés) es un sistema de representación de composicional que usa un metalenguaje, el cual es un subconjunto del lenguaje natural, junto a un subconjunto de definiciones de palabras, más un subconjunto de sus propiedades sintácticas. Las palabras incluidas en este metalenguaje suelen no ser muchas, y son llamadas *palabras principales* (*primes* según la bibliografía consultada), ya que son aquellas cuyo que están presentes en todos los lenguajes, es decir que son conceptos “universales” y comunes, básicos para la producción de la mayoría de los textos.

A continuación mostramos la figura 7, que muestra una tabla de primes para el lenguaje Inglés:

I, YOU, SOMEONE, SOMETHING/THING, PEOPLE, BODY	Substantives
KIND, PART	Relational substantives
THIS, THE SAME, OTHER/ELSE	Determiners
ONE, TWO, SOME, ALL, MUCH/MANY	Quantifiers
GOOD, BAD	Evaluators
BIG, SMALL	Descriptors
KNOW, THINK, WANT, FEEL, SEE, HEAR	Mental predicates
SAY, WORDS, TRUE	Speech
DO, HAPPEN, MOVE, TOUCH	Actions, events, movement, contact
BE (SOMEWHERE), THERE IS, HAVE, BE (SOMEONE/SOMETHING)	Location, existence, possession, specification
LIVE, DIE	Life and death
WHEN/TIME, NOW, BEFORE, AFTER, A LONG TIME, A SHORT TIME, FOR SOME TIME, MOMENT	Time
WHERE/PLACE, HERE, ABOVE, BELOW, FAR, NEAR, SIDE, INSIDE	Space
NOT, MAYBE, CAN, BECAUSE, IF	Logical concepts
VERY, MORE	Intensifier, augmentor
LIKE/WAY	Similarity

Figura 7. Tabla de palabras *prime* para el lenguaje Inglés, para usar como metalenguaje semántico natural.

NSM es una teoría cognitiva, cuyos adherentes defienden con la idea de que las palabras simples y ordinarias son un mejor medio de representación para tareas cognitivas, que las palabras de mayor nivel técnico de otros metalenguajes pertenecientes a otras teorías.

Los defensores de NSM también esgrimen que cualquier sistema de representación necesariamente tiene que estar fundado en lenguaje ordinario; y se oponen al uso de términos formales, tecnicistas y sofisticados, ajenos al uso cotidiano que se le da al lenguaje. Según esta teoría, la comunicación de cualquier forma, debería poder ser comprendida usando el lenguaje cotidiano.

Si bien el enfoque de NSM es uno de los mejores enfoques que se hayan desarrollado en los últimos años, hasta ahora su uso en NLP es mínimo.

El modo de representación de NSM es la explicación semántica, la cual consiste en parafrasear de manera reductiva. Esto significa que la explicación semántica intenta decir lo que un orador dice, pero con otras palabras, simplificando las formas, pero manteniendo el mismo significado de lo que el orador quiso decir.

Como desventaja, al enfoque de NSM se le reconoce que, para muchas palabras, no es posible explicarlas solo con *palabras principales,* o *primes*, por lo que se dificulta la reducción parafrástica para casos en donde se usen esas palabras. En estos casos, para explicar palabras complejas se necesita usar *primes* semánticos junto a significados léxicos complejos, llamados por los autores "moléculas semánticas", las cuales son conceptos que pueden anidarse unos a otros, creando cadenas de dependencia semántica.
Por ejemplo, las explicaciones para las palabras *correr* y *caminar* requieren de las moléculas semánticas *pie* y *suelo*. Pueden haber moléculas universales, pero muchas terminan siendo específicas de un lenguaje o una cultura.
El concepto de moléculas, incluyendo la opción de poder anidar conceptos, permite comprimir la complejidad semántica, y simplificar la comprensión del significado, ya que para entender todo lo que una palabra expresa, basta con seguir las distintas cadenas de dependencia semántica que la componen. Esta idea es realmente innovadora y muy diferente a otros enfoques tradicionales y más complejos, como los estructuralistas y los que se basan en reglas lógicas.

Otro enfoque interesante es el de **semánticas orientadas a objetos** (SOO), ya que es muy reciente. Hasta ahora se usa restringidamente en la representación de significados en algunos campos de la semántica, pero en tiene un futuro prometedor y podría ser de gran relevancia si se aplicara a NLP. Esto es en parte porque en el campo de las ciencias de la computación, se cuenta con décadas de investigación sobre sistemas orientados a objetos. De todo este conocimiento, ya sabemos que el sistema cognitivo humano reconoce mejor la realidad cuando se la representa con entidades, que tienen comportamiento y propiedades, y que se encuentran relacionadas con otras entidades, e interactúan entre sí. Esto se corresponde básicamente con la descripción de la **programación orientada a objetos**, en donde el concepto de objeto es central. Las similitudes entre este paradigma de programación y la forma en que el sistema cognitivo se organiza sugiere el uso de esta manera para representar significado dentro del análisis semántico.

Schalley [28][29] sugiere una representación a la que llama UER (Unified Eventity Representation), que se basa en el Lenguaje de Modelado Unificado, o UML, por sus siglas en Inglés. UML es un lenguaje ampliamente conocido en el campo de las ciencias de la computación, ya que es el estándar para representar sistemas pensados bajo un paradigma

Orientado a Objetos, a través de diagramas. Uno de los puntos fuertes de UML es su vasta variedad de elementos visuales para definir y explicar sistemas hechos en POO. UER aprovecha esa fortaleza, adoptando la naturaleza gráfica de los diagramas UML. Con UER se pueden definir entidades, relaciones, agregaciones, etc, permitiendo obtener así una representación léxica detallada.
Actualmente se continúa trabajando en formas de representación semántica orientadas a objetos. Uno de los desafíos más importantes para que este enfoque continúe avanzando, es el de alcanzar la habilidad de representar los significados verbales de manera detallada.
La estructura del significado verbal, debe ser instanciada puesta en contexto, para su total comprensión. Se cree que los enfoques orientados a objetos pueden adaptarse a esto perfectamente, debido a su dicotomía clase/instancia.

2.2.5 Análisis pragmático

La pragmática [30] se enfoca en estudiar la intención de los productores y receptores de una expresión escrita u oral, por lo que depende fuertemente en entender el contexto en el que las expresiones son emitidas. Esto indica que el análisis pragmático, a diferencia del sintáctico y semántico, va un paso más allá de la estructura del lenguaje.

El análisis pragmático se concentra en la teoría del acto de discurso, en el estudio de las conversaciones, y en las diferencias culturales, que juntas hacen a la comunicación y comprensión por parte de locutor e interlocutor, respectivamente. Es clave el estudio del uso del lenguaje con fines comunicativos, que nace de manera intencional de las personas, ya desde los primeros años de edad.

2.3 Análisis de sentimiento / Minería de opiniones

2.3.1 ¿Qué es?

Una opinión es un juicio o punto de vista personal y subjetivo sobre cualquier entidad cuestionable. Se caracteriza por ser un pensamiento individual abierto a la disputa frente a opiniones de otras personas Existen 2 tipos de opiniones: las *regulares* y las *comparativas*. Las opiniones *regulares* son aquellas emitidas sobre una sola entidad, mientras que las *comparativas* comparan dos o más entidades, destacando las similitudes y diferencias entre ellas. A veces pueden expresar la preferencia del autor por alguna de las entidades comparadas.

Dos conceptos importantes a la hora de analizar opiniones son los de **subjetividad** y **emoción**. Una frase subjetiva es aquella que expresa sentimientos personales o creencias, sin respaldo ni evidencia fáctica que los acredite como verdaderos o definitivos. La emoción, en cambio, es la manera en que los individuos se sienten respecto a objetos observables. Se dice que la emoción puede dividirse en 6 categorías: un individuo puede sentir amor, odio, sorpresa, miedo, rabia o alegría frente a las entidades que se le presenten.

Habiendo repasado estos conceptos, se presenta una definición sobre opiniones.

Una **opinión** puede verse como una quíntupla conformada por:

1. una **entidad** objeto de la opinión
2. un **aspecto** de aquella entidad,
3. el **titular**, o quien emite la opinión,
4. la **orientación** de dicho individuo acerca del atributo de la entidad. La orientación puede tener un **valor** positivo, neutral o negativo, y se pueden usar distintos niveles de intensidad dentro de esas categorías. Si la opinión cubre todos los atributos de la entidad, se dice es de carácter general, y
5. el tiempo en que fué emitida la opinión.

Estos cinco componentes son los escenciales para estudiar una opinión. Sin embargo, dependiendo del estudio que se realice, pueden agregarse otros elementos que sean de interés, como pueden ser el sexo y edad de cada persona dueña de la opinión. Estos elementos nos sirven como base para dar estructura a datos textuales, que como se mencionó previamente, son no estructurados.

La **minería de opiniones**, también conocida como análisis de sentimientos, es el campo que estudia opiniones, sentimientos, evaluaciones, valoraciones, actitudes y emociones de personas acerca de entidades tales como productos, servicios, individuos, organizaciones, eventos, y temáticas. Hoy en día se utilizan los términos "minería de opiniones" y "análisis de sentimientos" de manera indistinta, y también se han usado otros nombres para tareas ligeramente distintas, como *extracción de opiniones*, *minería de sentimientos*, *análisis de subjetividad*, *análisis de emociones,* y *análisis de revisiones* [6].
Este campo de estudio tiene sus orígenes en el Procesamiento de Lenguaje Natural, o NLP, por sus siglas en Inglés. Las técnicas de NLP se había estudiado y desarrollado bastantes años antes, pero la minería de opiniones, como rama de NLP, produjo interés recién a partir de la década del 2000. De hecho, el término minería de opiniones aparece por primera vez en la publicación *Dave et al* [1], presentada en una conferencia sobre la Web en el año 2003.

El interés por la minería de opiniones radica en la posibilidad de obtener **grandes volúmenes de información** sobre lo que un sector de la población piensa **en ese momento**. Resaltamos la cantidad de información, ya que hay una diferencia notoria frente a los métodos tradicionales, como en las encuestas. Una encuesta cuesta muchos recursos cuando se intenta llegar a un número significativo de la población. El costo radica no solo en la cantidad de encuestas que se pueden realizar, sino en la amplitud geográfica de la misma. Con minería de opiniones, prácticamente se puede saber lo que piensa cualquier persona en cualquier parte del mundo, con muy bajo costo.
También destacamos el tiempo en que es obtenida la información (el mismo momento en que se minan opiniones). Volvamos al caso de la encuesta: la realización de ésta puede llevar varios días. No solo se debe preguntar a los encuestados, sino que se debe recopilar toda la información y analizarla; esto suma aún más tiempo. Es posible que, al momento de tener el resultado final de la encuesta, hayan pasado bastantes días, y la misma no refleje 100% la realidad actual. Algunos de los encuestados pueden haber cambiado su opinión. Imaginemos el

caso en que se preguntó qué opinaba la población sobre un político, y que, en el tiempo en que se analizaban los datos y se producían los resultados de la encuesta, se descubre un caso de corrupción que lo involucra. Es muy probable que lo que la gente opine sobre él cambie, por lo que el resultado de la encuesta pierde credibilidad.
Si se hubiese usado Minería de Opiniones, no solo se podría haber obtenido el resultado en el momento previo al destape del caso de corrupción, sino que se podría hacer un seguimiento del impacto, comparando las métricas de antes y después del descubrimiento, sin tener que distribuir nuevamente encuestadores.

Técnicamente hablando, el análisis de sentimientos se encarga de descubrir varias instancias de la quíntupla previamente descrita en un conjunto de documentos escritos *D*. Para eso, se necesita realizar una secuencia de tareas, a saber:

1. Extraer las entidades y agrupar los sinónimos en clusters. Cada cluster obtenido representa a una entidad única en el proceso de minería.
2. Extraer todos los aspectos de las entidades obtenidas en el paso anterior, y agruparlos en clusters de aspectos para cada entidad.
3. Extraer al titular y el momento en que emitió la opinión. Este paso puede ser relativamente sencillo, dado que la obtención de estos datos va por fuera del procesamiento del texto, y en algunas redes sociales, pueden extraerse de una estructura de datos definida.
4. Determinar si la opinión tiene orientación positiva, neutral o negativa.
5. Generar la quíntupla, basándose en los elementos extraídos en los pasos previos.

Como en un estudio de estadística, en minería de opiniones se requiere analizar un gran volumen de datos, que nos indique generalidades sobre lo que piensa un sector de la población respecto a una entidad. Es por este carácter general que en estas aplicaciones se desea contar con una suerte de resumen de los resultados. Una forma muy popular es la de resumen de aspectos, en la cual, se recolectan opiniones y se clasifican según el aspecto de la entidad que tratan. Pueden haber opiniones generales, esto es, que tratan sobre la entidad como un todo y sin hacer hincapié en algún aspecto en particular. Luego, se determina la orientación de la

opinión, y se contabilizan. Finalmente, se obtiene un resumen como el siguiente ejemplo, para un modelo de guitarra de una conocida marca:

Fender Telecaster
Aspecto: General
Positivo: 200 <comentarios>
Negativo: 6 <comentarios>
Aspecto: Sonido
Positivo: 304 <comentarios>
Negativo: 2 <comentarios>
Aspecto: Precio
…

Este tipo de resumen es muy útil para presentar en forma de gráfico de barras o de torta, en algún tablero de control en ámbitos como la gerencia de una empresa que quiere averiguar sobre qué opinan los clientes sobre sus productos. Este tipo de presentación permite reconocer esto último rápidamente, concentrándose en el volúmen y en las tendencias, más que en las individualidades.

Otra manera de resumir, es describiendo textualmente los resultados de contabilizar opiniones por aspecto, como resumen cuantitativo: *"Al 60% de nuestros clientes les gusta el producto, mientras que a un 40%, no les gusta, por diversos motivos".*

Existen diversas formas de aplicar minería de opiniones, algunas más sofisticadas que otras, y con distintos objetivos, como se podrá observar. Naturalmente, cada una tiene aspectos en donde se destacan, como también sus puntos flacos. En este capítulo vamos a tratar los que nos parecieron los más relevantes.

2.3.2 Usos

Clasificación de sentimientos

En la clasificación de sentimientos, se intenta extraer la opinión de escritos como puede ser una reseña sobre un producto, un artículo, una columna periodística, etc. En este caso, se

toma al documento entero como una unidad de datos de donde se puede extraer información relevante, y se intenta detectar la opinión. Algunas de las técnicas usadas para esta tarea, es la de aprendizaje supervisado. En este tipo de algoritmos, los datos de entrada son llamados 'datos de entrenamiento', y se manejan etiquetas conocidas, para reconocer, en este caso, palabras negativas o positivas . A partir de estos datos de entrenamiento, se prepara un modelo en un proceso en el que se hacen predicciones, y se realizan correcciones cuando esa predicción es errónea. El proceso continúa hasta que se obtiene un modelo con un grado de precisión deseado para realizar esas predicciones.
Los datos de entrenamiento utilizados son artículos para los cuales ya se determinaron la orientación. Suelen usarse calificaciones del 1 al 5 para determinar que un artículo es positivo o negativo

Los algoritmos para clasificar opiniones, deben cumplir con las siguientes características:

- Deben poder detectar los términos mencionados con mayor frecuencia. En algunos casos es importante tener en cuenta la posición donde aparecen esas palabras dentro de una frase.
- Deben tratar especialmente a los adjetivos presentes en el texto, dado que se comprobó que son importantes en el reconocimiento de sentimiento.
- Tratar las *palabras de opinión*, que son aquellas que se usan regularmente para expresar sentimientos positivos o negativos. Algunos ejemplos son las palabras *bueno, malo, pobre, terrible, excelente*. También es importante tratar palabras que no son adjetivos, y también frases, que son también formas de representar opinión.
- La detección de negaciones es también importante, ya que su sola presencia en una frase puede cambiar la orientación de la opinión. Sin embargo, deben tratarse con cuidado, porque en una frase puede palabras de negación, como "*no"*, pero sin embargo la orientación puede no ser negativa.
- Por último, se debe considerar la dependencia sintáctica entre palabras.

Además de clasificar opiniones entre negativas y positivas, también se investiga cómo predecir el puntaje de revisiones.

Existen estudios orientados a la **clasificación no supervisada**, que buscan extraer la opinión semántica. Esto significa que no toma palabras aisladas para hacer análisis, sino que busca analizar la dependencia entre ellas para detectar la orientación. Concretamente, se buscan frases con adjetivos y adverbios, y se observa la posición de éstos en la misma, siguiendo una tabla de reglas. Es decir, la frase se extrae si detecta que se cumplen esas reglas de posición. Dados los adjetivos y adverbios, se estima la orientación usando una ecuación. Se debe tener en cuenta que el resultado es una probabilidad, por lo que puede no obtenerse la verdadera orientación de la frase.

Detección de subjetividad

Otra variante es la **detección de la subjetividad,** en la cual se busca determinar si una frase es subjetiva u objetiva. Nótese que en la variante anterior, lo que se buscaba era ver la opinión en una reseña o artículo. En este caso, se pone el foco en una sola oración. Para lograr esto, se realizan dos tareas. Primero, se estima si la frase es subjetiva o no, y luego, se extrae la opinión. Estos pasos suelen ser útiles para descartar aquellas frases que no sirven para el estudio, ya que no contienen opiniones.

Generalmente, en la subjetividad de oraciones también se utilizan algoritmos supervisados, ya que son problemas de clasificación, y suele trabajarse desde el supuesto de que la frase será lo más simple posible: que contendrá una opinión sobre una cosa, y tendrá un solo opinador.
Por supuesto, el punto flaco en esta suposición es la dificultad de analizar frases más complejas, con más de una opinión sobre más de un atributo. Algunos autores resaltan que en una misma frase puede haber componentes subjetivos y otros fácticos, por lo que la complejidad para implementar un método exitoso, aumenta.
Se han hecho múltiples estudios tomando conversaciones en foros de discusión en la Web. Los usuarios de foros, no solo expresan sus opiniones, sino que también interactúan entre sí, por lo que el contenido es de gran interés para su análisis en términos de hacer minería de opiniones.

Otra estrategia más simple, es la de aplicar **minería basada en un diccionario**. Básicamente, se trata de obtener a mano un pequeño conjunto de palabras que expresen opinión de un diccionario, generalmente sacado de Internet. Al hacer esa extracción de manera manual, se puede asignar la orientación de esas palabras. Con esa base, se hace crecer al conjunto de

palabras conocidas, minando diccionarios, en búsqueda de sinónimos y antónimos. Se puede hacer una revisión manual, al finalizar de ejecutarse el algoritmo, para eliminar las palabras cuyas orientaciones fueron mal asignadas. El diccionario resultante, puede utilizarse junto con otros métodos de minería supervisados.
Como puede observarse, este método es muy sencillo, aunque poco potente, además de es dependiente de supervisión manual. Además, carece de la capacidad para adaptarse a distintos dominios.

Existe una enfoque alternativo, **basado en cuerpos**, que tiene la potencia como para detectar el dominio, y adaptar cómo detecta la orientación. Este método busca y detecta patrones, además de una lista de palabras que reflejan opinión, que sirvan como base para detectar otras palabras adicionales en los textos. Esta detección se realiza cuando adjetivos y adverbios no "conocidos" están unidos a otros "conocidos" mediante palabras conjuntivas. Por ejemplo, si dentro de la lista de palabras conocidas, se tiene al adjetivo *resistente*, y el algoritmos se encuentra con la frase:

La carcasa de este celular está hecha con material resistente y duradero.

Entonces, tomará la palabra *duradero* y la agregará a la lista de palabras conocidas. La estrategia de tomar palabras porque se encuentran en conjunto con otras también sirve para determinar si expresan opinión positiva o negativa. De esta manera el algoritmo aprende nuevas palabras junto a su orientación, y aumenta su capacidad de interpretación. Por supuesto, esta idea no siempre es del todo consistente.

Por ejemplo, para el campo de los celulares, podemos encontrar frases como
La duración de la batería es larga y puede alcanzar hasta dos días sin recarga.

Como también, frases como:

El tiempo de respuesta para abrir aplicaciones es más largo que en sus competidores.

Entonces, puede observarse cómo la palabra *largo*, indica una opinión positiva o negativa, dependiendo del atributo del objeto que se está describiendo. Es por esto que un enfoque más

completo debe determinar la orientación de los adverbios o adjetivos de opinión, a partir del aspecto o atributo con el que están asociados.
Una contra en el enfoque basado en cuerpos, es que no es tarea fácil contar con un conjunto de datos lo suficientemente grande para que cubra todas las palabras de un idioma. El enfoque basado en diccionario, al almacenar menos datos, es una alternativa que sí podría cubrir las palabras de un idioma. Sin embargo, en donde el enfoque basado en cuerpos se destaca, es en la detección de contextos y dominios en donde aparecen las palabras procesadas.

En los inicios de este capítulo se había separado a las opiniones en dos categorías: regulares y comparativas. Hasta ahora, hemos desarrollado métodos de minería de opiniones regulares; pero debe tenerse en cuenta que también existen variantes para las opiniones comparativas.
Es importante resaltar que ambas formas de opinión difieren en cuanto a semántica y a su forma sintáctica. En las opiniones comparativas, siempre se pondera un atributo de una entidad, poniéndolo en contraste con el mismo atributo para otra entidad del mismo dominio. Las comparaciones pueden hacerse dos o más entidades, o a veces entre una misma entidad y versiones anteriores de la misma.

Este tipo de comparaciones es muy común en artículos tecnológicos, en donde se ponen a prueba las capacidades de dispositivos que son competencia entre sí dentro de un rubro, para finalizar con una conclusión respecto a los puntos en donde cada uno resalta.
En general, en una una oración comparativa se observa la relación entre dos objetos, basándose en similitudes o diferencias. Existen dos clases de comparaciones, la regular, en donde lo que se busca reflejar es que un objeto tiene más, menos o lo mismo de un atributo que otro objeto; y la superlativa, en donde se quiere definir que una entidad tiene el mayor o menor atributo, de un conjunto de entidades similares.

Identificación de preferencias

La identificación de preferencias se basa en el estudio de opiniones comparativas, de manera que ellas se están ponderando las propiedades de dos o más objetos o entidades. El objetivo de este tipo de análisis es el de reconocer por qué variante se inclinan las preferencias del locutor.
Si bien esta es una aplicación de minería de opiniones, como el texto a analizar es una comparación, posiblemente no exprese una opinión directa positiva o negativa; sino que la

opinión estará presente de una manera más sutil. Es por esto que, aplicar una clasificación de sentimientos directa, no tiene mucho sentido. En cambio, se podría buscar la preferencia a partir de la valoración que el locutor hace de atributos compartidos entre los objetos o entidades en cuestión.

Hasta el día de hoy no se realizaron suficientes estudios sobre identificación de preferencias, por lo que no se conocen muchos métodos para identificar preferencias en lenguaje natural. Un estudio fué realizado por Liu [34], quien llegó a proponer una metodología. La misma se vale de las palabras de opinión que típicamente se usan para comparar. Estas palabras pueden dividirse en dos categorías:

1. **Palabras comparativas**: son palabras comparativas que fácilmente permiten reconocer preferencias. Ejemplos son *mejor, peor, eficiente*, etc. También entran en juego los adjetivos a los que se les agregan adverbios tales como *más, menos*; aunque para este otro caso, se deben aplicar algunas reglas para no caer en falsos positivos.

2. **Palabras comparativas dependientes del contexto**: En este caso entran las palabras que se usan en frases en donde la interpretación de la opinión es ambigua, y depende del contexto. Por ejemplo, observemos la frase "*La durabilidad de la batería del celular X es mayor que la del celular Y".* En este caso, el objeto de comparación es la batería de dos teléfonos celulares distintos, y la palabra que compara es *mayor*. En este caso, sin conocimiento en el dominio de los teléfonos celulares, un algoritmo no sería capaz de saber si *mayor,* en este caso, es positivo o negativo. El conocimiento adecuado, aplicado a las palabras *mayor*, *duración* y *batería* son las que determinan la opinión. En esta categoría también entran las comparaciones con adverbios que refuerzan o describen a los adjetivos.

La propuesta de Ganapathibhotla y Liu, se basa en preparar a los algoritmos de minería con suficientes artículos y reseñas comparativas sobre un dominio en cuestión, como para que estos estén entrenados lo suficientemente como para conocer qué atributos suelen compararse, y qué adjetivos suelen estar emparejados a esas propiedades, conociendo el el valor en esa combinación.

Una vez entrenados, los algoritmos recorren el texto en búsqueda de palabras comparativas, que necesariamente entran en las dos categorías mencionadas previamente. Una vez

obtenidas las palabras de comparación, se simplifican en lo posible con la siguiente observación:

"Si un adjetivo o adverbio es positivo, o negativo, entonces su forma comparativa o superlativa también será positiva o negativa."

Un ejemplo de simplificación tomando por ejemplo esta observación puede encontrarse en las palabras *bueno* y *mejor*. Esta simplificación puede hacerse de manera automática con diccionarios de sinónimos y bases de datos léxicas. Una vez simplificadas, las palabras pueden categorizarse entre comparativas crecientes o decrecientes.

Una vez que se obtiene toda esta información, detectar el objeto de preferencia es simple. Si hay una comparación positiva sin negación, entonces el primer objeto mencionado es el preferido. Si hay una comparación negativa sin negación, entonces el segundo objeto mencionado es el preferido.

Detección de spamming de opiniones

El *spamming* en la Web, consiste en realizar acciones con la intención de engañar a los motores de búsqueda a través de medios ilegítimos, para posicionar a algunas páginas en lugares más altos de los que les corresponderían. Esto redunda en una degradación de las búsquedas. El motivo detrás del *spamming* en la Web es económico. Si un sitio que vende un producto, se encuentra en posiciones bajas en una búsqueda, tiene muy bajas probabilidades de ser hallado por un usuario. El posicionamiento depende de los algoritmos que emplea el motor de búsqueda, y generalmente se basan en la popularidad del sitio, y principalmente, de que el sitio pague por estar en lugares con mayor exposición.
Existen muchas empresas que se dedican a ayudar a otras a obtener mayor exposición en los resultados de motores de búsqueda, explotando algunas debilidades de los algoritmos que usan. Existen algunas empresas que usan medios éticos, pero en este caso nos interesamos por las que no lo son, ya que emplean *spamming* en la Web.

Respecto a las opiniones, ya hemos mencionado que con la llegada de la Web 2.0, se hizo común que los usuarios compartieran contenido en sitios, de manera que se hizo frecuente

que, alguien que quisiera averiguar sobre un producto o servicio, acceda a estas páginas y lea las opiniones y comentarios de otras personas. Por ejemplo, imaginemos que una persona desea viajar a México. Nunca fué, por lo que quiere averiguar los costos y las experiencias de otros viajeros que sí hayan viajado a dicho destino. Puede acceder a sitios como **tripadvisor.com**, y buscar comentarios sobre hoteles en la zona que desea visitar. Naturalmente, si el futuro viajero ve que la mayoría de los comentarios son positivos, probablemente compre los pasajes y reserve un hotel. Las opiniones en la Web adquieren un gran valor, ya que pueden ser determinantes a la hora de tomar una decisión o formar una opinión respecto a un objeto.

Para corroborar esto, en los últimos años se estudió [31] la determinación de las opiniones en la toma de decisiones de los usuarios. Lo que se buscó fue desarrollar una forma de reconocer el grado de utilidad de un comentario. Teniendo esta capacidad, se podrían mostrar los comentarios y revisiones que más le sirvan a un usuario en las primeras posiciones de una consulta. Algunos sitios, como **Amazon** o **StackOverflow** implementan estos resultados personalizados mediante el *feedback* de los usuarios. Para cada revisión, los usuarios tienen la opción de calificarla o indicar si les resultó útil o no. De esta manera, estos sitios pueden ordenar los resultados en función de la cantidad de votos positivos que tengan, y hasta pueden agregar información a cada uno, con comentarios del estilo *"a 32 personas les resultó útil"*.

La importancia que cobran los comentarios en la red, han llevado al uso poco ético del *spamming* de opiniones. Esta actividad consiste en generar comentarios con el objetivo de engañar y despistar a los lectores o a los algoritmos de minería de opinión, dando opiniones inmerecidas de carácter positivo sobre una entidad (producto, persona, servicio, institución, etc.), de manera instalar un sentimiento colectivo falso respecto a esa entidad. Lo mismo puede hacerse por la negativa, es decir, generando comentarios negativos, para difamar la reputación de una entidad objetivo. Las opiniones generadas con estos fines son también llamadas falsas opiniones.

En la medida en que más usuarios accedan a esta modalidad para formar opinión respecto a algo, más crítica será la tarea de identificar y controlar esta problemática.

Mencionamos que la explotación de minería de opiniones puede ser determinante para hacer que más personas decidan comprar un producto o adquirir un servicio; esto redunda en beneficios económicos. Sin embargo, esta acción puede llevarse a cabo con fines políticos.

Durante el desarrollo de este trabajo, se llevaron a cabo las elecciones presidenciales de Estados Unidos, cuyo ganador fue el candidato republicano Donald Trump. Durante las semanas posteriores a conocerse el resultado, surgieron noticias [32] sobre una posible campaña financiada por Rusia para irrumpir en sistemas norteamericanos y robar mails del partido demócrata, y también para distribuir comentarios difamatorios y noticias falsas sobre la candidata demócrata Hillary Clinton en diversos sitios. Respecto a estas notas periodísticas de índole política, sabemos que siempre puede haber parcialidad e intereses, es decir, que una nota así se debe tomar con cuidado, ya que puede no ser imparcial. Sin embargo, la probabilidad de que se haya podido influenciar en la elección de un país desde otro haciendo uso de *spamming* de opiniones, refleja el riesgo que conlleva esta actividad, porque ya no estamos hablando de beneficios económicos para una empresa, sino del poder incidir en la llegada de una fórmula electoral al gobierno de un país.

Volviendo al problema de detección de spam de opiniones, reconocemos ahora su importancia. Está claro que, sin filtrar opiniones falsas, se puede llegar a un punto en el futuro en que la mayoría de las opiniones en la Web sean *spam*, y por lo tanto se pierda la confianza en los sitios cuyo valor principal es generado por sus usuarios.

Según algunos autores, el enfoque que se debería tomar es clasificar los comentarios analizados en dos categorías: las que son *spam* y las que no son *spam.* Para realizar el debido análisis de clasificación, se debe tener en cuenta que existen tres distintos tipos de opiniones o reseñas. Jindal y Liu [33][34] reconocen tres tipos de reseñas *spam,* y son las siguientes:

- **Tipo 1 u opiniones engañosas**: son aquellas que buscan engañar a los lectores y a los sistemas de minería de opinión, a través de opiniones falsas, es decir, que la persona u objeto no merece dicha valoración. Los motivos detrás de las opiniones engañosas son dos: por un lado, las opiniones positivas falsas buscan promover los objetos, para convencer a los lectores de adquirirlo, y por otro lado, tenemos las falsas opiniones negativas, que buscan desprestigiar a la entidad en cuestión.

- **Tipo 2 u opiniones sobre marcas**: el tipo 2 abarca los comentarios que no se hacen sobre un objeto o producto específico, sino que van dedicados a las empresas o marcas

que manufacturan o venden dichos productos. También puede aplicar a servicios. Generalmente son consideradas como *spam* por su poca utilidad, dado a que son muy abarcativas y no pueden usarse para identificar la valoración de algo concreto. Además, estas opiniones suelen ser parciales.

- **Tipo 3 o no-opiniones**: este tipo de reseñas se caracterizan por aparentar tener una opinión, pero en realidad carecen de ella.

Normalmente, identificar los tipos 2 y 3 puede resolverse con algoritmos de Machine Learning debidamente entrenados. Por lo que el desafío es obtener un conjunto de datos o características que sirvan como modelo para que aprendan los algoritmos mencionados. Siguiendo con lo enunciado por Jindal y Liu [33][34], vemos que los investigadores reconocen tres grupos de características que pueden ser útiles para aplicar a Machine Learning.

- **Características de reseñas**: Son atributos relacionados al contenido de la reseña o comentario. Algunos ejemplos son la cantidad de palabras de opinión, que fueron detalladas en el capítulo de Clasificación de Sentimientos, la cantidad de nombres de producto o marca que son empleados, y la longitud de la misma.

- **Características de producto**: como su nombre indica, son las cualidades del producto objeto de la reseña. Algunos ejemplos son el precio y la valoración promedio.

- **Características de autor**: son características y atributos de la persona que realizó la reseña. Algunos ejemplos son la cantidad total de reseñas que el autor tiene.

Es difícil identificar las opiniones de tipo 1, ya que estas también pueden tener las características mencionadas anteriormente. Es decir, una opinión falsa comparte las mismas características que una verdadera.
Existen opiniones falsas más dañinas que otras. Los autores coinciden en concentrarse en estudiar las opiniones que pueden generar grandes daños. Un detalle que permite identificar a una revisión falsa es que estas suelen desviarse de la media. Esta desviación es una condición necesaria para buscar esparcir opiniones inmerecidas. Entonces, por ejemplo, si para un producto la mayoría de opiniones son buenas y cada tanto encontramos una opinión negativa,

se la debería tener en cuenta para analizar detalladamente, ya que podría ser falsa, o simplemente una opinión verdadera que puede salirse de la media.

Una vez detectados estos comentarios que se salen de la media, se debe determinar si estos son verdaderos o reseñas *spam*. Usar algoritmos de *Machine Learning* no es del todo eficiente, al ser es muy difícil entrenarlos, porque los comentarios generalmente serán elaborados a mano.
En su investigación, Jindal y Liu observaron una gran cantidad de reseñas duplicadas, o muy similares entre sí. Esto les resultó llamativo, y vieron que con frecuencia se relacionaba con que esas reseñas eran *spam*. De manera que, para detectar *spamming de opiniones*, se debería entonces detectar las reseñas cuyo valor de opinión se sale de la media y además se tiene duplicados. Se debe recordar que estas cualidades son un indicio de que la reseña es spam, aunque no lo confirma de manera absoluta. Por otro lado, muchas opiniones falsas son armadas manualmente una por una, de manera que no sean duplicados exactos entre sí, por lo que resulta muy difícil detectarlos de manera automatizada. Los resultados que se pudieran obtener teniendo en cuenta sólo estas características serían bastante pobres. Es por esto que queda claro que la detección de *spamming* de opinión todavía está en su infancia, y que se debe continuar con su desarrollo.

2.4 Herramientas de análisis actuales

En la actualidad, hay una buena cantidad de herramientas de análisis de sentimiento publicadas en la Web. Algunas pueden encontrarse en forma de librería o código fuente, para ser estudiadas o utilizadas en algún proyecto casero. Las implementaciones que se pueden hallar en este formato, en general no son lo suficientemente potentes; un ejemplo es **sentiment**, una librería para usar en **node.js,** que básicamente tiene un diccionario de palabras (la versión original usa un diccionario del idioma inglés, aunque nosotros usamos una versión de este software con un diccionario en español), en el cual para cada palabra, tiene un puntaje, que indica si la palabra es positiva, negativa o neutral. Para procesar el texto, **sentiment** usa ese diccionario para reconocer todas las palabras sueltas que pueda y sumar el peso, para definir la opinión expresada en la frase.

También existen herramientas de sentiment analysis disponibles en forma de servicios Cloud; la mayoría pagos. Recientemente, Google publicó *Cloud Natural Language API,* la cual incluye reconocimiento de entidades (identificar palabras como nombres, expresiones o ubicaciones, por ejemplo), análisis de sentimiento, y análisis de sintaxis. Algunos conocedores del tema coinciden en que lo que Google ofrece no es ninguna novedad, ya que hace años que otros proveedores, como IBM, Aylien, Lexalytics y Meaning Cloud, brindaban este servicio en forma de API.

Sin embargo, la novedad que trae la solución de Google, es su potencia en la detección de relaciones y atributos de entidades; además de buenos resultados en la interpretación semántica del texto. Sin embargo, algunos expertos resaltan que *Cloud Natural Language API* tiene las mismas imprecisiones que sus competidores, en cuanto a los resultados.
Uno de las desventajas de esta implementación, es que su uso está limitado a unos pocos idiomas, al día de la fecha, Inglés, Español y Japonés.

IBM es otra compañía que apuesta al análisis de texto con su propio servicio. Desde su sitio Web, IBM destaca que el 80% de los datos son no estructurados, incluyendo artículos, investigaciones, posteos en redes sociales y datos de sistemas empresariales. Se ha visto en capítulos anteriores, que para un sistema no es fácil obtener información de datos no estructurados. Es a partir de esa problemática, que IBM presenta a su servicio, llamado Watson, como una solución a esto.

Watson es un sistema informático *cognitivo* capaz de realizar análisis de lenguaje natural, recuperación de información, representación de conocimiento y machine learning. Algo para destacar, es que no se limita a aplicar estas técnicas sólo a texto, sino que también es capaz de realizar reconocimiento de imágenes, y procesamiento de voz. Es el núcleo del servicio *Watson Developer Cloud* que IBM vende a desarrolladores, científicos y gente de negocios, como una herramienta para procesar y extraer información de distintas fuentes de datos en la Red.

En 2011, IBM hizo una demostración de las capacidades de Watson, haciéndolo competir y ganar una partida del juego *Jeopardy*, frente a dos de sus campeones más reconocidos. Desde

aquel momento, IBM siguió mejorando y extendiendo Watson para venderlo como un servicio accesible en la nube, como *Cloud Natural Language API* de Google.

Dentro del contexto de este estudio, nos interesa desarrollar sobre dos herramientas de Watson Developer Cloud: el Analizador de Tono y el Clasificador de Lenguaje Natural.

El **Analizador de Tono** de Watson tiene por entrada cualquier texto, para el cual es capaz de interpretar y reconocer la emoción (enojo, miedo, tristeza, alegría), tendencia social (extroversión, agradabilidad, escrupulosidad) y estilo de lenguaje (confiado, analítico y tentativo). Algunos posibles usos son:

- Mejorar el tono de un discurso, presentación o email, para llegar más eficientemente a transmitir una idea y obtener un mayor impacto en el público destinatario.
- Podría ser usado por un bot encargado hacer atención a clientes. Pudiendo reconocer el tono de conversación del cliente, podría adaptarse para obtener mejores resultados. Por ejemplo, si detectara que el cliente está enojado, puede cambiar las palabras y el tono utilizado para buscar atenuar ese enojo y así satisfacer la necesidad del cliente.

El **Clasificador de Lenguaje Natural** (CLN) es muy útil, ya que tiene la capacidad de reconocer la intención de un texto, y de esta manera clasificarlo, según un entrenamiento previo que se le haya dado al servicio. Por ejemplo, CLN podría ser entrenado para identificar la orientación en opiniones extraídas de comentarios en redes sociales, artículos, etc., para luego clasificarlo entre positivo, negativo o neutral. Otro caso de ejemplo sería el de un cliente de email, que, usando CLN de Watson, podría interpretar las categorías en que entra cada mail: foros, compras, importante, trabajo, entre otros.

Otra herramienta que permite realizar minería de opiniones es **Amazon Web Services.**
AWS, Amazon Web Services, es una compañía subsidiaria de Amazon que ofrece un conjunto de servicios de computación cloud orientado a aplicaciones que requieran estos tipos de servicios bajo demanda. Estos servicios están geográficamente distribuidos, y en total suman más de 70 servidores distintos incluyendo temas tan variados como poder de cómputo, almacenamiento, redes, bases de datos, analytics, administración, móvil, internet of things y

machine learning entre otros. Su objetivo es poder brindar servicios especializados de baja latencia, altas prestaciones y bajo demanda a bajos costos.

AWS no provee servicios directos de opinion mining, en vez de eso, proveen servicios de machine learning con aprendizaje supervisado junto a distintos servicios de bases de datos de gran potencia según las particularidades de los datos que queramos analizar. Si bien en un principio nos parece una buena opción para el desarrollo de un motor de minería de opiniones, creemos que hay opciones más sencillas para montar un servicio de minería de opiniones que ésta.

Otra opción interesante al momento de usar una API de minería de opiniones es **Mashape**.
Mashape es un conjunto de cuatro herramientas. Ofrece una plataforma open-source junto con servicio cloud para administrar, monitorear y escalar API y microservicios.
Entre todas las APIs de minería de opiniones que pueden encontrarse en Mashape, **Twinword** fue la que más nos interesó. De todas formas, en Mashape encontramos más de 20 APIs que brindan este servicio.

Twinword funciona de una manera muy simple : provee una API REST a la cual se le puede consultar acerca de una palabra, y responde con puntaje subjetivo, objetivo y palabras claves. . Cabe mencionar que Twinword provee varias librerías para hacer análisis de lenguaje natural como ser asociación de palabras , taggeo para textos, clasificación de textos, recomendación de categorías, etc. siendo sentiments análisis uno de los tantos servicios que brinda. De todas formas nos estaremos refiriendo a la librería de SA cuando usemos la palabra Twinword.

Es una API de análisis de opiniones que brinda información rápida sobre el análisis de comentarios de los usuarios. La API se puede acceder mediante métodos HTTP POST o GET enviando la palabra clave a analizar. Respondió un JSON con el resultado del análisis.
Su fácil uso la hace una API ideal para la minería de opiniones, desafortunadamente sólo provee mediciones para frases en inglés. No se ofrece documentación de como funciona internamente.

Nombre	Consultas	Precio en dólares
básico	500 / mes $0.003 PER EXTRA	0/mes
pro	125,000 / mes $0.001 PER EXTRA	19/mes
ultra	750,000 / mes $0.001 PER EXTRA	99/mes

Por otro lado también encontramos la API de Azure Machine Learning provista por Microsoft Azure. Esta API para realizar minería de opinión utiliza el motor de Azure Machine Learning con algoritmos de máquinas de vector soporte. El servicio clasifica opiniones en tres niveles, positivo, neutral y negativo. También provee puntaje de confianza por si se desea más adelante ajustar la polaridad. De 0 a 0.45 es negativo, de 0.45 a 0.6 neutro y desde 0.6 a 1 es positivo.

Nuestra experiencia al usar el servicio no fue del todo alentadora, la documentación encontrada fue muy confusa, desactualizada al punto de encontrar material que ya no tenía soporte. De esta forma decidimos no trabajar con el motor de ML de Microsoft.

2.5 Visión a futuro

Dado el crecimiento sostenido en la generación de datos en contenidos online, es de esperar que se requerirán métodos de minería de texto y análisis de sentimiento más potentes que los disponibles actualmente.

Según Bing Liu [6], los problemas que tenemos en la actualidad para llegar a una solución de minería de opiniones efectiva, residen en la forma en que se está encarando; haciendo una fuerte apuesta en algoritmos de Machine Learning. Si bien estos algoritmos han logrado avances y una precisión aceptable, sería sano intentar el uso de otros conceptos o tecnologías, no como reemplazo, sino como refuerzo que sirva para contraste en un intento de integración

de resultados.

De la naturaleza interdisciplinaria de la minería de textos, más el reconocimiento que está recibiendo, están surgiendo nuevos algoritmos y métodos. Los autores reconocen las siguientes posibles direcciones futuras del análisis de texto:

- **Métodos escalables y robustos para entender al lenguaje natural:** la mayoría de los enfoques actuales tratan al texto como una colección de palabras, la cual es una representación simple. Lo ideal sería contar con métodos que permitan extraer información de texto irrestricto. Los métodos disponibles hoy en día funcionan bien si tienen suficiente supervisión y entrenamiento, lo cual restringe sus usos.

- **Adaptación de dominio y transferencia de aprendizaje:** el correcto funcionamiento los métodos actuales de minería de textos depende de la supervisión y entrenamiento de los mismos. Estos son limitantes, ya que preparar todos los datos de entrenamiento implica un gran esfuerzo, y además los algoritmos sólo quedan preparados para interpretar texto con algunas limitaciones. Es decir que, a pesar del entrenamiento, no están preparados para analizar texto irrestricto. Los conceptos de adaptación de dominio y transferencia de aprendizaje evitarían estos limitantes, dado que el entrenamiento podría disponibilizarse dentro de dominios que puedan usarse en distintos contextos.

- **Análisis contextual de datos de texto:** el texto y los datos contenidos en él dependen del contexto, es decir que depende del autor y de la época y de la red de conexiones que tenga con otros textos. Por eso es muy importante de cara al futuro incorporar a los métodos actuales la capacidad de entender el contexto.

- **Minería de texto paralela:** en la actualidad, es mucha la cantidad de texto a procesar por los algoritmos de text mining, y se prevé que va a aumentar en el futuro. Es por esto que, para ser viables, los algoritmos de text mining probablemente tengan que correr en forma de tareas paralelas un cluster de computadoras. La explotación de esta característica se ve favorecida con las capacidades actuales y futuras de cloud computing.

2.6 Conclusiones

El análisis de texto es un campo muy activo en la actualidad, que viene con un desarrollo sostenido en los últimos años. Si bien nace de las ciencias de la computación y la estadística, cruzó fronteras y alcanzó a otras disciplinas o áreas, como el marketing y la administración, dado sus usos prácticos para reconocer la orientación y respuesta de un público o clientes respecto a un servicio o producto.

Estos usos prácticos, sin embargo, no obtienen resultados completamente satisfactorios. Se vió en este trabajo que esto se debe a distintas complejidades que se presentan en el contenido textual que debe analizarse. El mayor desafío, en este caso, es el del procesamiento natural del lenguaje; se estima que todavía no tenemos un entendimiento suficiente del problema. Por lo tanto, las soluciones disponibles hoy en día cubren parcialmente lo que se espera de ellas.

En este capítulo, cubrimos los aspectos generales del análisis de texto, haciendo especialmente foco en las redes sociales y otras fuentes de datos presentes en la Red, de las cuales se puede extraer información con AM. Hicimos una presentación sobre minería de opiniones, describiendo qué es una opinión, qué elementos la componen que son de interés para su análisis. Describimos brevemente distintos métodos y técnicas para realizar minería de opiniones, resaltando fortalezas y desventajas. También nombramos algunas herramientas de código abierto disponibles como librerías para que puedan ser usadas en desarrollo de software, como también listamos servicios implementados por grandes compañías, disponibles en la Nube para ser consultadas. Finalmente, recolectamos los posibles caminos que los expertos creen que puede tomar este área.

Coincidimos con los autores y expertos en que todavía falta maduración en las técnicas utilizadas, pero creemos que con el aporte de las distintas disciplinas interesadas, más el siempre vertiginoso crecimiento de las herramientas de software disponibles, gradualmente se conocerá mejor la naturaleza del análisis de texto, y se van a llegar a implementar soluciones superadoras que obtengan resultados más cercanos a los que promete esta disciplina.

3. Desarrollo propuesto

3.1 Descripción y objetivos

En los capítulos anteriores, describimos que de las redes sociales pueden extraerse un volumen considerable de comentarios de usuarios, y que, usando una herramienta automatizada para interpretar las opiniones generales en esos comentarios, se podría contar con información valiosa para realizar estadísticas y estudios de mercado a un bajo costo. A su vez, repasamos las dificultades y desafíos que representa la interpretación automatizada de texto en lenguaje natural, y listamos algunas tecnologías y conceptos que son parte del estado del arte de estas técnicas. Algunas de las grandes compañías de software, como Google, Microsoft e IBM las están utilizando, y apuestan a ellas sabiendo que aún tienen mucho potencial por descubrir.

Uno de los objetivos de este trabajo de tesina es el de desarrollar una herramienta Web con foco en la temática de minería de opiniones. En concreto deberá cumplir con los siguientes requerimientos:

1. Se debe conectar con una red social para obtener de ella comentarios sobre un mismo tema durante un tiempo determinado. La extracción de comentarios se hará en “lotes” iguales de tiempo.
2. Se debe integrar con servicios externos de análisis de sentimientos, de los que extraerá el valor de opinión de cada comentario.
3. Por cada comentario obtenido, además, se extraerán los siguientes datos:
    a. Género del usuario que comenta.
    b. Dispositivo usado para emitir el comentario.
4. Almacenará los resultados en una base de datos.
5. Deberá servir como herramienta para comparar los resultados obtenidos. Para esto, incluirá gráficos que permitan evaluar la evolución de los resultados a obtenidos. Contará con:
    a. Un gráfico de líneas para mostrar cómo progresa la proporción de resultados positivos, negativos y neutrales para cada herramienta, con el transcurso del

tiempo. Contará con filtros para seleccionar los resultados según género (femenino, masculino, todos), y los resultados hechos desde un mismo dispositivo (navegador web, Android, iPhone, todos).

b. Un diagrama de Venn para ver la cantidad de comentarios donde coincidieron las herramientas respecto al valor de opinión inferido.
c. Otro gráfico de líneas para mostrar la evolución del valor de opinión de los comentarios de un usuario durante un tiempo determinado.
d. Un gráfico de cajas y bigotes, cuyo uso será el de mostrar los cuartiles de la clasificación de las librerías.

Al sistema que pensamos y desarrollamos le dimos el nombre de **Opinator.** A partir de ahora, lo llamaremos por ese nombre en distintos pasajes del trabajo.

Decidimos extraer comentarios de la red social *Twitter*, ya que nos pareció que contaba con una API interesante para lo que queríamos lograr, además de estar bien documentada. En particular, vamos a usar un llamado a la API, a la cual se le debe enviar un *String* con un término cualquiera; a partir de ese momento se mantiene abierta una conexión, en donde la API retorna en tiempo real un *stream* de tweets que contienen ese término. Otro factor que nos interesó de Twitter es que solo admite comentarios de 140 caracteres, lo cual nos da un marco acotado y un mejor control sobre el texto que vamos a analizar.

El sistema se conecta con Twitter para buscar lotes de una cantidad limitada de comentarios de usuarios aleatorios. La frecuencia con la que Opinator se conecta para obtener esos lotes, la duración de la conexión con la API de Twitter, y el tamaño del lote son configurables.

Todos estos comentarios obtenidos tratarán sobre un mismo tema provisto por el usuario. En la medida que reciba los tweets, el sistema los enviará a analizar por distintos servicios de análisis de sentimientos integradas. Cada uno de estos servicios usan estrategias distintas para su fin. Una vez obtenidos los resultados del análisis realizado, los guardaremos en una base de datos, para luego mostrarlos con gráficos que ayuden al usuario a hacer comparaciones, pudiendo aplicar filtros para hacer más específicos a los resultados. En las secciones que vienen a continuación, pasamos a detallar la arquitectura y el funcionamiento del sistema. También mencionaremos los servicios utilizados.

3.2 Arquitectura y tecnologías utilizadas

En este capítulo explicaremos la arquitectura y tecnologías utilizadas del desarrollo propuesto. El mismo es una aplicación Web, cuya organización en capas puede visualizarse en el diagrama de la figura 8:

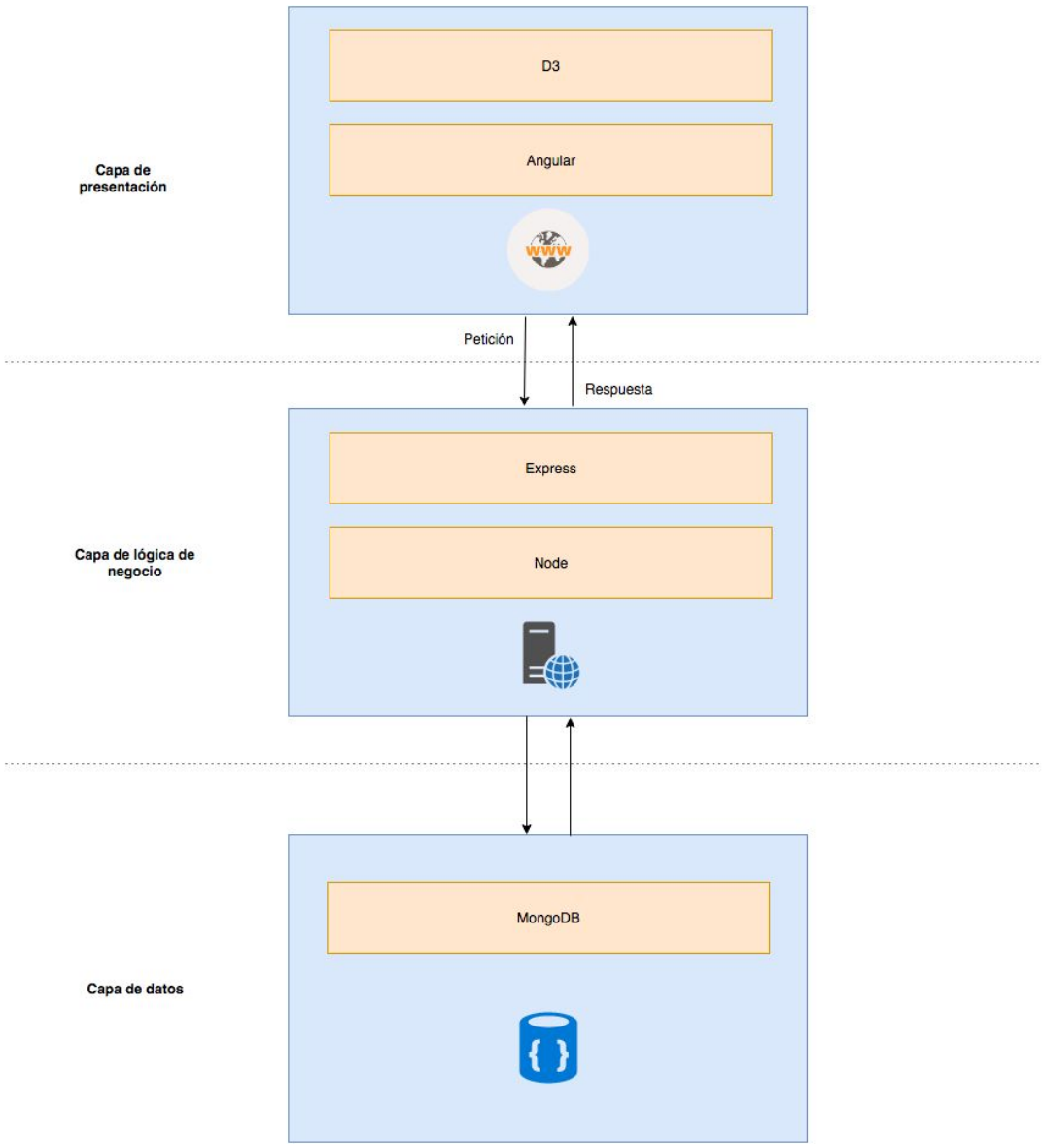

Figura 8. Diagrama de arquitectura de capas de Opinator.

Opinator cuenta con tres capas: de **presentación**, de **lógica de negocio**, y de **datos**. A continuación explicamos cada una, su rol y las tecnologías que las componen:

3.2.1 Capa de presentación

En la **capa de presentación** (frontend) usamos **Angular**, un framework Javascript *open source* mantenido por Google y por colaboradores externos. Brinda facilidades para programar interfaces web tipo Single Page Applications (SPA), que se caracterizan por crear una sola página HTML, manejar el contenido de la misma para que cambie de manera dinámica, sin tener que recargar, ni cargar páginas nuevas, dando una sensación de fluidez en la navegación de la misma.

La actualización de contenido se realiza a través del uso de HTML5 y peticiones AJAX para comunicarse con distintas APIs que proveen los datos necesarios. Este tipo de aplicaciones se ejecutan completamente en el navegador, y tienen cierta autonomía del servidor, en el sentido de que pueden residir en servidores diferentes, y cuentan con rutas propias (rutas "de frontend"), es decir que las rutas del cliente no significan peticiones al servidor, sino que definen la navegabilidad dentro del mismo. Para mostrar las distintas páginas, sin embargo, se requieren de datos brindados desde backend, para lo que Angular brinda herramientas para hacer peticiones AJAX programáticamente. Las respuestas a esas peticiones deben venir serializadas en formato JSON.

Otra característica de Angular es que provee una arquitectura model–view–controller (MVC) para el lado del cliente. Tradicionalmente, el uso esta arquitectura se reservaba para servidores y aplicaciones de *backend*. Agregando MVC en el frontend, se puede repartir mucha de la lógica que antes sólo residía en el backend, en ambos componentes, equiparando responsabilidades entre ambos.

Angular fué diseñado con la idea de que la programación declarativa (en HTML) es la mejor alternativa para implementar interfaces gráficas, mientras que el paradigma imperativo (en Javascript) tiene un mejor uso en la definición de lógica de negocio. Años atrás, la línea divisoria entre el desarrollo de interfaces y de lógica de negocio era difusa; se usaba Javascript no sólo para implementar lógica, sino también para manejar aspectos de GUI tales como aplicar efectos, y mostrar u ocultar elementos, manipulando el DOM programática y explícitamente. El código javascript que manejaba esas cuestiones de UI podía estar embebido dentro del mismo HTML, como también en archivos `.js` separados de manera arbitraria a

criterio del desarrollador, o peor aún, mezclado en mismos archivos junto a lógica de negocio. Esto generalmente redundaba en proyectos en donde no sólo se mezclaban lenguajes, paradigmas y responsabilidades; sino que también eran relativamente difíciles de mantener.

La solución que Angular provee para mantener adecuadamente separados a ambos mundos, es la de extender al HTML tradicional, agregando nuevas etiquetas y atributos para presentar contenido de manera dinámica, a través de un mecanismo de sincronización automática entre la vista y los datos subyacentes, conocido como *enlace de datos de dos vías* (two-way data binding en Inglés) (ver figura 9). El nombre del mecanismo viene por su naturaleza bidireccional: en ella, un modelo de datos (un objeto javascript), se enlaza o relaciona con distintos componentes de la vista, de manera que cada cambio que se produce sobre el objeto, modifica lo que se muestra en la vista, de manera automática. A su vez, algunas propiedades de ese objeto pueden enlazarse a inputs HTML, de manera que, cada cambio producido por el usuario en esos inputs, impacta también modificando las propiedades del objeto.

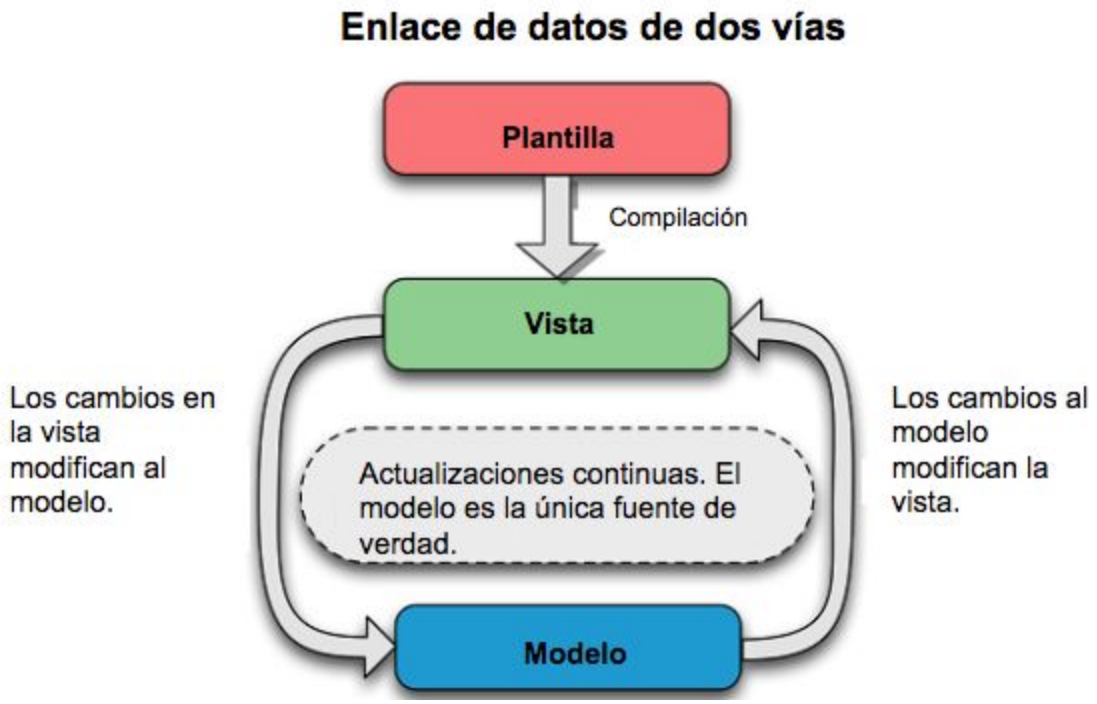

Figura 9. Diagrama representativo del mecanismo de *enlace de datos de dos vías* de Angular.

Mediante el mecanismo de *enlace de datos de dos vías* (ver Figura 9), Angular desenfatiza el

manejo explícito del DOM, permitiendo al usuario programador usar javascript sólo para el manejo de objetos de datos que conforman al modelo, definir servicios, recursos, etc.

Angular funciona interpretando archivos HTML extendidos, es decir, que además de contar con los nodos *default* del lenguaje HTML, tiene algunos definidos por el framework. Estos nodos son conocidos como directivas de Angular. Más adelante daremos más detalle sobre las mismas.

De la interpretación de directivas, Angular permite definir la estructura y comportamiento de la vista de manera declarativa: se puede relacionar un modelo de objetos con partes de la vista, mostrar u ocultar secciones cuando se dan determinadas condiciones, iterar modelos de tipo colección, entre otros.

Volviéndonos a centrar en Opinator, dijimos que tendría una interfaz gráfica tipo *dashboard* o tablero, con gráficos y tablas para poder comparar y comprender con facilidad los resultados obtenidos. Para implementar los gráficos, decidimos incluir una librería javascript llamada *D3* (*Data-Driven Documents*), cuyo fuerte es que usa un modelo JSON de donde extraer los datos para realizar gráficos, y que si se modifica programáticamente ese modelo, el gráfico responde dinámicamente, cambiando su forma. Otra característica que nos llevó a elegir D3, es que permite un alto grado de libertad para definir gráficos programáticamente en JavaScript, lenguaje que puede observarse como predomina en todo el proyecto.

Un factor importante en esta selección fue la fácil integración entre Angular y D3, por medio de *directivas* de Angular. Como mencionamos anteriormente, una *directiva* es un nodo o atributo HTML que extiende a dicho lenguaje, y ayuda a definir comportamiento de manera declarativa, sin la necesidad de incluir código Javascript que maneje el DOM de manera explícita. Angular provee sus propias directivas *built-in*, pero también provee al usuario la opción de implementar directivas propias. Al implementar una directiva, lo que se está haciendo es asociar funcionamiento o comportamiento a un nombre, que luego podrá ser referenciado desde un documento HTML, para declarar que esa parte del DOM tendrá ese comportamiento. El usuario de una directiva puede usarlas para enriquecer vistas, abstrayéndose de cómo está desarrollada.

A continuación mostramos un ejemplo simple de directiva hecha con Angular para Opinator:

```
angular
    .module('app.results')
    .directive('deviceIcon', function() {
        var icons = {
            'web': $('<i class="fa fa-globe"></i>'),
            'iphone': $('<i class="fa fa-apple"></i>'),
            'android': $('<i class="fa fa-android"></i>'),
            'windows': $('<i class="fa fa-windows"></i>')
        };
        return {
            restrict: 'E',
            scope: {
                val: '='
            },
            link: function(scope, element, attrs) {
                var icon = icons[scope.val];
                $(element.parent()[0])
                    .append(icon)
                    .append(' ' + scope.val);
            }
        };
    });
```

En la misma, declaramos el nombre que llevará la directiva, que es `'deviceIcon'`. Este nombre es el que usaremos después para llamar a la directiva desde el HTML, es decir escribir `<device-icon val="'web'"></device-icon>` en donde queramos que se ejecute el comportamiento, que es aquel definido en la función de callback declarada en la línea `link: function(scope, element, attrs) {...}`.

En este caso, la función obtiene un identificador desde la propiedad del objeto `scope.val`, busca un ícono asociado a ese identificador, y de existir, lo inserta en el mismo lugar del DOM en donde la directiva fué llamada.

Otras directivas que creamos contienen el código necesario para renderizar gráficos y *data visualization*. Uno de ellos se llama `barChart`, e implementa un gráfico de barras genérico. La ventaja que obtenemos con haber metido todo ese código en una directiva, es que, cada vez que queramos mostrar un gráfico de barras, podremos incluir un nodo `<bar-chart val="{values}"></bar-chart>` en nuestro HTML (donde `values` sea un arreglo que contenga datos para graficar) y obtendremos el gráfico. De esa manera entonces, podemos abstraernos de la implementación, reusar código y mantener el HTML "limpio" de Javascript y declarativo. A continuación mostramos cómo quedó parte del HTML de Opinator, con directivas:

```
<div class="row">
      <div class="col-md-3 col-md-offset-1">
            <h3>Géneros</h3>
            <gender-pie-chart val="genders"></gender-pie-chart>
      </div>
      <div id="devices-bar-chart" class="col-md-8">
            <h3>Dispositivos usados</h3>
            <bar-chart val="topDevices"></bar-chart>
      </div>
</div>
```

De esta sección se puede deducir con facilidad que los tags marcados en naranja y verde renderizan un *pie chart* o gráfico de torta, y un *bar chart* o gráfico de barras, respectivamente. En ninguna parte del documento hay secciones de código Javascript, y no es necesario entender la implementación de ambas directivas, para saber con facilidad qué es lo que hacen.

3.2.2 Capa de lógica de negocio

Node

La **capa de lógica de negocio** (backend) corre en un servidor Web, y está desarrollada en **Node**, el cual es "un entorno de ejecución para Javascript que se ejecuta de manera asincrónica, pensado para desarrollar aplicaciones de red eficientes y escalables" [36]. La eficiencia y escalabilidad de Node se relaciona con lo simple que es para manejar concurrencia, gracias al uso de *callbacks* y un proceso que los autores llaman *Bucle de eventos* (ver Figura 10). Este manejo contrasta con el tradicional en tecnologías de más años como Java, que consiste en usar múltiples *threads* del Sistema Operativo. A continuación explicamos los conceptos de *funciones de retorno (callback* en Inglés*),* y del mecanismo de *Bucle de eventos.*

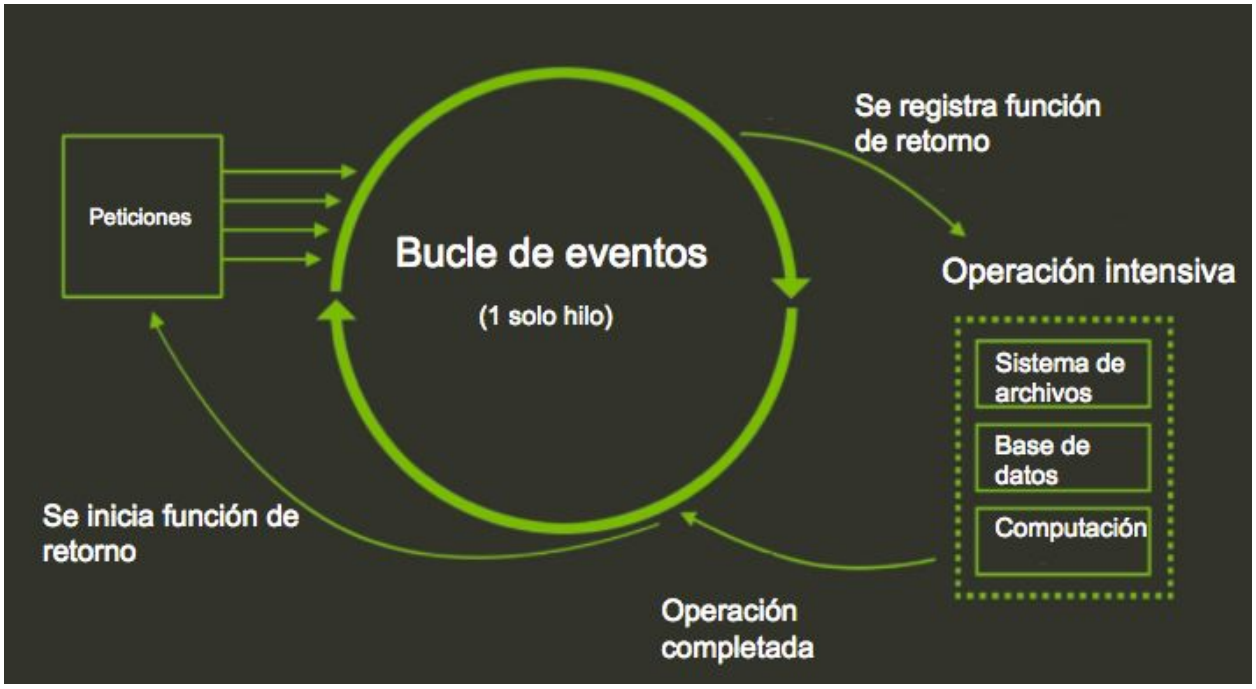

Figura 10. Diagrama de funcionamiento del mecanismo de *bucle de eventos* de Node.

En el modelo de Node, una operación que puede bloquear al único hilo que maneja el entorno, es delegada para que sea ejecutada fuera de ese hilo. Un ejemplo de estas operaciones "bloqueantes" son: lectura/escritura de archivos, conexiones a bases de datos, y uso de red

para comunicaciones. La delegación de estas tareas pueden realizarse al kernel del SO, o a otro componente, para liberar el *thread* de Node. Para cada una de estas operaciones bloqueantes, el *bucle de eventos* de Node se encarga de registrar una *función de retorno*, que puede ser provisto por el usuario. Un *callback* es un bloque de código ejecutable, que es pasado como argumento a otro bloque. De un *callback* se espera que sea ejecutado en un momento determinado. Utilizando un patrón de diseño *Observer* [5], el callback queda asociado a la operación, de manera que se ejecute eventualmente ante el evento de finalización de dicha operación. En pocas palabras, la ejecución de esa función de *callback* sirve para declarar qué acciones se deben tomar una vez que termina una operación asincrónica, sin importar cuándo termina; eventualmente ocurrirá.

De esta manera, se pueden realizar tareas de entrada/salida, sin bloquear el funcionamiento de Node, haciéndolo idóneo para el desarrollo de servidores HTTP, por ejemplo, en donde es necesaria una alta disponibilidad para atender múltiples requests a la vez.

Express

Express.js se describe a sí mismo como un "un framework para el desarrollo de aplicaciones web mínimo y flexible, para Node.js”[37]. Ayuda a agilizar el desarrollo de aplicaciones Web. Si bien un desarrollo así puede realizarse solamente con Node, Express es de utilidad, ya que internamente usa el módulo HTTP de Node, pero abstrae a sus usuarios de los detalles técnicos, brindándoles características realmente muy fáciles de comprender y utilizar para tener un servidor Web funciona de manera rápida. Entre las características más importantes que aporta Express, tenemos la posibilidad de usar *middleware* [38]. Una aplicación Express es, esencialmente, una serie de llamadas a funciones *middleware.* Estas funciones tienen acceso a los *request* y *response* HTTP, como también a la siguiente función en la cadena de *middlewares*, y cada una representa una tarea a realizar por cada interacción entre el cliente y el servidor. Un ejemplo de ellas es la de autenticar al usuario que realiza un request, con una función middleware que acceda al *request* HTTP y valide claves en él.

Las funciones *middleware* sirven entonces para realizar distintas cosas, como definir una función *handler*, que se ejecute ante cada petición HTTP que ocurre en el servidor; sirve también realizar cambios en los request/response, para agregar *headers*, realizar validaciones, como también *logging*.

A continuación mostramos un ejemplo de *middleware* en Express, para hacer *logging* del tiempo en que fué recibido un request en el servidor. Una vez realizado el *log*, se llama a una función `next()`, que se encarga de obtener la próxima función *middleware* en la cadena, para seguir atendiendo ese request.

```
var app = express();
app.use(function (req, res, next) {
  console.log('Time:', Date.now());
  next();
});
```

Puede apreciarse la sencillez y transparencia del framework; es muy poca o nula la "preparación" o *set up* para usar sus características.

Otra característica notable de Express es la posibilidad de definir rutas. Definir una ruta implica determinar qué acciones tomar ante una petición a un *endpoint* específico(una URI), para un método HTTP específico, en una instancia de servidor Web. Usando una cadena de *middlewares*, una ruta puede tener más de una función *handler* que se encargue de "atajarla" y definir la respuesta.

A continuación mostramos dos ejemplos muy sencillos de definición de rutas con Express. Uno atiende un método HTTP GET al *endpoint* "raíz" o /, el segundo define como responder a un método HTTP POST en el endpoint */users*. El tercer ejemplo es sacado de Opinator, que dice como se debe atender un método HTTP GET en el *endpoint /sentiments*.

```
//EJEMPLO 1
app.get('/', function (req, res) {
  res.send('Hello World!');
});


//EJEMPLO 2
```

```
app.post('/users', function (req, res) {
  res.send('Got a POST request at /users endpoint');
});


//EJEMPLO 3
app.get('/sentiments', QueryController.sentiments);
```

Nótese nuevamente la poca o nula configuración necesaria, para tener un servidor funcional. Cabe destacar también que en el tercer ejemplo, `QueryController.sentiments` es un método de clase para la clase QueryController, y que dicho método actúa como *callback*: es ejecutado sólo cuando el servidor recibe una petición a */sentiments*; la definición del método es `QueryController.sentiments = function(req, res) { ... }`; recibe los objetos request (`req`) y response (`res`), para atender esa petición.

3.2.3 Capa de datos

Motor de Base de Datos

Para la capa de datos decidimos usar MongoDB, una base de datos noSQL que graba documentos en formato JSON. Esta característica fue determinante en la elección de la tecnología de base de datos, ya que la respuestas que debe realizar el servidor al cliente son también en JSON, por lo que no hay que realizar ningún mapeo de la información en el grabado y obtención de datos como sí ocurriría con una base de datos SQL o Relacional.

Datos interpretados de la API

Como se mencionó anteriormente, consumimos tweets en formato JSON desde una de las APIs de Twitter. De ese tweet usamos algunos nodos. A continuación mencionamos cuáles son, para que los interpretamos, y mostramos ejemplos del formato que tienen en el JSON. Al final detallamos nodos que no usamos, pero que consideramos que tienen valor para enriquecer los resultados en un desarrollo futuro.

Campo “text”

Contiene el texto del tweet escrito por el usuario. Es el escrito que mandamos a analizar por Watson, Cognitive Services y Sentiment para obtener la opinión expresada en el mismo.

```
{
    "text": "RT @PostGradProblem: In preparation for the NFL lockout, I will be spending twice as much time analyzing my fantasy baseball team.",
    ...
}
```

Campo “source”

Tiene información sobre el dispositivo o fuente desde donde se emitió el tweet, como por ejemplo, un iPhone, un Android, o desde la Web. Dado que la información viene como un string con formato HTML, aplicamos una lógica para extraer el nombre del dispositivo.

De esta manera, por cada tweet analizado, hacemos un conteo de dispositivos y lo grabamos en nuestra BD, para mostrar luego en un gráfico de barras en el frontend.

```
{
    ...,
    "source": "<a href=\"http://twitter.com/\" rel=\"nofollow\">Twitter for iPhone</a>",
    ...
}
```

Nodo “user”

Representa al usuario que publicó el tweet. Internamente tiene muchos subnodos y campos. A continuación detallamos aquellos que usamos:

Campo “user.lang”

Indica el lenguaje del usuario. Lo usamos para filtrar los tweets que recibimos desde la API, para quedarnos sólo con los que están escritos en Español.

Campo “user.name”

Representa el nombre del usuario. No debe confundirse con el campo “**screen_name**”, que representa el nombre de la cuenta. El campo “**name**” lo procesamos con una librería JS de nombre **gender**, que mediante un diccionario de nombres, intenta inferir el género. El resultado

de ese proceso puede ser *hombre, mujer,* o *desconocido*, en caso de no haber podido reconocer un género, y lo almacenamos en uno de los esquemas de base de datos que describiremos próximamente.

Campo “user.screen_name”

Es el nombre de usuario con el que puede ser identificado unívocamente. Siguiendo el ejemplo de la figura anterior, el **screen_name** del Twitter de la facultad de Informática de la UNLP, es **InformaticaUNLP.** A cada tweet que obtenemos de la API de Twitter, lo mandamos a analizar por diferentes servicios y librerías. Una vez obtenidos los resultados, los grabamos agrupados junto al texto del tweet e información del usuario, en la base de datos. Uno de los datos de usuario almacenados es el **screen_name,** para representar en la capa de presentación al autor de un tweet.

Esquemas de datos

A continuación presentamos el esquema que utilizamos en nuestra base de datos. Vale recordar que usamos MongoDB, la cual es una base de datos no relacional. El grabado de la información se realiza en documentos con formato JSON. Para cada esquema mostraremos un ejemplo, y explicaremos decisiones de diseño.

Esquema tweets

El esquema **tweets** contiene los datos que extraemos de los usuarios autores de los tweets obtenidos por la API de Twitter. A continuación mostramos un ejemplo de un documento de este esquema:

```
{
    "_id" : ObjectId("590e19fbad94c538e947533e"),
     "text" : "Bernasconi: “El armado de listas debe representar la esencia y el espíritu de Cambiemos ” https://t.co/6G6vLT9fz6 https://t.co/3iYyZDiY8h",
    "date" : ISODate("2017-05-06T18:46:18.380Z"),
    "user" : {
        "name" : "Primicias Quilmes",
```

```
        "screen_name" : "QuilmesPrimicia",
           "description" : "Portal de Noticias de distintos medios sobre la
Ciudad de Quilmes.",
        "lang" : "es",
        "location" : "Quilmes, Argentina",
        "gender" : "unknown",
        "device" : null,
                                                    "profile_image_url"      :
"http://pbs.twimg.com/profile_images/768528963765403650/NJgtTLvY_normal.jpg"
    },
    "sentiments" : {
        "sentimentJS" : "neutral",
        "watson" : "neutral",
        "msCognitive" : "neutral"
    }
}
```


Descartando el campo "_id", que es generado por MongoDB para identificar documentos, el resto es armado por Opinator. En la sección **Datos interpretados de la API** se describió cómo se obtienen los campos de usuario y del tweet.

Cada documento del esquema **tweets** representa a un tweet, y va acompañado con datos del usuario que lo creó, y los resultados de las herramientas de análisis de opinión. Cada documento del esquema se compone por los siguientes campos y objetos:

- Campo **text**: contiene el comentario que hizo el usuario para un tweet.
- Campo **date:** contiene la fecha en que se emitió el tweet.
- Objeto **user:** representa al usuario que escribió el tweet. Está formado por los siguientes campos internos:
    - Campo **name:** Es el nombre del usuario.
    - Campo **screen_name:** representa la cuenta del usuario de manera unívoca.
    - Campo **description:** Descripción que el usuario pone sobre su perfil.
    - Campo **lang** : Idioma del usuario.
    - Campo **location**: Es la ubicación que el usuario declara.

    - Campo **gender**: Género del usuario. Este dato no es provisto por Twitter, sino que lo creamos a partir de una herramienta de inferencia de sexo a partir de nombres de personas.
    - Campo **device**: Nombre del dispositivo con el que el usuario creó el tweet.
    - Campo **profile_image_url**: URL de la imágen de perfil del usuario.
- Objeto **sentiments:** representa los resultados obtenidos por las distintas herramientas de análisis de opinión para el tweet. Cada uno de estos campos puede tener un valor entre los siguientes: *positive, negative,* o *neutral*. Está formado por los siguientes campos internos:
    - Campo **watson:** Resultado obtenido con el servicio Watson de IBM.
    - Campo **msCognitive:** Resultado obtenido con el servicio Cognitive Services de Microsoft.
    - Campo **sentimentJS:** Resultado obtenido con la librería Sentiment.

3.3 Servicios y librerías externas

En esta sección vamos a describir los servicios externos y librerías usados para el análisis del texto, y el servicio de Twitter desde el cual consumimos tweets. Cada uno de ellos tiene diferentes formas de uso, como también diferentes respuestas: algunas, por ejemplo, retornan un puntaje que puede variar entre 0 y 1; otras responden con un string como *positive, negative* o *neutral*. Un ejemplo de diferencias respecto al llamado de cada uno, es que para el caso de una librería, alcanza con llamar a una función, que se encarga de analizar el texto; mientras que en el caso de un servicio externo, es necesario hacer una Petición HTTP a un servidor con *headers* específicos.

Ante este escenario, para unificar los llamados y las respuestas de cada librería o servicio, implementamos objetos que nosotros llamamos *estrategias*, inspirados en el patrón de diseño conocido como *Strategy* [5]. Cada objeto *estrategia* encapsula en un método la forma en que se utiliza el servicio, y el resultado, de manera en que sea igual para todas las estrategias usadas, independientemente de que se use internamente.

A continuación presentamos los servicios externos y librerías (a partir de ahora, estrategias) que usamos en Opinator:

Nombre	Tipo	Autor	Enfoque/Estrategia
Watson	REST API	IBM www.watson.ibm.com/	*NLP/Machine Learning*
Cognitive Services	REST API	Microsoft microsoft.com/cognitive-services	*NLP/Machine Learning*
sentiment-spanish	Librería	npmjs.com/~goalkeeper112 lfbu.112@gmail.com	*Diccionario de palabras*

Dos estrategias son servicios externos: **Watson** de IBM y **Cognitive Services** de Microsoft. Ambos exponen una API Web, y cuentan con estrategias de Machine Learning y NLP, por lo que tienen la capacidad de analizar el sentimiento interpretando las palabras en el texto y el contexto en el que son utilizadas. Estos servicios cuentan con otras capacidades, tales como extracción de entidades de un texto, y análisis de imágenes y sonido. Sin embargo, dado que esta tesina se concentra en el análisis de texto y sentimientos, vamos a concentrarnos en esa faceta de dichos servicios.

Una de las estrategias es una librería llamada **sentiment-spanish**, que intenta extraer el sentimiento en el texto en función de las palabras sueltas que encuentra en el mismo. Esta estrategia es muy simple y se diferencia de las otras dos, dado que no realiza un estudio del contexto de la opinión, ni observa si hay particularidades como negaciones y sarcasmo, entre otros factores, que puedan modificar el sentido de las palabras.

APIs de Twitter

En secciones previas mencionamos que elegimos Twitter como red social de donde sacar comentarios en tiempo real sobre un término, o conjunto de términos. Esta decisión fue tomada por dos razones. La primera es el límite que impone Twitter para que los tweets tengan hasta 140 caracteres, lo cual nos pareció favorable para todos los comentarios que mandamos a

procesar tuvieran medianamente la misma longitud.

La segunda razón detrás de esta decisión fué el conjunto de APIs que Twitter ofrece, y su documentación. En el análisis que hicimos inicialmente, vimos que estas APIs ofrecían una gran variedad de consultas posibles que nos eran de utilidad para nuestra implementación. Cabe destacar que la documentación también nos pareció muy completa. Nos ayudó a entender qué podíamos obtener con la API y cómo utilizarla.

Investigando más detalladamente, observamos dos variantes de las API de Twitter: una llamada *Public Search* y otra de tipo *Streaming*. Si bien hallamos valor en ambas, nos terminamos inclinando por usar la variante *Streaming*. A continuación nombramos las características y diferencias entre ellas, y qué vimos en la API *Streaming* para considerar que era la que mejor se adaptaba a nuestra idea.

Ambas API identifican aplicaciones y usuarios a través de un conjunto de claves que provee Twitter, y para las cuales hay que registrarse en su sitio. A través de esa clave, no solo identifican quiénes le están haciendo las diferentes API *calls*, sino que también controlan el acceso a las mismas.

API *Public Search*

En palabras de la página oficial de Twitter, su API de *Public Search* “permiten acceder de forma programática a leer y escribir datos” en dicha plataforma [35]. Algunas de las peticiones permitidas por la API permiten conseguir información de usuarios: por ejemplo, como obtener datos de su perfil, de sus seguidores, su geolocalización (si aceptó compartirla).

Otras peticiones permiten obtener una colección de cantidad fija de tweets relevantes, según una palabra que forme parte de la consulta. Esta petición nos resultaba muy útil para las necesidades que teníamos para implementar Opinator. Dentro de los parámetros de la consulta, además del término o palabras buscadas, podían solicitarse cosas como que todos los tweets sean de un lenguaje específico; o que se obtengan los tweets de un lugar geográfico específico. Esta consulta también cuenta con un parámetro que indica cuántos tweets se espera que retorne en ese llamado.

Básicamente, con esta consulta teníamos todo lo que necesitábamos para implementar la obtención de tweets, que luego enviaríamos a procesar. Sin embargo, vimos una limitante. Esta era que, para obtener tweets de manera contínua con esta API, necesitábamos hacer varios llamados de manera repetida durante un período de tiempo. El siguiente diagrama (figura 11) refleja el uso de esta API:

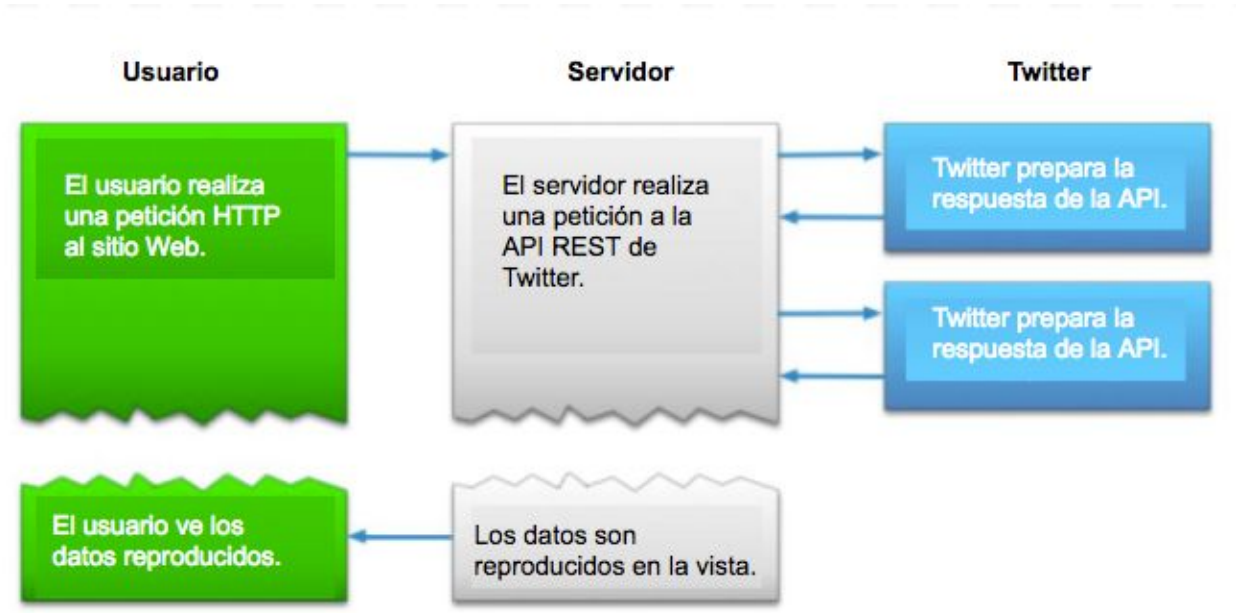

Figura 11. Diagrama de uso de la API *Public Search* de Twitter.

Puede apreciarse como, el primer paso es una petición del lado del usuario/cliente. Eso representaría una consulta realizada desde nuestro prototipo. El servidor de nuestro prototipo, entonces debería realizar no una, sino varias peticiones a la API *Public Search* de Twitter, para obtener resultados. Consideramos que esto iba a perjudicar la velocidad de recolección de tweets, ya que entre cada llamado hay un tiempo de demora entre que se envía la petición y se recibe la respuesta. Por cada respuesta obtendríamos una cantidad fija de tweets, que posteriormente tendríamos que procesar. Hasta no terminar ese procesamiento, no podríamos hacer otra petición a la API, lo cual podría hacernos perder la oportunidad de conseguir más tweets. Por otro lado, nada nos aseguraba que con cada llamada a la API no obtuviéramos tweets repetidos.

Buscando alternativas, afortunadamente hallamos dentro de los servicios de esta red social, uno que se adaptaba mejor a nuestras necesidades: la API de Streaming, que describiremos a continuación.

API de Streaming

Leyendo la documentación de Twitter, nos encontramos con que la API de Streaming nos daba la opción de obtener tweets de manera contínua y realizando una sola conexión. En palabras del sitio oficial de esta red social, a un cliente de la API de *Streaming* "se le es enviado tweets, de manera fluída y sin la necesidad ni el costo asociado a estar haciendo *polling*", esto es, sin tener que estar haciendo reiteradas peticiones. Esto era justo lo que buscábamos.

La API de Streaming funciona de manera similar a su par *Public Search*, en el sentido de que también se accede a través de peticiones HTTP, y también los clientes son identificados a través de claves de OAuth. La diferencia entre ambas radica en que en la variante *Public Search*, se abre una conexión, se obtiene un objeto JSON con una cantidad fija de datos, y luego la conexión se cierra. Para obtener varios conjuntos de datos, es obligatorio hacer varios *request* o peticiones (*polling*). En el caso de la variante *Streaming*, una vez que se abre la conexión, la API empieza a enviar JSONs al cliente hasta que el cliente decide cerrar la conexión. A continuación mostramos la figura 12 para explicar este caso de uso:

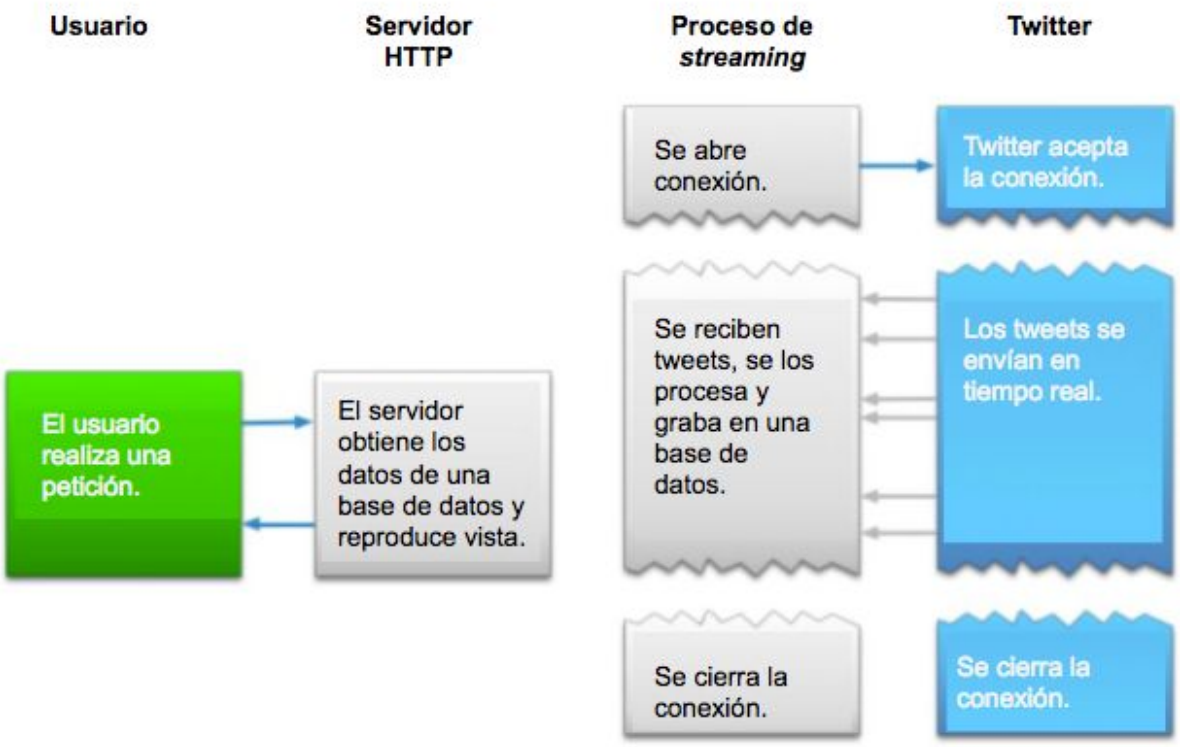

Figura 12. Diagrama de uso de la API de streaming de Twitter.

Puede observarse como el Servidor Web recibe una consulta desde el cliente/usuario. A partir de ese pedido, se hace una sola petición desde el servidor hacia la API de Streaming y se mantiene la conexión abierta, a través de la cual la información es enviada en sentido API → Servidor Web. Cabe destacar que, si el servidor Web hiciera la petición de manera directa y sincrónica, se vería bloqueado y no podría atender más peticiones de clientes. Por eso decidimos seguir la propuesta de Twitter: como puede verse en el diagrama, la idea para solventar esa dificultad es, crear un proceso paralelo que se encargue de abrir la conexión, recibir los tweets y aplicar toda la lógica siguiente, para así liberar el proceso del servidor Web.

El funcionamiento de este servicio, en el cual los datos “fluyen” y llegan en tiempo real, nos resultó el adecuado para lo que queríamos implementar, y es por esto que lo elegimos. De esta forma, podríamos mantener vivo un mismo proceso, en el cual obtendríamos tweets, y los enviaríamos a procesar por distintos servicios de Análisis del Lenguaje Natural, también expuestos a través de APIs, de manera que HTTP sería el único protocolo de comunicación

utilizado. Definimos que el proceso de recolección podría mantenerse “vivo” recolectando tweets hasta una cantidad o ventana de tiempo predefinidos.

Ya se mencionó en secciones previas que, en cada tweet obtenido, viene incluida la información del usuario que lo publicó, por lo que no era necesario hacer peticiones a la API REST.

3.4 Opinator

3.4.1 Funcionamiento y arquitectura

El software que desarrollamos es una aplicación Web que recolecta comentarios de Twitter, para luego obtener la opinión de ellos, haciendo uso de servicios externos tales como Watson de IBM y Cognitive Services de Microsoft, y de la librería para Node llamada **sentiment**.
La aplicación tiene también la capacidad de extraer algunos rasgos de los usuarios que hicieron los comentarios, tales como su género y el dispositivo utilizado. Los resultados obtenidos se usan para implementar distintas visualizaciones para sacar conclusiones y comparar datos.

Desarrollamos un proceso que recolecta todos los datos necesarios y que se ejecuta de manera paralela al funcionamiento de la aplicación web.
El proceso de recolección de datos llama a la API de Twitter con el término seleccionado por el usuario. En la medida que llegan los tweets de la API, se realizan los siguientes pasos:

1. Se extrae el género del usuario en función de su nombre, con una librería llamada **gender**, que cuenta con una base de datos de nombres recopilados a partir de censos.
2. Se extrae información sobre el dispositivo utilizado para crear el tweet. Este dato es provisto por Twitter.
3. Se envían los comentarios a ser analizados por las distintas estrategias de minería de opiniones mencionadas en la **sección 3.3**.
4. Una vez obtenidos los resultados de cada paso, se graba la información obtenida en en una base de datos.

Opinator es una aplicación Web, esto indica que su arquitectura es del tipo *cliente-servidor*. Del lado del cliente o navegador, se realizan llamadas al servidor web. Las respuestas que da el servidor de Opinator contienen los distintos conjuntos de datos obtenidos a partir del procesamiento de los tweets, mencionados previamente. Los datos son enviados en formato JSON. Al recibirse estos datos en el cliente, se usan para implementar gráficos con D3 [39] y Three.js [40]. Algunos de estos gráficos cuentan con filtros que sirven para aplicar un nivel de detalle mayor en los datos que se grafican. En la medida en que se aplican los filtros, se vuelven a hacer las llamadas al servidor para traer los conjuntos de datos filtrados por el criterio seleccionado. Los gráficos se actualizarán automáticamente, sin necesidad de actualizar la página, gracias a características tanto de D3, Three.js y de Angular, mencionados anteriormente en éste capítulo.

La interacción completa entre cliente, servidor, proceso de recolección de tweets, APIs y base de datos, puede apreciarse en la figura 13.

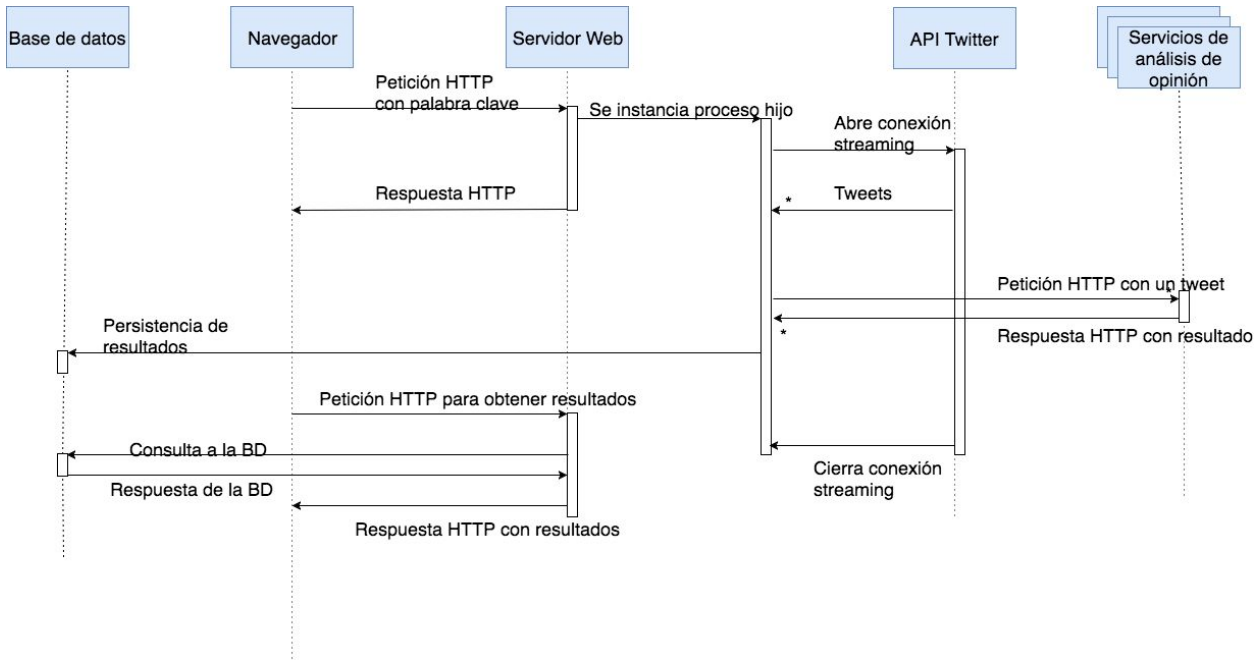

Figura 13. Diagrama que explica el flujo de búsqueda, procesamiento y presentación de resultados.

3.4.2 Recolección de lotes de comentarios

Opinator permite recolectar tweets que hacen mención de un mismo tema (1), por lotes iguales de tiempo (2), cuya recolección de tweets se activa en horarios específicos (3), durante una ventana de tiempo determinada(4). Los 4 puntos destacados son configurables. Esto quiere decir que el usuario de Opinator tiene las opciones de elegir la temática a buscar, el tiempo que van a durar los lotes de recolección, el horario en que debe comenzar la recolección para cada lote, y el tiempo total en el que se hará la recolección de lotes.

La idea detrás del concepto de lotes, es la de contar con unidades o bloques de tiempo iguales, que sirvan para hacer comparaciones entre ellos y sacar conclusiones a raíz de esa comparación.

3.4.2 Gráficos

A continuación describiremos la pantalla que presenta Opinator a su usuario, las distintas secciones de gráficos que la componen y sus funciones.

La primera sección contiene un gráfico de líneas que muestra cómo cambiaron, a lo largo del tiempo, los valores de opinión en cada lote (ver figura 14). El eje *x* representa el tiempo, mientras que el eje *y* indica el porcentaje de valores de opinión por lote, obtenido por una de las herramientas de análisis de opinión integradas. Los valores en el eje *y* son presentados en una escala porcentual para poder comparar correctamente los lotes, ya que en cada uno se puede tener una cantidad de tweets distinta. El volumen de usuarios activos en Twitter no es constante, por lo que lotes de distintas fechas y horarios pueden tener una cantidad diferente de comentarios.
El cálculo realizado para obtener un porcentaje *p* para un valor de opinión *v de* un lote *N* fué:

$$p = \frac{cantidad\ de\ comentarios\ v\ del\ lote\ N * 100}{total\ de\ comentarios\ del\ lote\ N}$$

Donde *v* representa un valor de opinión: positivo, negativo o neutral.

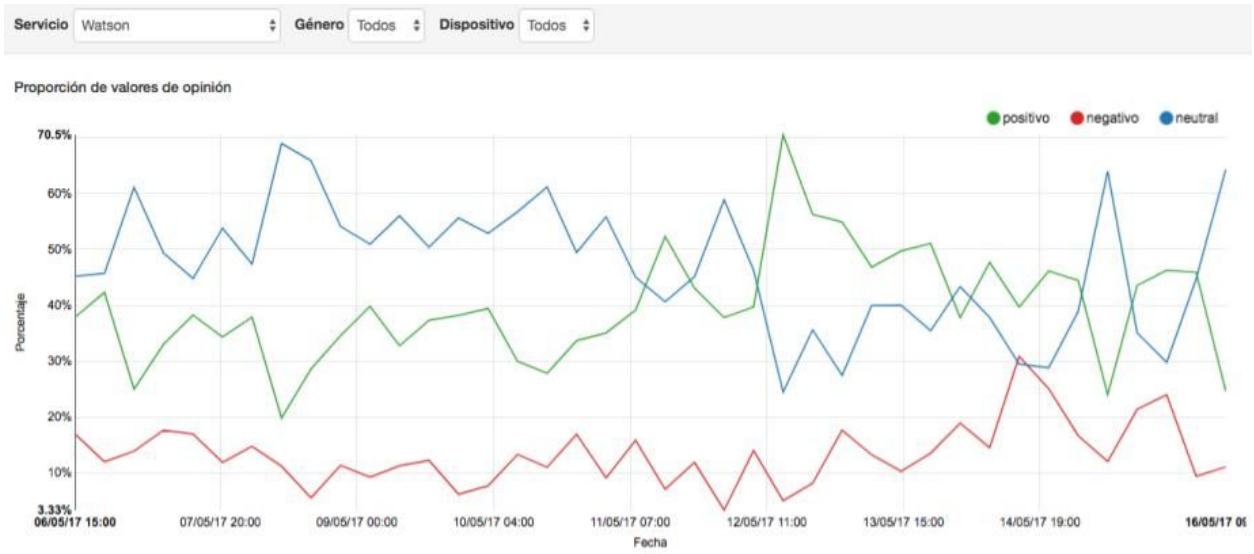

Figura 14. Línea de tiempo con los porcentajes de los distintos resultados de opinión para cada lote obtenido, según Watson.

En todo momento, se muestran los resultados de un servicio de los tres utilizados para extraer opiniones. El gráfico cuenta con líneas de color verde, rojo y azul para mostrar la evolución de los porcentajes positivos, negativos y neutrales, respectivamente. En el ejemplo de la figura 14, se muestran los resultados obtenidos con Watson.

Esta sección cuenta con tres filtros, que sirven para hacer más específico el conjunto de datos que se visualiza:

- **Servicio:** con él se puede seleccionar uno entre los tres servicios utilizados.
- **Género:** permite especificar si se desea ver los comentarios de usuarios de un género en particular.
- **Dispositivo:** permite especificar si se desea ver los comentarios de usuarios de un dispositivo en particular.

En la medida en que se seleccionan filtros, el gráfico se actualiza de manera automática, sin necesidad de refrescar la página.

Hay un gráfico de barras dentro de la misma sección, que representa el total de comentarios que recolectamos por cada lote, independientemente del valor de opinión que pudiera tener cada uno. El mismo puede observarse en la figura 15:

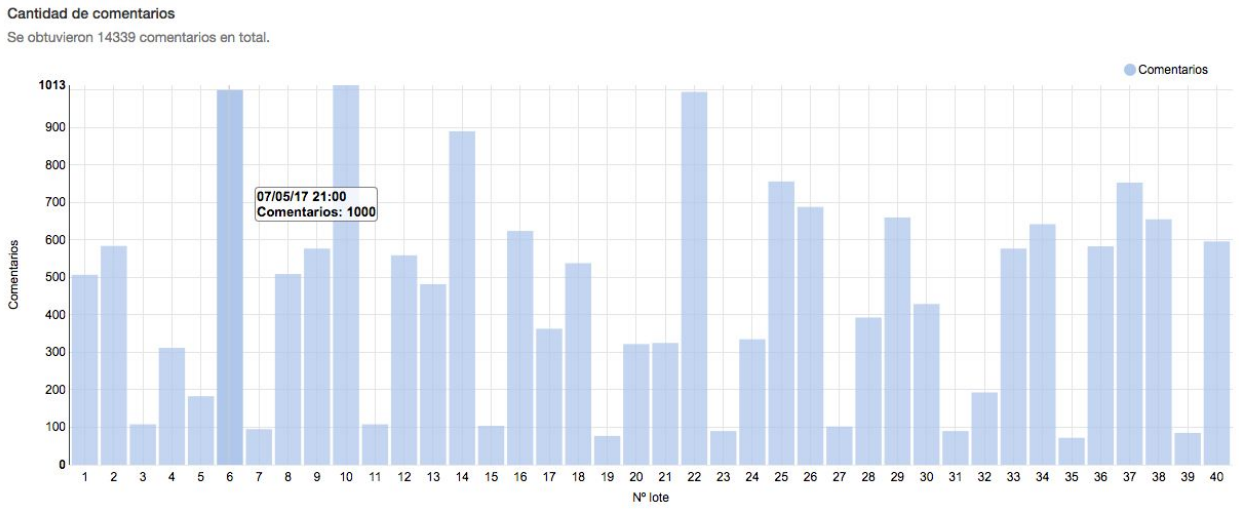

Figura 15. Gráfico de barras con la cantidad de comentarios de cada lote.

El eje *x* representa el número de lote (van del 1 al 40, y están ordenados por tiempo, desde el primer lote recolectado al último). El eje *y* representa la cantidad de comentarios. La idea detrás de este gráfico de barras es la de mostrar la diferencia de comentarios que se pudieron recolectar para cada lote.

La sección que le sigue puede observarse en la figura 16:

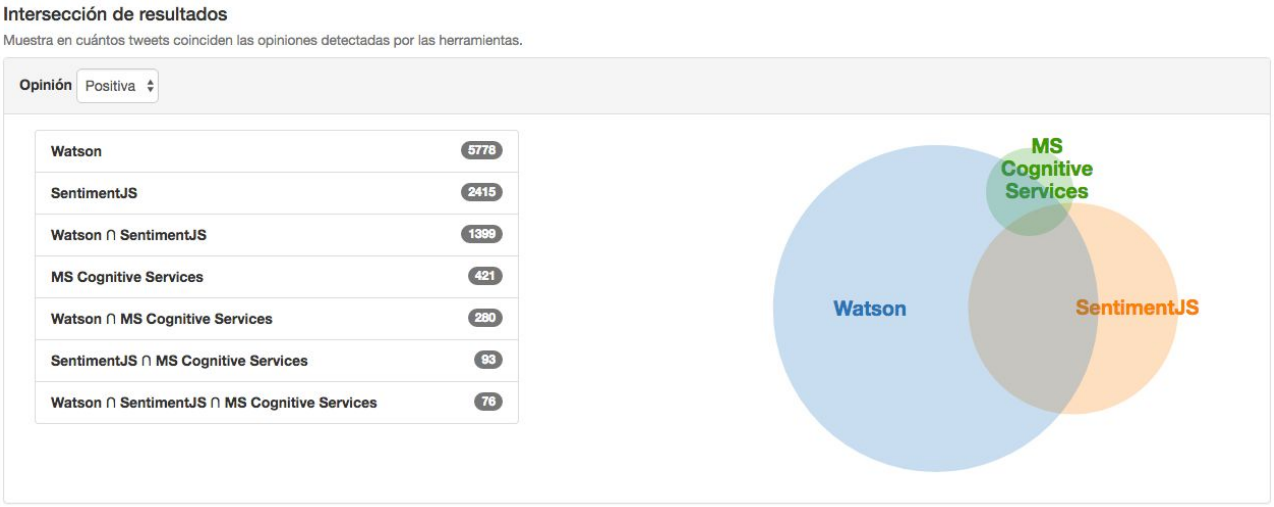

Figura 16. Diagrama de Venn que muestra la cantidad de comentarios en donde las herramientas coincidieron en cuanto al valor de opinión que detectaron.

Dicha sección contiene un diagrama de Venn, que fué incluido para observar en cuantos comentarios coincidieron las herramientas respecto al valor de opinión que infirieron. También sirve para identificar la diferencia de proporción en la cantidad de comentarios con un valor de opinión para las distintas herramientas.

Además del gráfico, la sección cuenta con una tabla referencial que muestra las cantidades de cada conjunto exhibido en el diagrama, incluyendo las intersecciones.

El gráfico tiene características funcionales similares al de la figura 14. Esto quiere decir que puede filtrarse, y que también se actualiza cuando se hacen cambios en el filtro. En este caso, puede filtrarse por:

- **Servicio:** con él, se puede cambiar el valor de opinión. Esto quiere decir que se puede elegir observar la intersección en los resultados de las herramientas para comentarios positivos, negativos, o neutrales.

La sección siguiente en la aplicación se ve en la figura 17:

Figura 17. Gráfico de líneas que muestra el valor de opinión que detectó una herramienta en cada tweet para un usuario específico.

Esta sección fue pensada para mostrar el valor de opinión de los tweets obtenidos de un usuario. Cuenta con un gráfico de líneas, en donde el eje *x* se usa para representar el orden temporal de los tweets, y el eje *y* tiene los tres valores de opinión posibles. Al situar el cursor sobre cada punto de la curva, aparece un descripción con la hora exacta en que fué emitido y la opinión del comentario que representa ese punto (ver figura 17).

La idea detrás de este gráfico, es la de mostrar la evolución de la opinión en los comentarios de un usuario, percibida por una herramienta. La función es similar la visualización de la figura 14, pero el enfoque es distinto, ya que se centra en un sólo usuario.

Se tienen dos filtros en esta sección:

- **Usuario:** Se carga con los 5 usuarios que más comentarios tuvieron en la ventana de tiempo de recolección de tweets. Con él, se puede seleccionar un usuario para ver sus comentarios.
- **Servicio:**. Permite seleccionar una de las tres herramientas usadas en Opinator, para ver los valores de opinión recolectados con una en particular. Se muestran los valores de una herramienta a la vez.

Por último encontraremos una sección que contiene un gráfico de cajas y bigotes utilizado para mostrar tendencia central y dispersión en la clasificación de las librerías, presente en la figura 18.

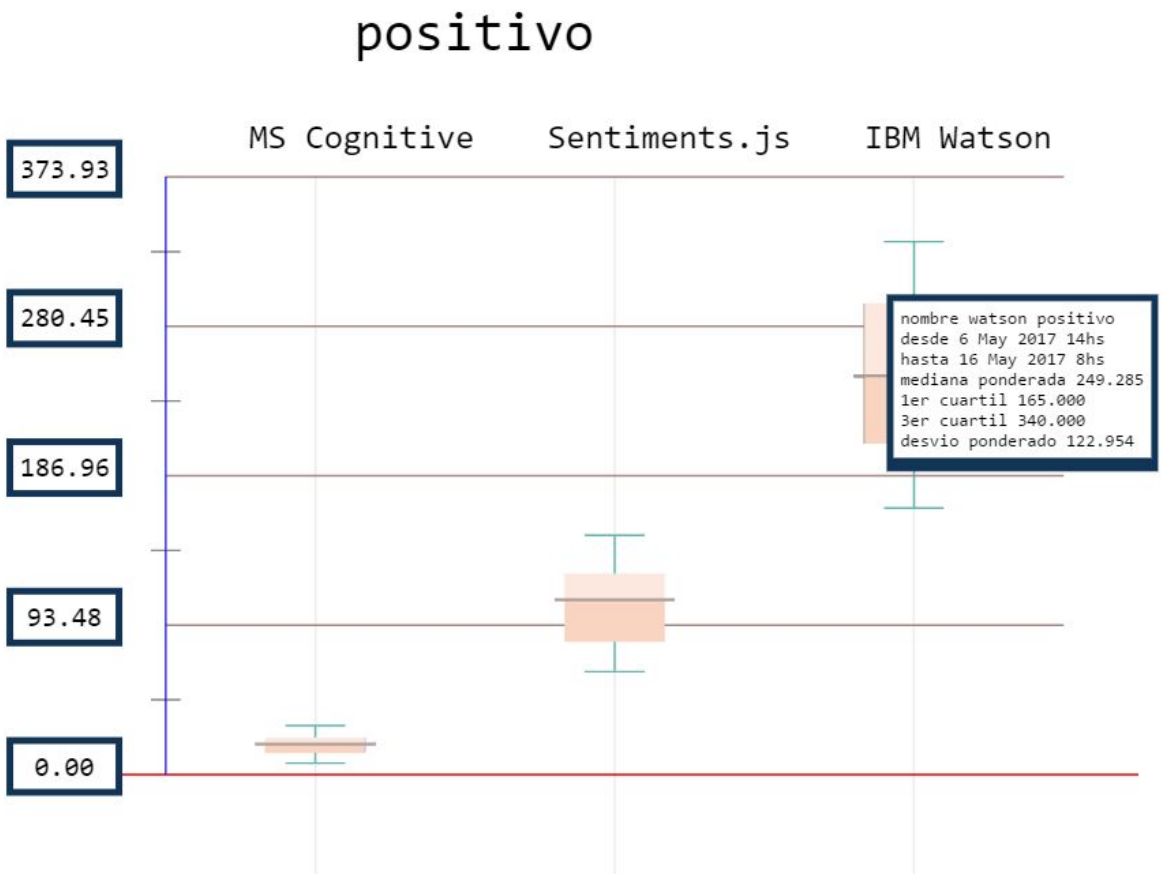

Figura 18.1. Gráfico de cajas y bigotes que muestran la tendencia central en forma de mediana ponderada y la dispersión en forma de primer y tercer cuartil con su desvío.

Figura 18.2. Tooltip del gráfico de cajas y bigotes que muestran la tendencia central en forma de mediana ponderada y la dispersión en forma de primer y tercer cuartil con su desvío.

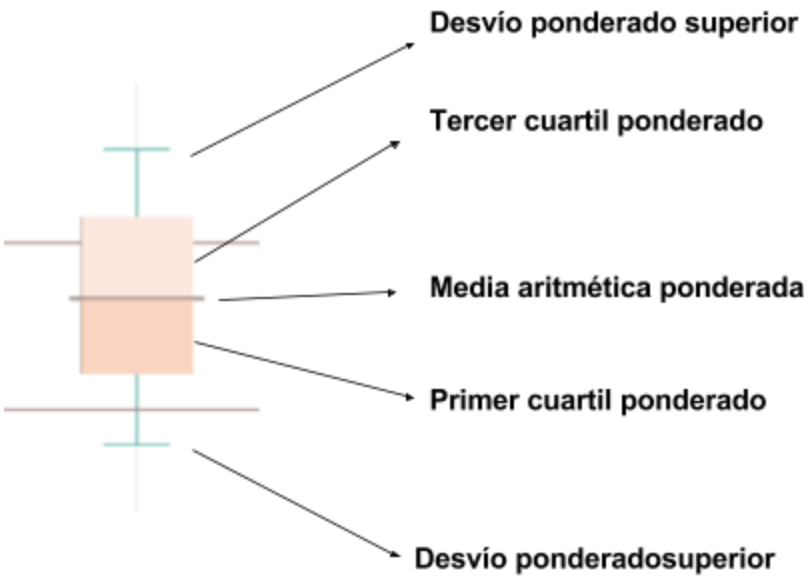

Figura 18.3. Descripción de un componente del gráfico de cajas y bigotes.

Para la generación del gráfico de cajas y bigotes obtuvimos la mediana ponderada de los 40 lotes con su correspondiente escalado para cada librería en una valoración dada. Es decir, se promediaron los positivos, negativos y neutrales de los 40 lotes en las tres librerías. De este modo se generaron valores de tendencia central y dispersión, utilizando los valores del mediana ponderada, desvío estándar y primer y tercer cuartiles.

Con los valores obtenidos, se visualizan tres gráficos mostrando en cajas y bigotes los mediana ponderada, desvíos, primer y tercer cuartiles de las valoraciones para las tres librerías para una valoración dada por vez. Estos gráficos pueden intercambiarse utilizando un selector de valoraciones, de este modo podemos observar los valores de tendencia central y dispersión para una clasificación dada por las tres librerías y observar a cuántos puntos de distancia están los valores y sacar conclusiones con respecto al tipo de clasificación que tienen las librerías. Lo que se quiere observar en este gráfico es si se pueden ver tendencias en el tipo de clasificación de las librerías. Por ejemplo, si una librería tiende a clasificar más positivos que otra. Más tarde, profundizaremos sobre esto en la sección de ensayos realizados.

La herramienta puede generar de forma dinámica este tipo de gráficos, posee funcionalidades básicas para cálculos de valores estadísticos, normalización de datos, parseo de formatos que permiten ingresar valores en el formato estándar de Opinator y ser graficado. El gráfico fue realizado para Opinator utilizando tecnologías web.

Como en la recolección de datos puede ocurrir que no se complete un lote por cantidad, sino por tiempo, nos encontramos con lotes de distinta cantidad de comentarios. Como nos interesa poder comparar y estudiar la evolución de los comentarios en dichos lotes, es necesario establecer un marco común de cantidades para relacionar los lotes. Ante esta problemática utilizamos la técnica de escalado. Esta técnica permite que todos los lotes estén en la misma escala al momento de su análisis. Para esto se divide la cantidad de comentarios de una clasificación determinada sobre el total. Cada clasificación tendrá entonces un valor entre 0 y 1 dependiendo su proporción sobre el total. Esto nos permitiría poder analizar conjuntamente lotes de tamaños variable, ya que no estaremos comparándolos en su valor absoluto sino en un valor proporcional igual para todos los lotes [41].

Mediana ponderada de datos: el gráfico de cajas y bigotes utiliza mediana ponderada para comprar la clasificación que realizan las librerías. Esto se debe a que queremos obtener una entre lotes de distintos tamaños, por ende, cada lote deberá estar ponderados de acuerdo a su tamaño: aquellos con mayor tamaño tendrán más peso en los cálculos. De esta forma se utiliza la mediana ponderada [42], el desvío ponderado[43] y percentiles ponderados[44], en particular el percentil ponderado 25 y el percentil ponderado 75, es decir los cuartiles ponderados 1 y 3. Estos valores estadísticos tiene la particularidad de tener en cuenta no sólo el valor de cada elemento del dataset, sino el valor total del momento de la toma. Así, por ejemplo, una toma de 200 elementos tendrá menos importancia al momento de tomar sus

valores en los cálculos que una toma de 1000 elementos. Esto es así considerando que tiene mayor peso estadístico aquellas tomas con mayor volumen de datos.

Funciones estadísticas utilizadas: las funciones de tendencia central y dispersión utilizadas y el gráfico de visualización fueron realizadas para los ensayos y visualización de datos de tendencia central. Las funciones estadísticas serán descritas a continuación.

Función de mediana ponderada $\overline{X} = (\sum_{i=1}^{n} w_i * x_i) \ / \ \sum_{i=1}^{n} w_i$

Función de desvío ponderado, donde $\overline{X}$ es la mediana ponderada $\delta = \sqrt[2]{(\sum_{i=1}^{n} w_i * (x_i - \overline{X})^2) / \sum_{i=1}^{n} w_i}$

La función de cuartiles ponderados es la función de percentiles ponderados que se describe a continuación. Consiste en ordenar el conjunto de datos de menor a mayor según su valor. Luego sumar la cantidad total de datos, luego este total se multiplica por p * 0.01, donde p es el percentil a calcular, y se guarda en una variable W_p. Luego se hace una sumatoria de los totales ordenados hasta que el valor sumando sea mayor que W_p, en ese punto se obtendrá tal valor de percentil.

4. Ensayo realizado

4.1 Caso de estudio

En este capítulo hablaremos del ensayo que hicimos para validar el funcionamiento del software que desarrollamos para este trabajo, y también para comparar los resultados arrojados por las distintas herramientas que utiliza, mencionadas en el capítulo 3, sección 3 (*Servicios y librerías externas*).

Recordemos del capítulo anterior que Opinator puede ser configurado para recolectar tweets sobre un mismo tema (1), de a lotes iguales de tiempo (2) que se “activan” en horarios

específicos (3), durante un período de tiempo (4). Para nuestro experimento, configuramos esos cuatro puntos de la siguiente manera:

1. El tema buscado fué “Cambiemos”,
2. El tiempo de duración de cada lote fue de 1 hora cada uno,
3. La recolección de lotes se “activaba” cada 6 horas: a las 9:00hs, las 15hs, las 21hs y las 3:00hs,
4. El período de tiempo total de recolección fué configurado para que dure 10 días. Se inició el 6 de Mayo de 2017 y finalizó el 16 de Mayo de 2017.

Esto nos permitió recolectar en total **40 lotes** correspondientes a distintos horarios, que mencionan a *Cambiemos*, el partido político que ejerce el Gobierno Nacional Argentino al momento del desarrollo de este trabajo. Decidimos buscar este tema, ya que es una palabra para la cual podíamos obtener una cantidad razonable de comentarios de durante un período de 10 días. Habíamos considerado otras temáticas de la actualidad Argentina, pero observamos que la atención en torno a ellas podía durar pocos días, lo que se traduciría en pocos comentarios para extraer en todo el período de tiempo planteado.

4.2 Observaciones

4.2.1 Cantidad de comentarios

La primera observación que haremos será sobre la cantidad de comentarios recolectados. Una vez completado el tiempo de recolección de tweets de 10 días y obtenidos los 40 lotes, observamos que conseguimos un total de **14339 comentarios**. Valiéndonos del gráfico de barras de la figura 19, sabemos que el lote en el cual más tweets se pudieron juntar fué el número 10, que corresponde a las 21:00 horas del 8 de Mayo de 2017. En él se recolectaron 1013 comentarios. El lote en el que menos tweets consiguieron fué el número 19, que corresponde a las 03:00 horas del 11 de Mayo de 2017. En él se recolectaron 77 comentarios.

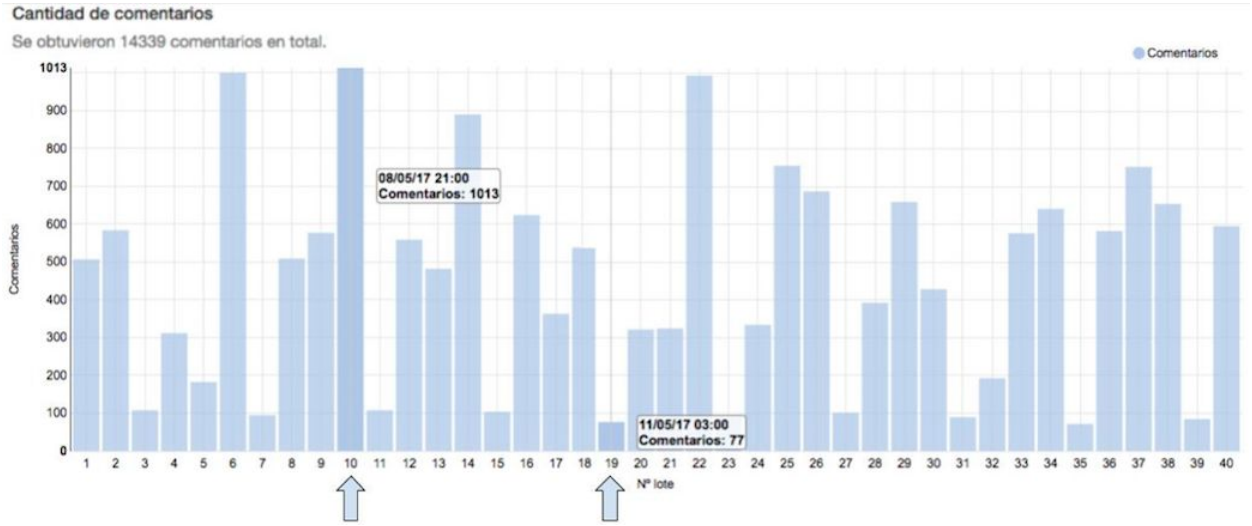

Figura 19. Se indican los lotes número 10 y 19, con la cantidad máxima y mínima de tweets, respectivamente.

Algo para destacar, es que los lotes de las 03:00 (el número 3, el 7, el 11, etc) tienen todos una cantidad de tweets menor, si se los compara con lotes de otros horarios. Lo que esto sugiere, es que, respecto al tema buscado, hubo notablemente menos participación de los usuarios en los lotes de las 03:00 de la mañana, en comparación a los demás de las 09:00, 15:00 y 21:00 horas.

4.2.2 Valores de opinión

A continuación haremos una descripción de los resultados obtenidos con cada servicio de análisis de opinión, mostrando las figuras correspondientes, y realizando observaciones para cada una.

Aquí haremos tres análisis, tomando todos los lotes:

- Sin aplicar filtros,
- Filtrando por género,
- Filtrando por dispositivo.

De la descripción de este gráfico en el capítulo 3, vale recordar que en el eje *y* mostramos una escala porcentual, que representa el porcentaje para cada valor positivo, negativo o neutral, en relación al total de comentarios de cada lote. Usamos este valor porcentual en la representación para estandarizar los resultados de los cuarenta lotes y así poder compararlos, ya que cada uno podía tener una cantidad de tweets distinta.

Centrémonos primero en el análisis sobre todos los lotes, sin aplicar filtros. La figura 20 muestra las proporciones de los 40 lotes de los resultados de Watson.

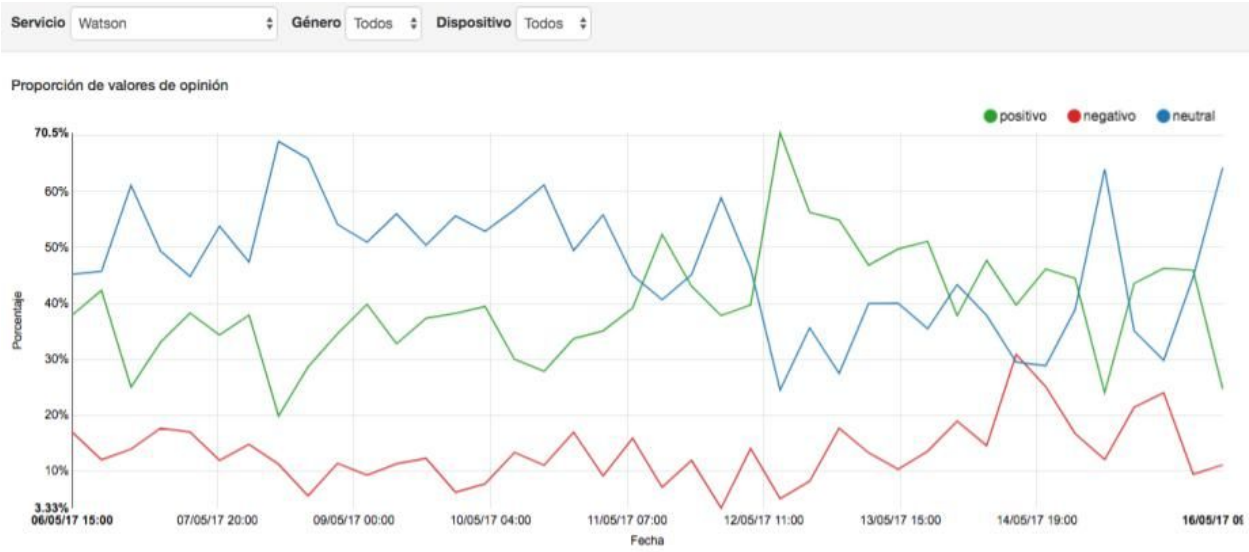

Figura 20. Evolución temporal de los resultados sin filtrar según Watson.

Los resultados de Watson pueden describirse por etapas. La primera etapa, que se extiende desde las 15:00 horas del 06/05/17 hasta las 09:00 horas del 11/05/17, se caracteriza por tener a los tres indicadores de valores de opinión marcadamente separados, y en donde la mayoría de comentarios detectados fueron neutrales, seguido en cantidad por el grupo de sentimientos positivos. La variante negativa es la que menor cantidad registra. En esta etapa, no se detectan cambios significativos en las 3 líneas.

La segunda etapa en los resultados de Watson dura desde las 15:00 horas del 11/05/17 hasta las 09:00 horas del 14/05/17. Ya en el primer lote de esta etapa se observa como crece la variable positiva para llegar a ser el grupo mayoritario, y decrece el grupo de comentarios neutrales, para quedar en un segundo lugar. Si bien esta situación se revierte en el lote siguiente, el indicador positivo vuelve a ser el mayoritario en los lotes siguientes, y se mantiene durante poco menos de dos días.

Finalmente, la tercer etapa se extiende durante dos días, desde las 15:00 horas del 14/05/17, hasta la misma hora del 16/05/17. La misma se destaca porque las 3 variables oscilan de forma marcada, llegando el indicador de comentarios neutrales a quedar como el grupo minoritario, superado por su par negativo.

Pasemos ahora a los resultados de Microsoft Cognitive Services (figura 21).

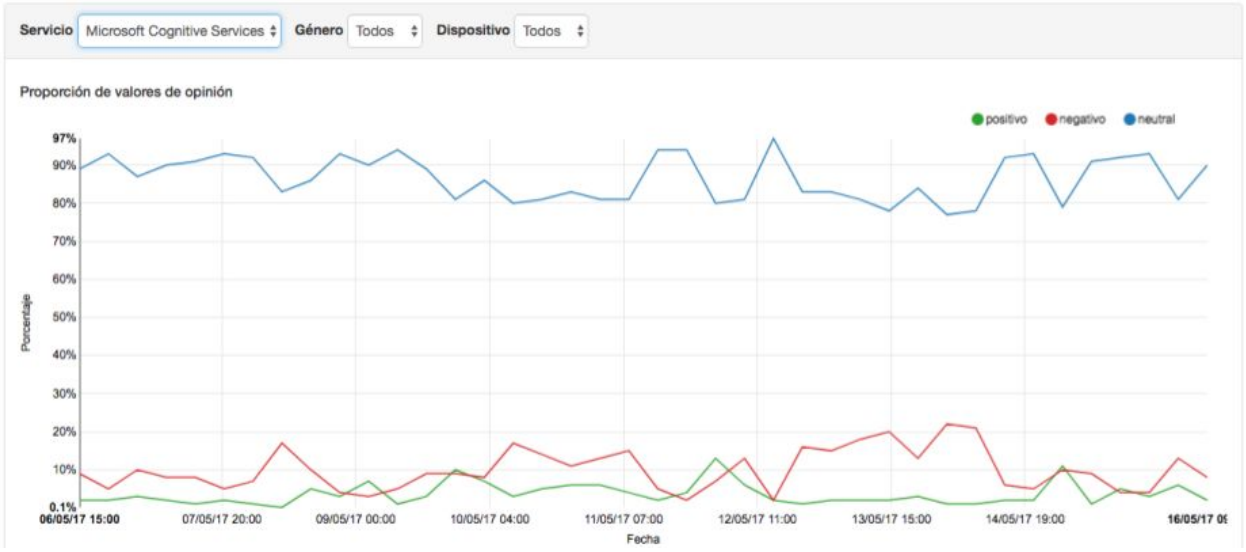

Figura 21. Evolución temporal de los resultados sin filtrar según Cognitive Services.

La gráfica nos muestra que la amplia mayoría de comentarios fue considerada neutral por esta herramienta. El porcentaje de valores neutrales tuvo entre un mínimo de 76.67% y un máximo de 96.69%, mientras que los comentarios positivos y negativos oscilaron, sumando ambos grupos, entre un mínimo de 3.31% y un máximo de 23.33%. Los comentarios positivos suelen ser los que menos cantidad registran, salvo en algunos pocos lotes, en donde superan a la cantidad de sentimientos negativos. Observamos que esto sucede en lotes aislados. Cuando la cantidad de positivos supera a los negativos, es solo durante un lote, y la diferencia no se mantiene en los lotes siguientes.

No se observan picos de alza o caída abruptos en las líneas, pero se destaca que a las 09:00 horas del 12/05/17, el porcentaje de opiniones negativas era de 1.59% y luego, a las 15:00 hs del mismo día, el mismo indicador subió hasta un 16.28%, para luego caer a un 6.24% a las 15:00 del 14/05/17. Durante este tiempo, esta variable se mantuvo sin caídas y alcanzó picos de 20.30% y 22.22%. Este comportamiento se extendió por dos días, y coincide con el período más bajo de tweets neutrales, siempre según el criterio de detección de opiniones de Microsoft Cognitive Services.

Veamos los resultados de Sentiment (figura 22):

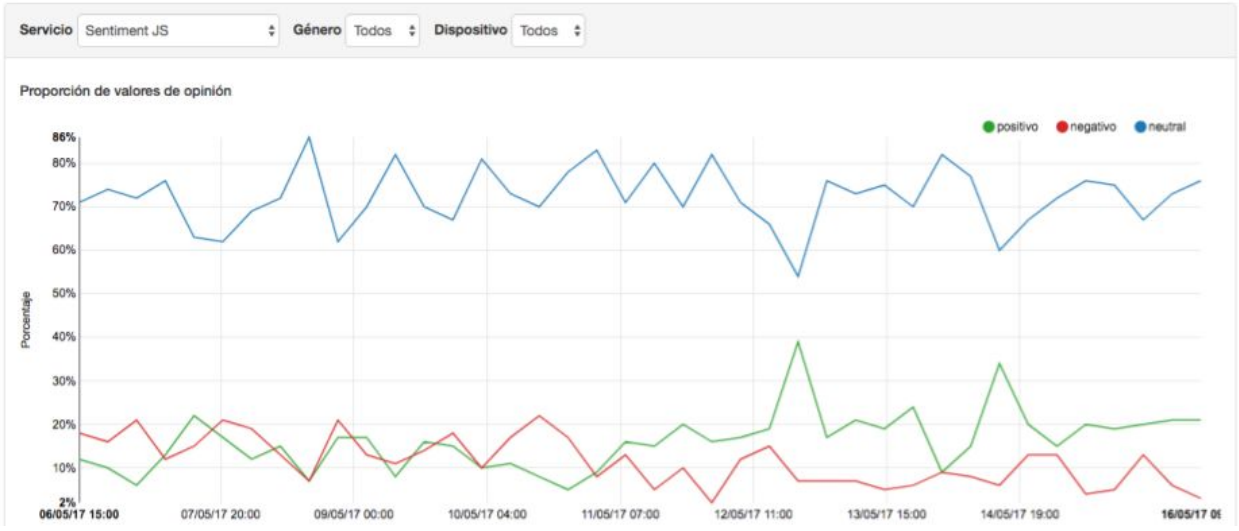

Figura 22. Evolución temporal de los resultados sin filtrar según Sentiment.

El gráfico nos muestra que esta herramienta halló a la mayoría de comentarios como neutrales en cada uno de los lotes, similar a Cognitive Services. De todas formas, hay algunos picos en donde el porcentaje del grupo de comentarios neutrales bajó considerablemente, mientras que subió el porcentaje de comentarios positivos. Hay dos casos notorios que describimos a continuación.

El primero de los casos se da en el **lote 26**, correspondiente a las 21:00 horas del 12/05/17. En el lote anterior (de las 15:00 horas del 12/05/17), observamos que los valores fueron los siguientes: comentarios positivos: 18.78%, negativos: 15.08%, neutrales: 66.14%. En el **lote 26**, se observa un marcado ascenso de los comentarios positivos a un 38.95%, lo que significa una diferencia de poco más del 20% con el valor anterior del mismo indicador. Naturalmente, descendieron los otros dos indicadores; siendo los comentarios neutrales los que tuvieron una caída más marcada a un 54.07%, mientras que la variable negativa bajó a un 6.98%.

El otro caso en donde esto ocurre es en el **lote 33**, que corresponde a las 15:00 horas del 14/05/17. El lote anterior (de las 09:00 horas del 14/05/17) tuvo los siguientes valores: comentarios positivos: 14.51%, negativos: 8.29%, neutrales: 77.2%. En el **lote 33**, los comentarios positivos ascendieron a un 34.49%. Los indicadores negativos y neutrales descendieron a 5.72% y 59.79%, respectivamente.

Cabe destacar que en ambos casos, el cambio se produce solo por un lote, y no se mantiene en lotes sucesores.

Continuemos con el estudio de los resultados filtrados por género. Para este caso, filtramos los resultados para ver las opiniones de todos los usuarios cuyo género fue detectado como “Hombre”. Obtuvimos **7408 comentarios** de usuarios catalogados en este género. Las figuras 23, 24, y 25 tienen las gráficas de Watson, Cognitive Services y Sentiment, respectivamente:

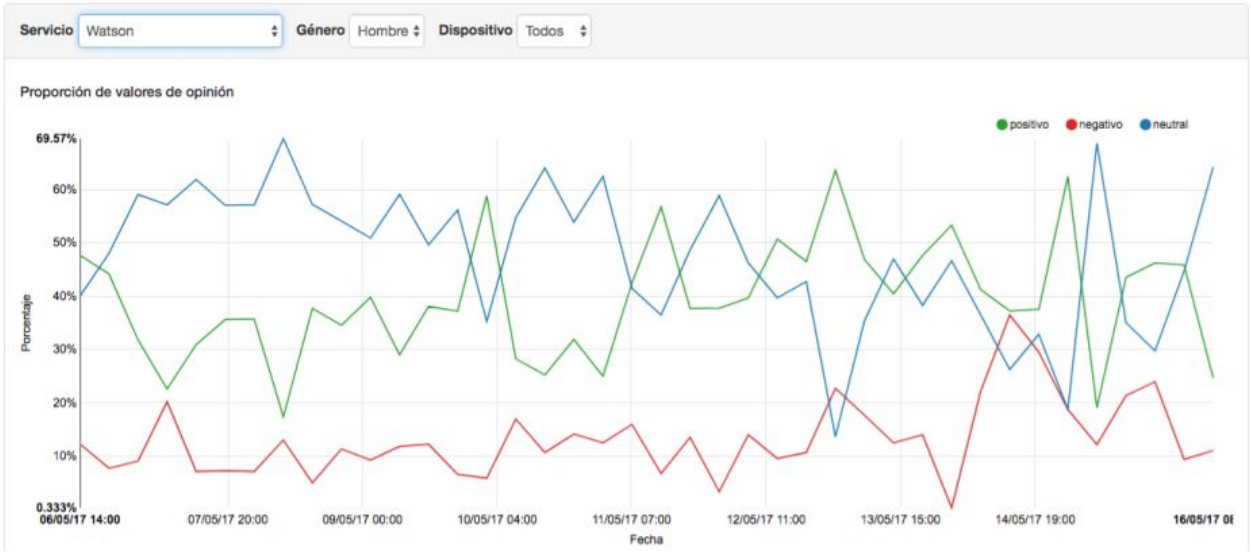

Figura 23. Evolución temporal según Watson de los resultados filtrados por género.

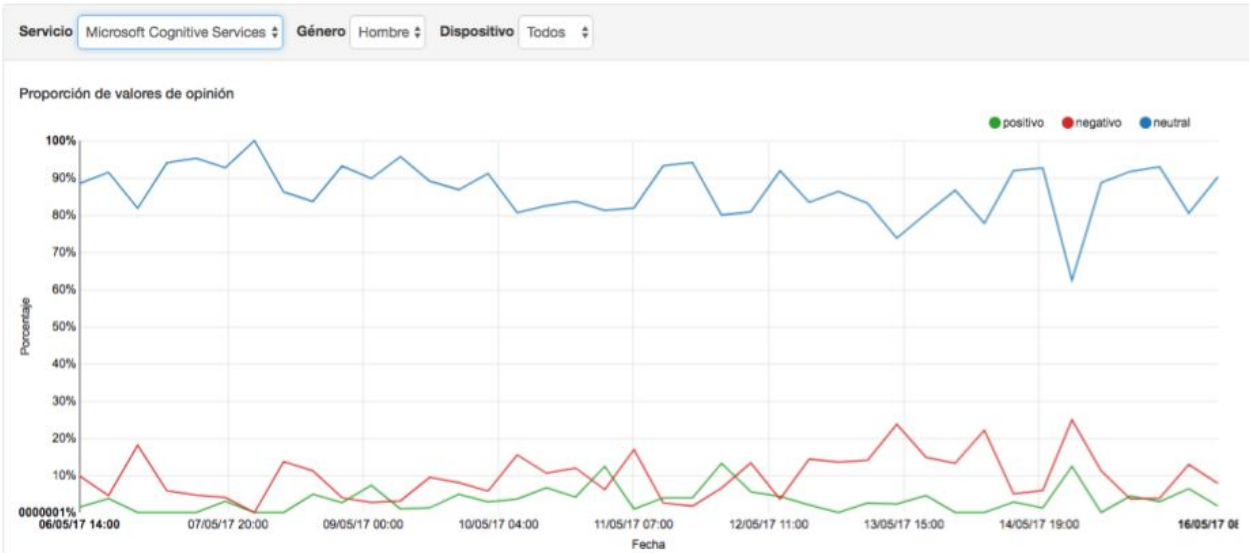

Figura 24. Evolución temporal según Cognitive Services de los resultados filtrados por género.

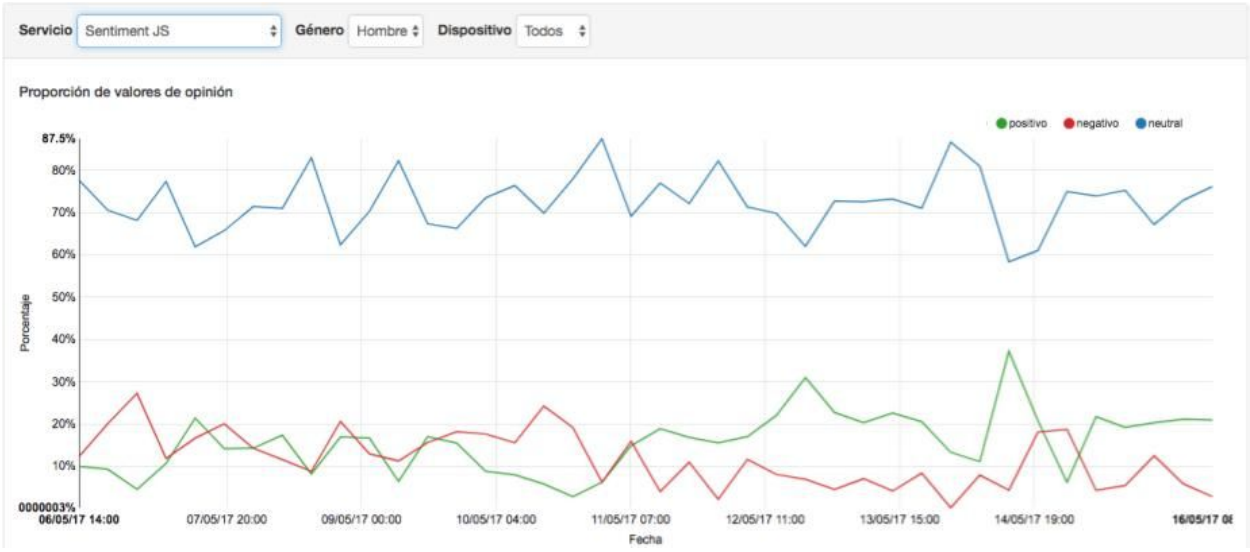

Figura 25. Evolución temporal según Sentiment de los resultados filtrados por género.

Podemos notar escenarios similares a los que notamos con los datos no filtrados. Para Cognitive Services y Sentiment, la mayoría de comentarios fueron detectados neutrales, quedando los valores positivo y negativo oscilando en cantidades bajas. Aquí también notamos en ambos una ventana de tiempo en donde se da una baja considerable de los valores neutrales, teniendo en cuenta los valores que venían teniendo en lotes anteriores. En el caso de Cognitive Services (figura 24), la caída de valores neutrales se produce el día 15 de Mayo a las 03:00, y se elevó la cantidad de valores negativos. Con Sentiment (figura 25), la caída de opiniones del mismo valor se produce el 14 de Mayo a las 15:00, y se elevan los valores positivos. Esto sugiere que pudieron haber factores en ambas fechas que llevaron a las dos herramientas a obtener resultados notoriamente distintos a los que venían teniendo.

En el análisis de lotes no filtrados, habíamos observado que Watson era la herramienta que más oscilación de resultados había. Filtrando por género, con Watson (figura 23) observamos que las variaciones en los tres valores de opinión son aún más evidentes.

Pasemos finalmente a los resultados filtrados por dispositivo. Para este caso seleccionamos el valor “Android”, para ver los valores obtenidos de comentarios emitidos desde esa plataforma móvil. Obtuvimos **11893 comentarios** de usuarios cuyo dispositivo usa ese sistema operativo. Las figuras 26, 27 y 28 presentan los valores de Watson, Cognitive Services y Sentiment, respectivamente.

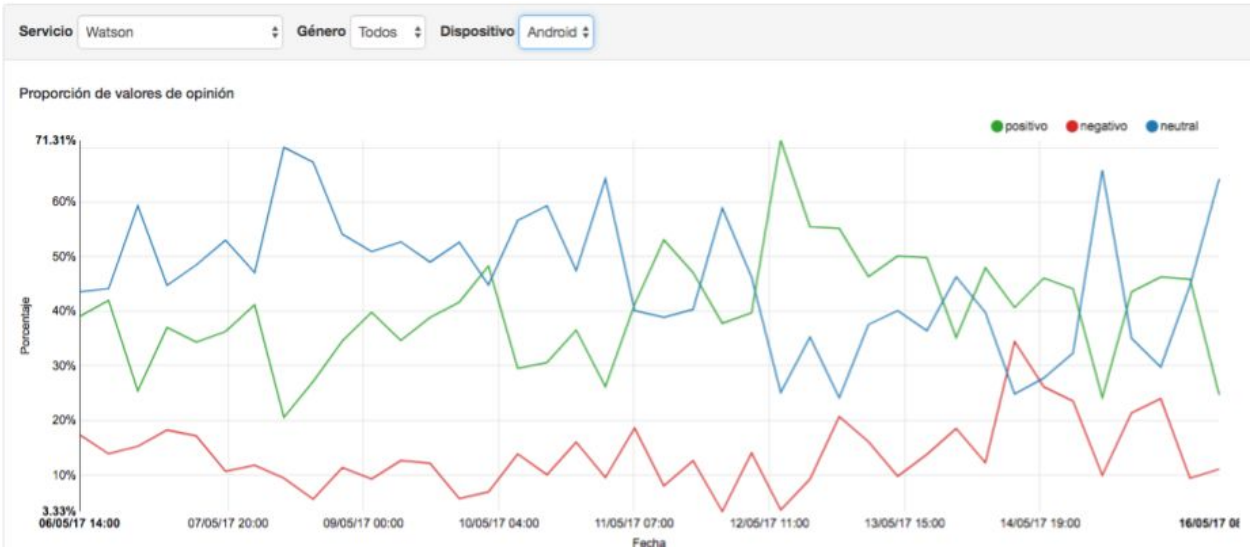

Figura 26. Evolución temporal según Watson de los resultados filtrados por dispositivo.

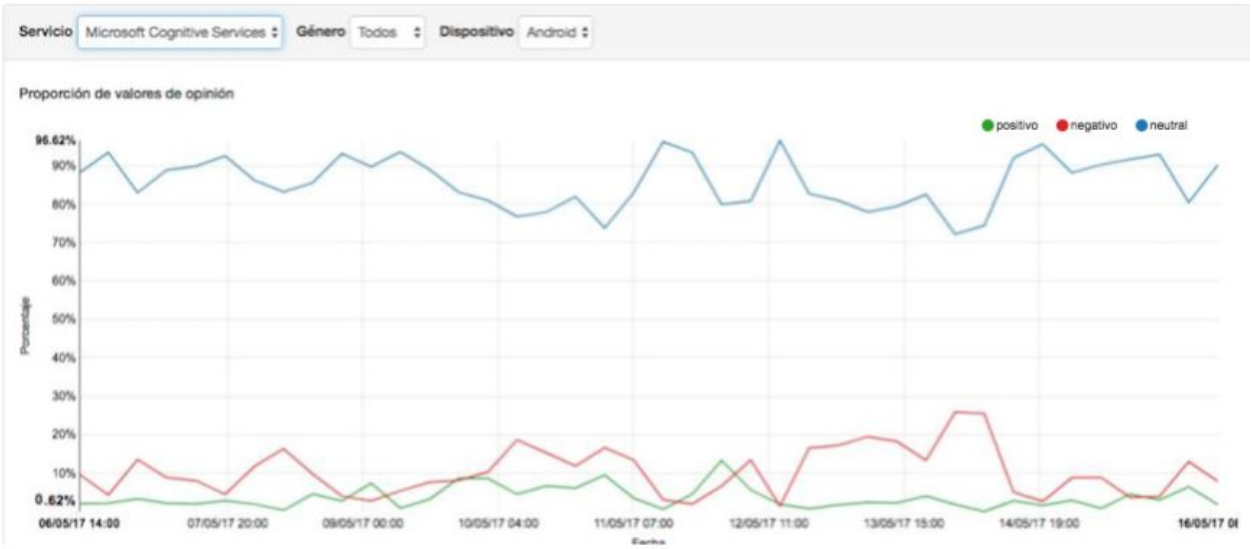

Figura 27. Evolución temporal según Cognitive Services de los resultados filtrados por dispositivo.

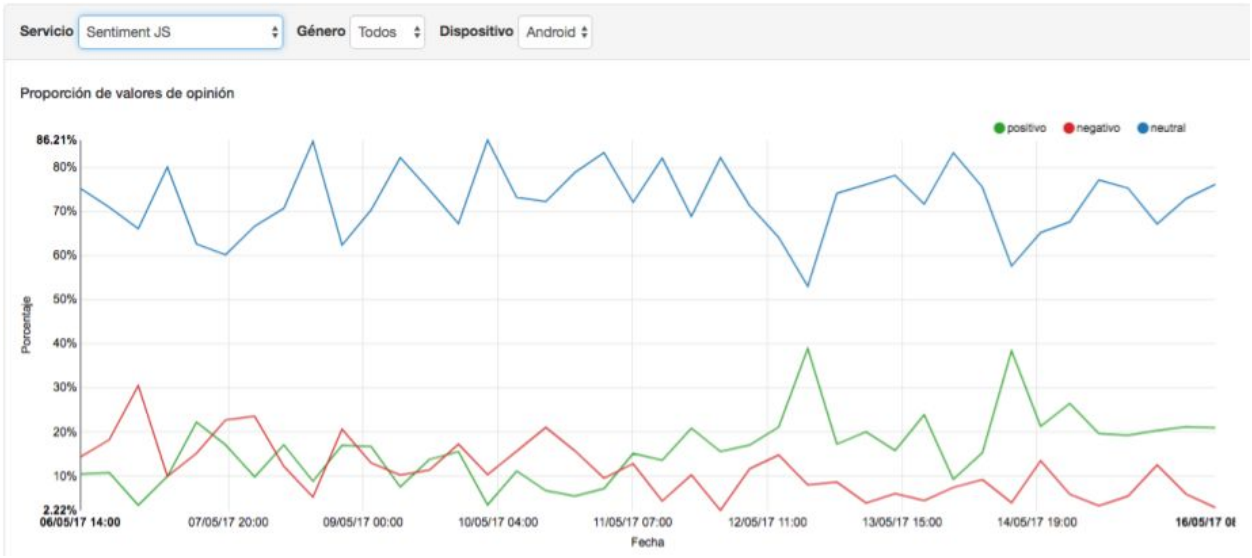

Figura 28. Evolución temporal según Sentiment de los resultados filtrados por dispositivo.

Estas gráficas se asemejan mucho a las de los resultados sin filtrar. Observemos, por ejemplo, que en los resultados de Sentiment (figura 28) se observan dos picos de subida de comentarios positivos, en detrimento de la cantidad de neutrales, puntualmente los días 12 de Mayo a las 21:00 y el 14 de Mayo a las 15:00, como también sucede en la gráfica de lotes sin filtro.

Lo mismo ocurre si comparamos la gráfica de Cognitive Services (figura 27) con filtro de dispositivo y sin filtro. En ambas se observa una subida de valores negativos, en detrimento de los valores neutrales, que comienza el 12 de Mayo a las 21:00, y se mantiene durante algunos lotes para bajar el 14 de Mayo a las 09:00.

El caso de Watson filtrado por dispositivo (figura 26) también es muy similar a la gráfica de lotes sin filtrar. Aplica la misma descripción por “etapas” que realizamos para la figura 20.

En resumen, se puede observar que las tres herramientas tuvieron algún cambio en el “comportamiento” de las líneas en distintos momentos de los últimos 3 días de recolección, si se lo compara con los lotes de días anteriores. Llegar al fondo del asunto para entender por qué sucedió esto, está fuera del análisis que queremos realizar en este trabajo, ya que para esto, habría que estudiar el contexto en el que se dieron los comentarios en esa etapa de tiempo. Lo que rescatamos de estas observaciones, es que las tres herramientas, ante un mismo conjunto de datos de entrada, tuvieron resultados notoriamente diferentes entre sí. Otra

observación que nos parece importante es que Cognitive Services y Sentiment catalogaron a una amplia mayoría de comentarios como neutrales.

4.2.3 Intersección de resultados

En esta sección, centraremos el análisis en el diagrama de Venn que implementamos, ya mencionado en el capítulo 3.

Cabe recordar que dicho diagrama muestra, para un valor de opinión, los conjuntos de resultados de las tres herramientas, y sus intersecciones. Principalmente quisimos observar en qué cantidad de comentarios las herramientas coincidieron respecto al valor de opinión inferido.

Consideramos que, si la cantidad de elementos en las intersecciones de este diagrama se acercaba a la cantidad de los conjuntos tomados de manera independiente, esto podría tomarse como un indicio que sugiere que las herramientas coincidieron en la mayoría de comentarios.

Observemos las gráficas obtenidas al aplicar los datos recolectados en el lapso de 10 días. La figura 29 corresponde al diagrama de intersección de comentarios positivos para las tres herramientas:

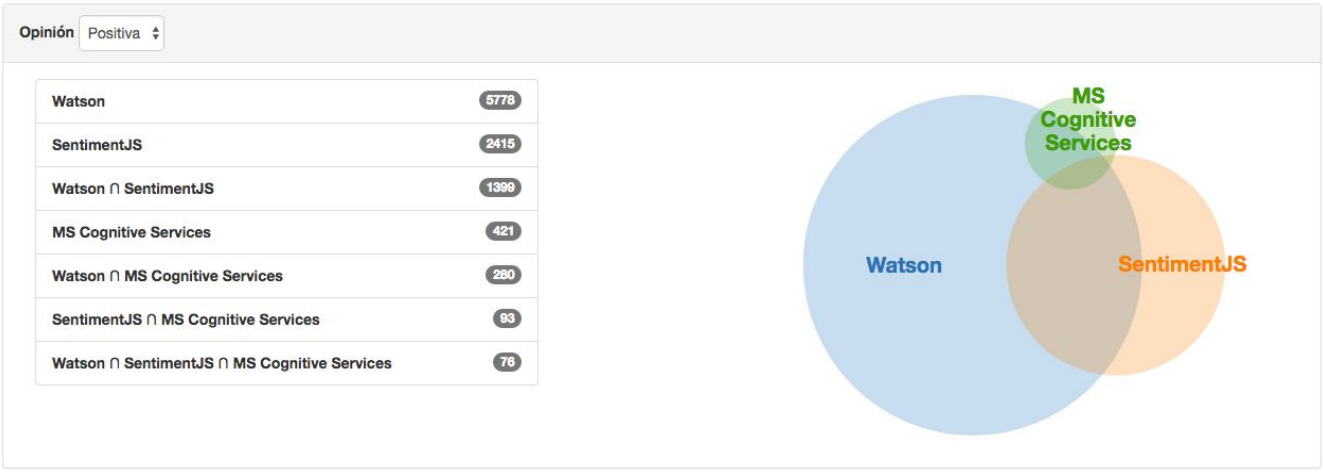

Figura 29. Diagrama de Venn con los conjuntos de comentarios positivos detectados por cada herramienta, y sus intersecciones.

Podemos ver como Watson fue el servicio que más comentarios positivos detectó, con 5778 comentarios, seguido por SentimentJS en segundo lugar con 2415. Microsoft Cognitive Services fué la herramienta que menos comentarios positivos detectó, con un total de 421.

Algo para destacar, es que la intersección **Watson ∩ SentimentJS** tiene más ocurrencias (1399) que el conjunto de Cognitive Services tomado de manera independiente. Estos resultados son coherentes con los observados en la sección 4.2.2, en donde Watson fue la herramienta que más comentarios positivos detectó, mientras que la variante de Microsoft fué la que menos obtuvo.

Finalmente, observamos que la intersección **Watson ∩ SentimentJS ∩ MS Cognitive Services** tiene solo 76 elementos, por lo que podemos decir que las herramientas coincidieron en que los comentarios eran positivos en un **0.53% del total de 14399 comentarios**.

La figura 30 muestra la intersección de comentarios negativos para las tres herramientas:

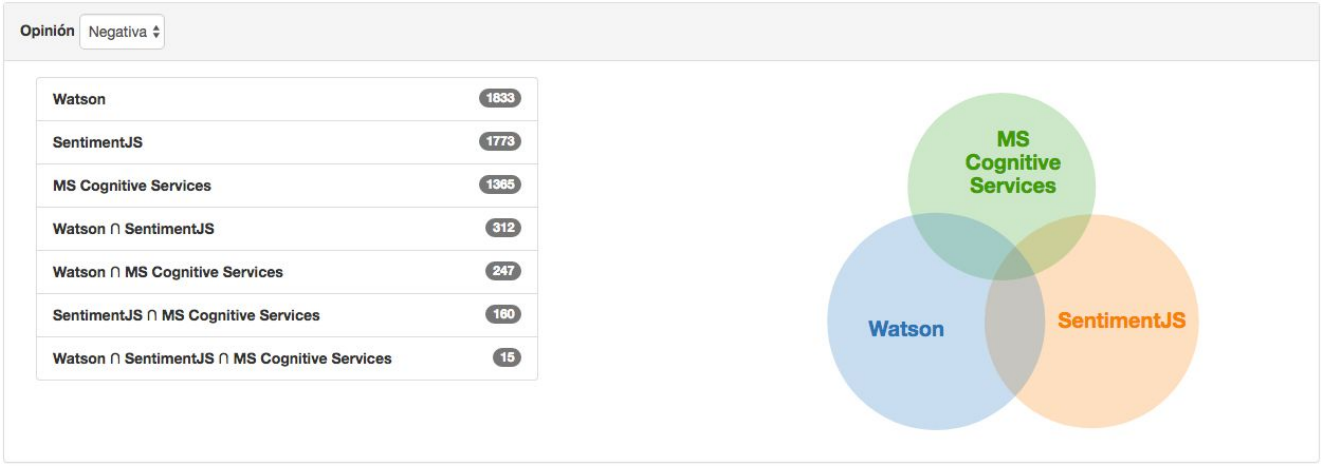

Figura 30. Diagrama de Venn con los conjuntos de comentarios negativos detectados por cada herramienta, y sus intersecciones.

En este caso, notamos que Watson detectó 1833 tweets negativos, SentimentJS reconoció 1773, y Microsoft Cognitive Services, 1365.

La intersección **Watson ∩ SentimentJS ∩ MS Cognitive Services** tiene solo 15 elementos. Siendo esta cantidad un **0.10% del total de 14399 comentarios**, podemos también decir que las herramientas tuvieron baja coincidencia respecto a los comentarios negativos.

Finalmente, tenemos la figura 31 que muestra la intersección de comentarios neutrales:

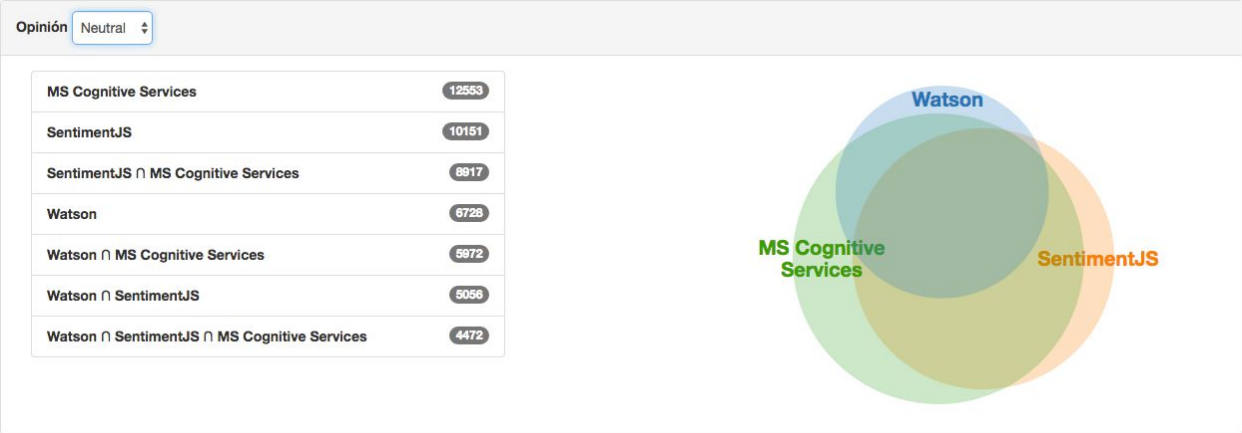

Figura 31. Diagrama de Venn con los conjuntos de comentarios neutrales detectados por cada herramienta, y sus intersecciones.

Aquí podemos observar que la herramienta que mayor cantidad de comentarios neutrales detectó fué Microsoft Cognitive Services con 12553 ocurrencias, seguido por SentimentJS con 10151. Estos números son coherentes con el análisis hecho en la sección 4.2.2, en donde observamos que el valor de opinión mayoritario era neutral, para estas dos herramientas.

Otro detalle a destacar es que el valor de opinión en donde más coincidieron las tres herramientas fué el neutral. Esto se deduce de la cantidad de elementos que tiene la intersección **Watson ∩ SentimentJS ∩ MS Cognitive Services**: 4472. Esto se traduce en un **31.06% del total de 14399 comentarios**.

4.2.4 Opinión de usuarios

A continuación, vamos a mencionar las observaciones que realizamos respecto al seguimiento de opinión de usuarios. Como se podrá recordar, implementamos un gráfico de líneas, cuyo funcionamiento y características describimos en la sección 3.4.2 de este documento. Para describir este ensayo, haremos las observaciones sobre los resultados que las 3 herramientas obtuvieron respecto a las opiniones de dos usuarios.

En un período corto de tiempo, sería de esperar que la clasificación que hagan las herramientas sobre los comentarios de una misma persona sobre un tema, mantengan el mismo valor: neutral, positivo o negativo. En nuestra gráfica de seguimiento de opinión de usuario, esto se vería como una línea recta, sin cambios entre diferentes valores de opinión. Un escenario de estas características podría sugerir que las herramientas mantuvieron el mismo criterio a la hora de catalogar los comentarios de un mismo usuario.

Tomamos dos usuarios para realizar este estudio. Uno de ellos hizo veinticinco comentarios entre los cuarenta lotes. La figura 32 muestra los resultados que Microsoft Cognitive Services obtuvo de los tweets de este usuario.

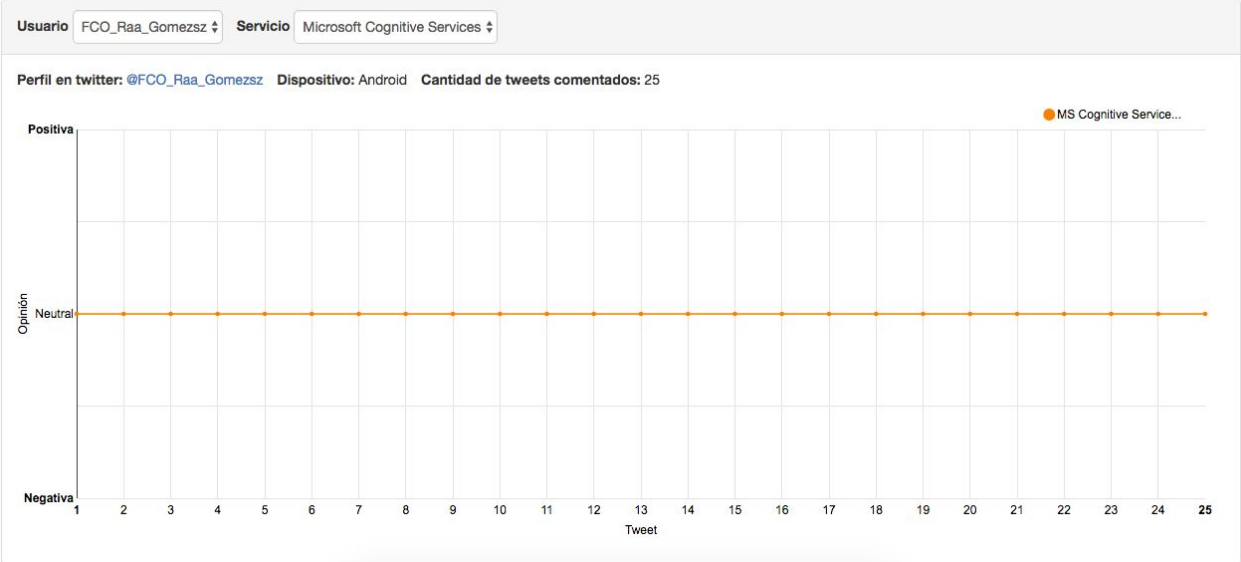

Figura 32. Evolución de la opinión según Microsoft Cognitive Services para el primer usuario estudiado.

Esta gráfica muestra que Cognitive Services detectó a los 25 comentarios como neutrales. Como dijimos, esto puede indicar que Cognitive Services mantuvo el criterio al catalogar todos los tweets de este usuario. Algo similar ocurre con la figura 33, correspondiente a los resultados de SentimentJS, en el cual, si bien existe una oscilación en los últimos cuatro valores, los veintiún restantes conservan un valor neutral. Para ambas herramientas cabe la observación de que, el hecho de que hayan clasificado a los comentarios como neutrales en su mayoría, si bien sugiere que han mantenido un mismo criterio para hacer esa clasificación, también podría evidenciar que no han podido decidir si eran positivos o negativos.

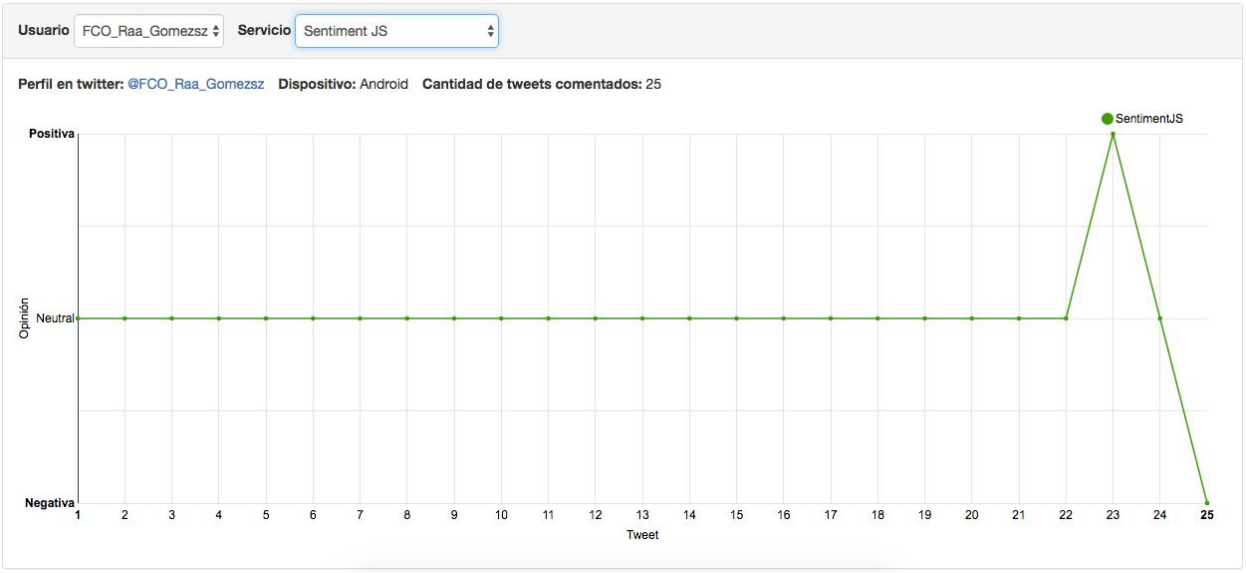

Figura 33. Evolución de la opinión según SentimentJS para el primer usuario estudiado.

Los resultados en la figura 34, correspondientes a Watson, muestran cambios entre valores neutrales y positivos para ventanas de dos o tres tweets. Esta diferencia en relación a las otras gráficas evidencia que Watson utiliza un criterio distinto para clasificar opiniones, respecto a las otras dos herramientas.

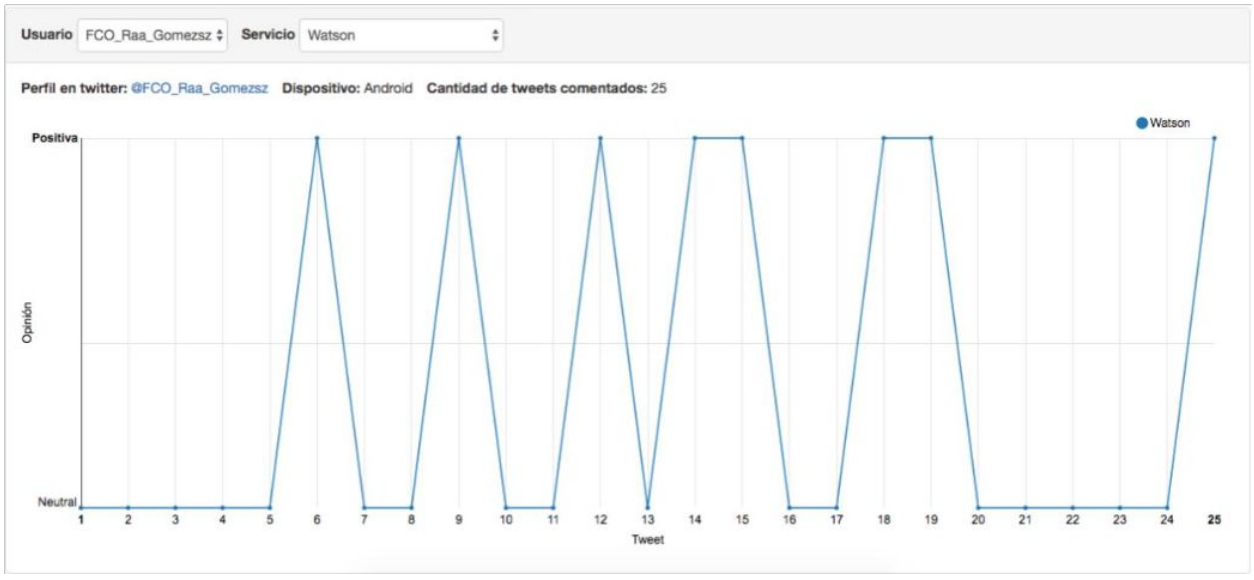

Figura 34. Evolución de la opinión según Watson para el primer usuario estudiado.

Otro usuario seleccionado para analizar publicó 21 tweets en los 40 lotes. Las figuras 35, 36 y 37 muestran los resultados que Cognitive Services, Watson, y SentimentJS obtuvieron de los tweets de dicho usuario, respectivamente.

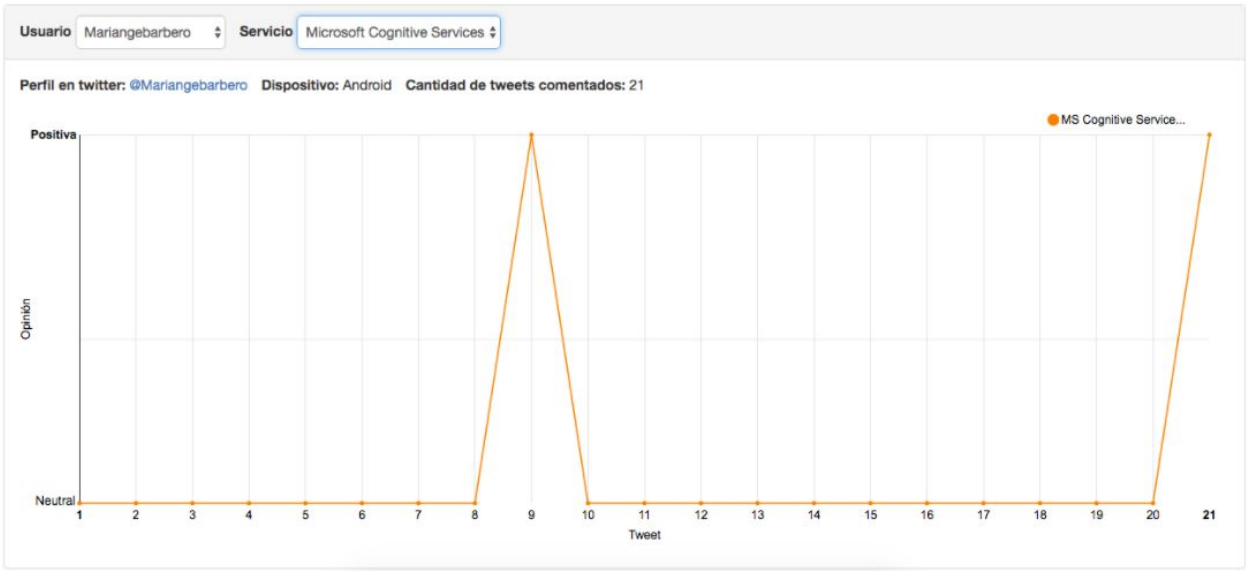

Figura 35. Evolución de la opinión según Microsoft Cognitive Services para el segundo usuario estudiado.

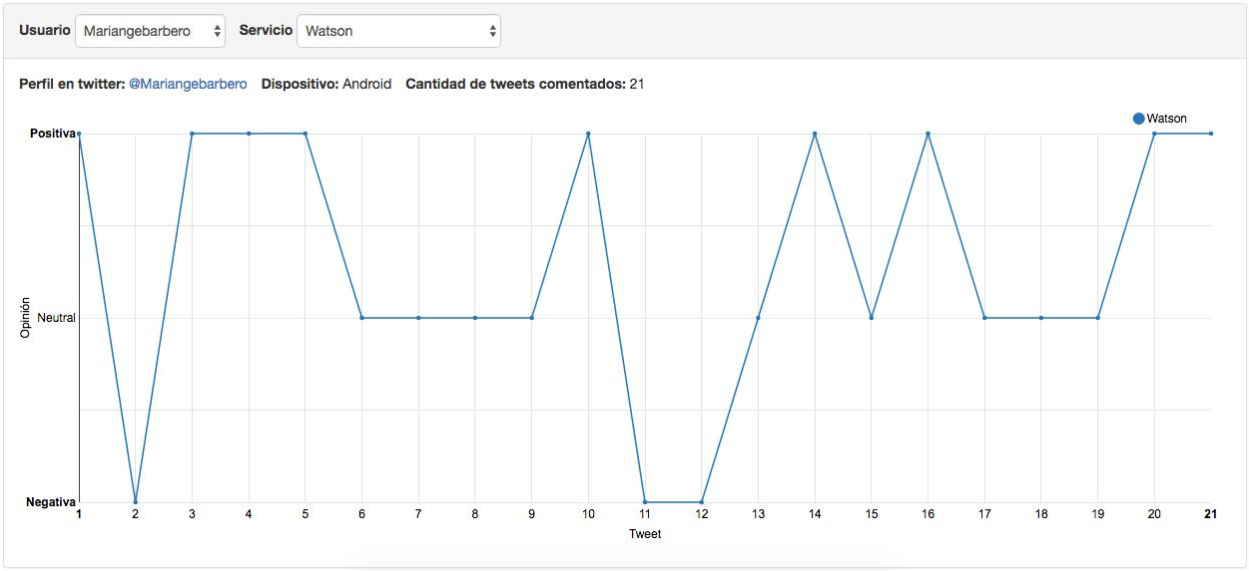

Figura 36. Evolución de la opinión según Watson para el segundo usuario estudiado.

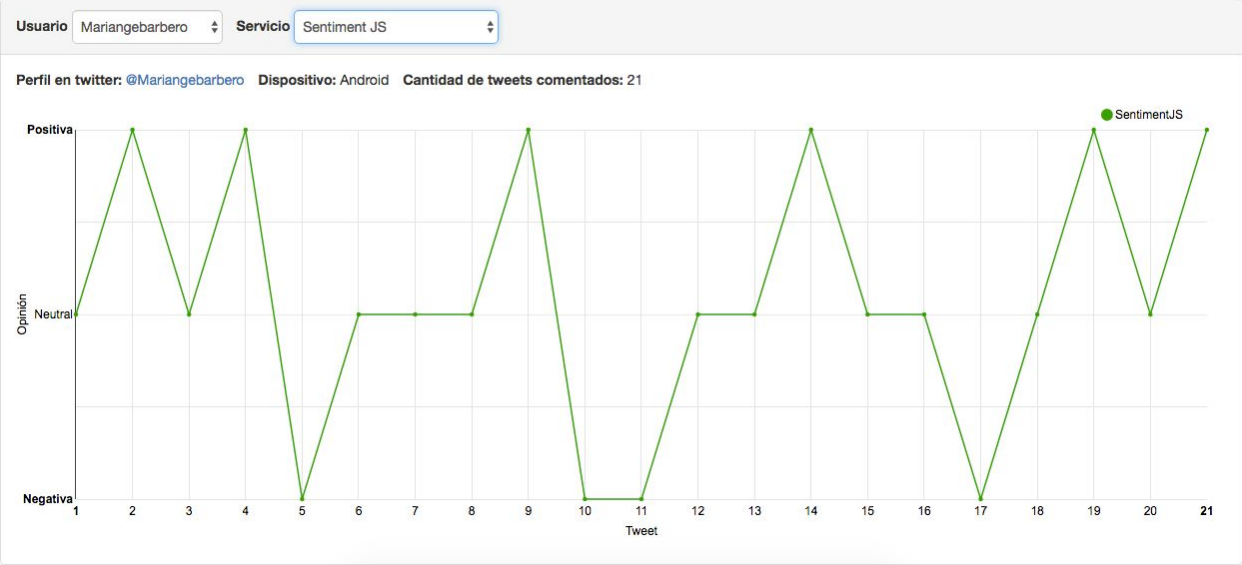

Figura 37. Evolución de la opinión según SentimentJS para el segundo usuario estudiado.

La figura 35 nos muestra cómo Cognitive Services catalogó a la mayoría de los comentarios como neutrales, mientras que apenas dos comentarios aislados fueron tomados como positivos. Este escenario es compatible con lo observado en el diagrama de Venn, y la gráfica de evolución temporal de los lotes, en donde Cognitive Services clasificó a la mayoría de los tweets como neutrales. La curva se mantiene mayoritariamente en un mismo valor de *y*, que representa el valor de opinión. Esto sugiere que la herramienta mantuvo un mismo criterio para clasificar.

Hagamos ahora foco en los resultados de Watson y SentimentJS en las figuras 36 y 37, respectivamente. Puede notarse cómo en ambos el valor de opinión oscila entre los tres valores a lo largo de los 21 comentarios. En algunos casos, el valor pasa de negativo a positivo de un comentario a otro. Cabe destacar que algunos de estos tweets se hicieron con diferencia de segundos entre uno y otro. Por ejemplo, analicemos la hora y valor de opinión según Watson y SentimentJS de los tweets del mismo usuario del día 11/05/17:

Según Watson

- Tweet número 9: Hora **21:29:01**, opinión **neutral**.
- Tweet número 10: Hora **21:32:40**, opinión **positiva**.
- Tweet número 11: Hora **21:32:56**, opinión **negativa**.
- Tweet número 12: Hora **21:48:08**, opinión **negativa**.

Según SentimentJS

- Tweet número 9: Hora **21:29:01**, opinión **positiva**.
- Tweet número 10: Hora **21:32:40**, opinión **negativa**.
- Tweet número 11: Hora **21:32:56**, opinión **negativa**.
- Tweet número 12: Hora **21:48:08**, opinión **neutral**.

Puede notarse que en un lapso de 19 minutos, Watson infirió que la opinión osciló entre los tres valores posibles. Otro aspecto notable es que entre el tweet 10 y el 11 del usuario, transcurrieron solo **16 segundos** y el valor de opinión en uno y otro son **positivo** y **negativo**.
La fluctuación de los valores podría insinuar que hubo factores que causaron que las herramientas cataloguen de manera dispar los comentarios.

4.2.5 Medidas de tendencia central y dispersión para comparar valoraciones de las librerías

A continuación, vamos a mencionar las observaciones que realizamos respecto a la tendencia central y dispersión. Como se podrá recordar, implementamos un gráfico de cajas y bigotes, cuyo funcionamiento y características describimos en la **sección 3.4.2** de este documento.

Los valores obtenidos pueden verse en las siguientes figuras:

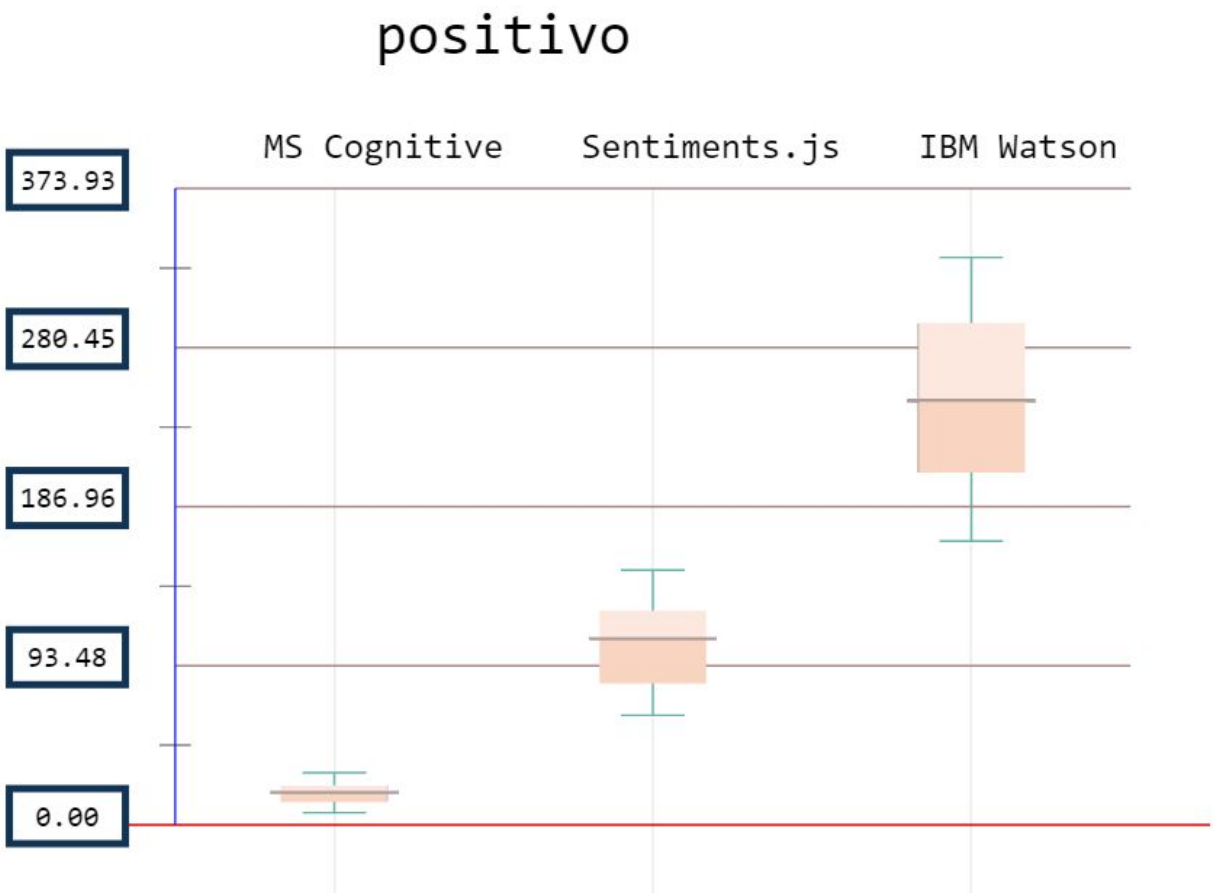

Figura 38. Diagramas de caja correspondientes a la cantidad de comentarios positivos detectados en Microsoft Cognitive Services, Sentiment JS, y Watson.

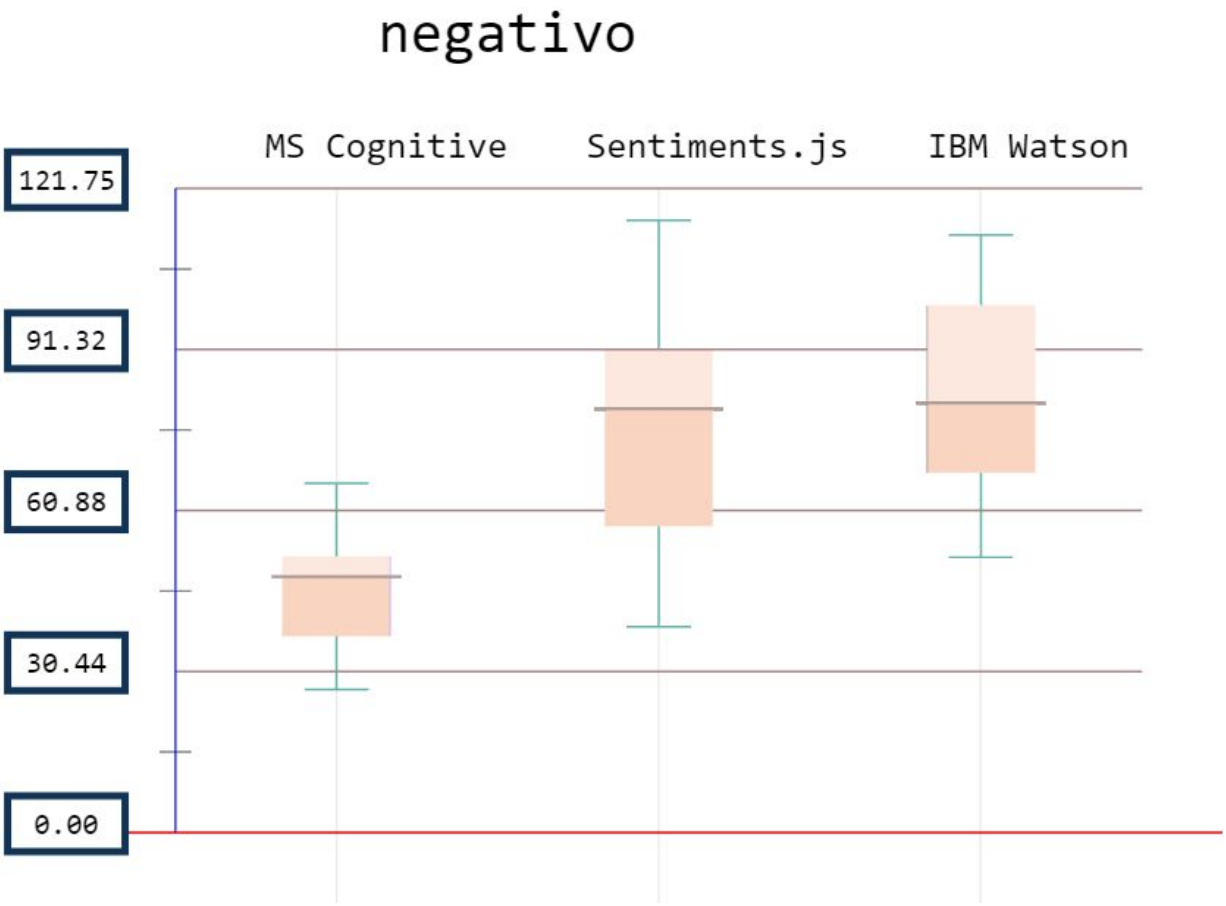

Figura 39. Diagramas de caja correspondientes a la cantidad de comentarios negativos detectados en Microsoft Cognitive Services, Sentiment JS, y Watson.

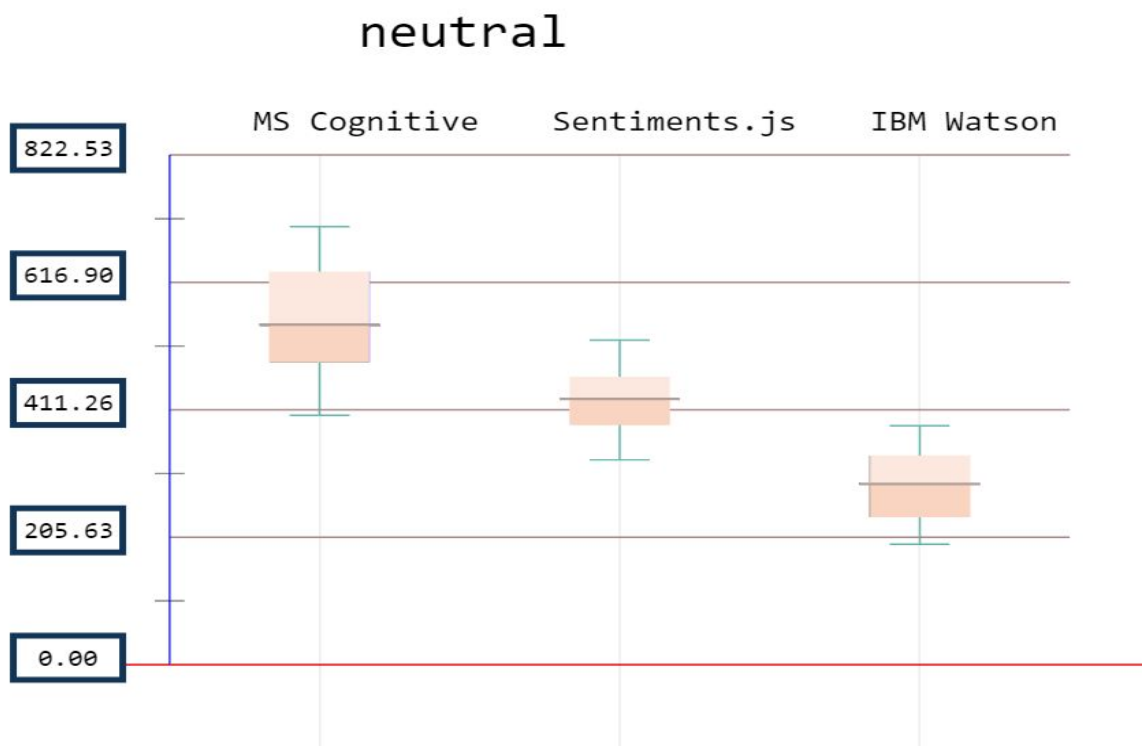

Figura 40. Diagramas de caja correspondientes a la cantidad de comentarios neutrales detectados en Microsoft Cognitive Services, Sentiment JS, y Watson.

Los datos mostrados en las figuras 40, 41, 42 también pueden visualizarse en la siguiente tabla:

clasificación	librería	**Mediana ponderada**	Desvio ponderado	1er cuartil ponderado	3er cuartil ponderado
positivo	MS Cognitive	19.324	18.805	8	27
	Sentiments.js	109.544	63.979	57	142
	IBM Watson	248.285	122.954	165	340
negativo	MS Cognitive	48.381	31.747	26	56
	Sentiments.js	80.082	60.338	36	102
	IBM Watson	81.169	45.184	55	118
neutral	MS Cognitive	548.353	234.526	428	719
	Sentiments.js	428.646	156.602	345	500
	IBM Watson	291.93	143.669	185	383

Figura 41. Tabla con valores de tendencia central y dispersión

Como podemos ver en el gráfico 40, hay una diferencia notable en cómo las librerías clasifican los positivos. La imagen permite pensar que Cognitive Services tiene menor tendencia que Watson en clasificar como positivos. Dado que la dispersión no es significativa ante esta diferencia de las medianas ponderadas, podemos presumir que habría una tendencia en Watson en clasificar positivos mayor que la de Cognitive.

Como podemos ver en el gráfico 41, para la clasificación de valores negativos, no hay diferencia notable entre las tres librerías. Todas presentan valores de medianas ponderadas, desvío y cuartiles muy similares. Dada esta situación no podríamos sacar conclusiones con respecto a tendencias de las librerías.

En el gráfico 42, para los valores neutrales, no hay diferencia notable entre las tres librerías. Todas presentan valores de medianas ponderadas, desvío y cuartiles muy similares. Dada esta situación no podríamos sacar conclusiones con respecto a tendencias de las librerías.

En ninguno de los gráficos podemos obtener algún tipo de conclusión con respecto a posibles tendencias mayores o menores de clasificación entre la librería Sentiments y Watson o Cognitive.

Las conclusiones obtenidas mediante este gráfico deberían ser soportadas por estudios estadísticos de mayor peso para dar mayor rigor a la conclusión, por el momento sólo podemos decir que se aprecian las tendencias mencionadas arriba. Aun así este tipo de conclusiones escapan de los objetivos propuestos en este trabajo, y sólo han sido agregados para enriquecer el análisis hecho sobre las librerías y sus comportamientos.

5. Conclusión

5.1 Repaso y observaciones a futuro

A lo largo de este trabajo de investigación y desarrollo, observamos que el campo del análisis automatizado de opiniones es muy reciente y tiene mucha actividad. Esto es evidente por la creciente cantidad de empresas que se dedican a este área y venden servicios relacionados, y por la bibliografía contemporánea con la que contamos en el estudio teórico de este trabajo. La necesidad de saber qué opinan los demás es importante desde hace muchos años, desde que gobiernos y negocios vieron la importancia de la retroalimentación constante, y la necesidad de no aislarse para con la población. Desde el punto de vista de los clientes y consumidores, conocer la opinión de otros en las redes también es importante, como cuando quieren conocer las experiencias y opiniones de otras personas respecto a un producto que están analizando comprar. La práctica de buscar reseñas sobre productos en Internet antes de comprarlo, se ha vuelto una práctica recurrente en muchas personas.

Debido a los usos que se le pueden dar, la minería de opiniones ha ganado importancia y ha pegado un gran salto desde sus orígenes en las Ciencias de la Computación, a otros campos impensados años atrás, como los de Administración de Empresas, el del Marketing y Publicidad, entre otros.

Durante este trabajo recorrimos desde la teoría los diferentes usos que puede darse a la Minería de Opiniones, en particular la clasificación de sentimientos, el manejo de la subjetividad, hasta también conocer las preferencias de una persona por sus expresiones. También vimos como el análisis de sentimientos puede servir como herramienta para combatir el spamming de opiniones.

El entendimiento de los problemas de la subjetividad, la intención, el contexto, las abstracciones y el sarcasmo es aún limitado, y son los principales obstáculos a sortear. A lo largo del capítulo 2 tratamos distintos autores, que mencionan estos inconvenientes y coinciden en que aún no se han desarrollado soluciones absolutas, por lo que los resultados de las implementaciones actuales tienen aspectos para mejorar. Por ejemplo, vimos como para detectar *spamming* de opiniones, se tienen en cuenta consideraciones que pueden indicar que un comentario es *spam*. Sin embargo, la detección de estas características no es suficiente, por

lo que se pueden obtener falsos positivos, y de esta forma algunos comentarios podrían ser tratados como *spam* cuando en realidad eran totalmente válidos.

Antes de comenzar con este trabajo, nos atrajo la posibilidad de utilizar tecnología capaz de comprender subjetividad e identificar el valor de opinión de texto escrito en lenguaje natural. En la medida en que fuimos investigando, vimos que no solo hay pequeñas y medianas empresas dedicadas a brindar servicios de minería de opiniones, sino que también grandes jugadores como Google, IBM y Microsoft han comenzado a ofrecer sus propios servicios de esta índole en los últimos años. El hecho de que estas grandes empresas dediquen recursos al análisis de sentimientos indica que, efectivamente, consideran que este área tiene valor. Creemos que esto es una ventaja para el área de análisis de opiniones, ya que estas empresas cuentan con recursos humanos y técnicos para lograr avances significativos. Si a esto se le suman los esfuerzos académicos, se puede esperar que en los próximos años las soluciones y los resultados de minería de opiniones peguen un salto cualitativo, respecto a cómo están hoy.

5.2 Conclusiones del ensayo y posibles trabajos futuros

En esta sección describiremos las conclusiones a las que llegamos a partir del ensayo descrito en el capítulo 4. Junto a cada conclusión, mencionaremos posibles tareas de análisis que creemos pueden tomarse en trabajos futuros para profundizar lo observado.
Al final de esta sección, incluimos un listado de tareas técnicas que pueden llevarse a cabo para extender las capacidades de Opinator.

A lo largo del capítulo anterior, hicimos observaciones de los gráficos de Opinator para el caso de estudio descrito en la sección 4.1. En la sección 4.2.2, hicimos un análisis alrededor de los valores de opinión catalogados a lo largo del tiempo por las tres herramientas integradas.

Vimos como con Cognitive Services, la mayoría de comentarios fueron detectados como neutrales en todos los lotes, mientras que para Watson, los indicadores de los tres valores de opinión variaron notablemente durante las 40 tomas o lotes, especialmente en los últimos tres días. El caso de Sentiment fué más similar al de Cognitive Services, ya que la mayoría de comentarios fueron detectados como neutrales en todos los lotes, con la salvedad de que se observaron algunos picos decrecientes para este valor de opinión.

El hecho de que Cognitive Services y Sentiment hayan catalogado a la gran mayoría de comentarios como neutrales insinúa que ambas herramientas no pudieron decidir si éstos eran positivos o negativos.

Por otro lado, el hecho de que los resultados de cada herramienta hayan sido tan disímiles entre sí, sugiere que las tres herramientas trabajan con criterios diferentes.

Además, en los lotes de los últimos tres días, se produjeron cambios en las líneas para todas las herramientas, si se las compara con los lotes de los días anteriores. Esto induce a pensar que pudieron existir factores en los comentarios de esas fechas que produjeron esos cambios en los resultados. Creemos que un posible trabajo a futuro sería estudiar el contexto en el que se dieron los comentarios en esa etapa de tiempo, para obtener una mejor comprensión del cambio de “comportamiento” de las líneas.

El análisis de la sección 4.2.3 está dedicado a las coincidencias que las tres herramientas tuvieron al inferir los valores de opinión. En este caso, se nos presenta un escenario compatible con el del análisis de la sección 4.2.2. Los bajos porcentajes en donde intersecan los resultados positivos (0.53% del total) y negativos (0.10% del total), nos sugieren que los servicios manejaron distintos criterios para decidir qué valor de opinión tiene cada comentario.

Respecto a los valores neutrales, Cognitive Services y Sentiment coincidieron en 8917 tweets, un 61.93% del total de comentarios. Esto se mantiene en la misma línea que el análisis de la sección 4.2.2, en donde sacamos que el valor de opinión mayoritario fue el neutral, para estas tres herramientas. Creemos que esto es un indicio de la incapacidad de las herramientas en catalogar entre positivo o negativo, especialmente en los casos de Sentiment y Cognitive Services.

El análisis de la sección 4.2.4 se enfoca en el seguimiento de opiniones de usuarios. En uno de los ensayos, obtuvimos distintos valores de opinión de comentarios de un mismo usuario con diferencia de apenas segundos. Los cambios en los valores que obtuvimos podrían insinuar que hubieron factores que afectaron al criterio de las herramientas. Creemos que para entender mejor qué fué lo que produjo esto, debería hacerse un análisis de caja blanca de las tres herramientas, para conocer sus estrategias. Lamentablemente, Watson y Cognitive Services son productos cerrados, por lo que en este trabajo nos limitamos a hacer un estudio de caja negra.

Sin embargo pensamos que, si más adelante, las empresas detrás de estos servicios dan a conocer los criterios y/o algoritmos utilizados, se puede afrontar a este análisis como tarea a futuro.

Por otro lado, al comparar los resultados de cada herramienta para los tweets de un usuario, vimos que las gráficas eran muy distintas entre sí. Creemos que esto evidencia que las tres herramientas tienen diferentes criterios o estrategias al momento de catalogar un tweet. Estas diferencias introducen la necesidad de poder reunir los resultados y mostrarlos de una manera que facilite su análisis. Aquí es donde Opinator agrega valor, ya que cubre este requisito.

Con el gráfico de cajas y bigotes pudimos observar que las librerías parecen tener cierta tendencia al momento de clasificar. En particular pudimos observar cierta tendencia en Watson a clasificación más positiva que el resto de las librerías, y en particular más que Cognitive. Por otro lado pudimos observar que estas dos librerías también tiene tendencias dispares en clasificar neutrales: Cognitive podría tener tendencias más altas de clasificar por neutrales si lo comparamos con el resto de las librerías, en particular con Watson. No pudimos sacar algún tipo de conclusión al momento de analizar cómo era la dispersión y el valor central en neutrales, dado que las tres librerías tuvieron tendencias similares (aunque si mirásemos este gráfico conjunto con el diagrama de Venn podríamos observar que estas clasificaciones generan conjuntos disjuntos, lo que permite pensar que es una coincidencia que las tendencias de neutrales sean parecidas, sino iguales).

Hay que mencionar, no obstante, que estas conclusiones están basadas solamente en la observación de los gráficos y los valores dispuestos en las tablas. Sería necesario realizar algún análisis estadístico que permita sacar conclusiones más definitivas para poder realizar algún tiempo de conclusión pesada al respecto de tendencias en las librerías. Quedará para un trabajo futuro realizar este trabajo estadístico.

Uno de los principales objetivos de este trabajo de tesina fué el de implementar una herramienta que sirva para analizar los sentimientos respecto a un tema, a través de minería de opiniones. Pensamos a esta herramienta como un potencial producto. Un caso de uso que habíamos pensado era que pudiera ser utilizada por la gerencia de una empresa para conocer qué opina una población sobre ella. Este conocimiento podría servir como apoyo para tomar decisiones, respecto a sus productos o su publicidad, por ejemplo.

En nuestro sistema integramos distintas herramientas y comparamos sus resultados. Los distintos análisis que realizamos en el capítulo 4 sugieren que las tres herramientas manejan distintas formas de detectar opinión. Esto nos lleva a pensar que la elección de un servicio de análisis de opinión puede ser determinante en los resultados que le llegan al usuario, y por tanto a las decisiones que pueda tomar en función de éstos.

A continuación listamos las posibles tareas que creemos que pueden realizarse en un futuro para mejorar y extender las capacidades de Opinator. Cabe destacar que en el desarrollo intentamos aplicar buenas prácticas de programación para que desarrollos venideros tengan una curva de aprendizaje poco pronunciada.

Las tareas técnicas son:

- **Manejo de usuarios y guardado de resultados.** Al ser Opinator un prototipo para hacer experimentos, no consideramos inicialmente el manejo de distintos usuarios en la misma. Sin embargo, creemos que sería una característica importante, ya que le podría permitir al usuario guardar los resultados que va obteniendo, para cargarlos y estudiarlos más adelante. Cada usuario podría acceder a su propio conjunto de resultados y visualizarlos.

- **Más herramientas de análisis de opinión.** Opinator integra a Watson de IBM, Cognitive Services de Microsoft y a Sentiment para procesar opiniones. Diseñamos a nuestra aplicación para que sea sencillo añadir nuevos servicios, en el caso que sea necesario, y alentamos a que en futuros desarrollos sea así. Esto hará a una mayor riqueza en los resultados que puedan obtenerse y en las comparaciones.

- **Selección de herramientas de análisis de opinión.** Hoy Opinator está pensado para usar un conjunto fijo de herramientas de análisis de opinión. En trabajos futuros se podría permitir al usuario elegir qué herramientas usar antes de iniciar el proceso de recolección de comentarios. Esto podría permitir usar una sola herramienta en vez de todas las disponibles, por ejemplo.

- **Uso de otras herramientas para caracterizar a la población analizada.** Como se mencionó ya en otras secciones, recurrimos a una librería para inferir el género de cada persona cuyo tweet recibimos desde Twitter. Si bien hemos tenido resultados con este enfoque, en muchos casos la librería no logra determinar el sexo, por lo que en los gráficos tenemos números relativamente grandes con género desconocido.
  Pensamos que podrían integrarse otras herramientas para reemplazar a las que nosotros incluimos. También creemos conveniente que en el futuro se extraigan otros datos de usuario. Estos datos podrían abrir puertas a nuevos gráficos y filtros, como por ejemplo un "mapa de calor" mostrando de qué lugares del mundo provienen comentarios con mayor frecuencia.

- **Dar soporte a otros idiomas.** En este desarrollo acotamos los tweets a analizar, centrándonos sólo en aquellos cuyo idioma sea el Español. Consideramos que si se puede asegurar que todas las estrategias pueden procesar otro idioma, podría extenderse la aplicación Opinator para que el usuario pueda configurar el idioma de los tweets que quiere analizar.

5.3 Palabras finales

A lo largo de este trabajo, hemos investigado el estado del arte de la minería de opiniones y el análisis de sentimientos aplicados a redes sociales. Vimos que la mayoría de implementaciones se inclina por el uso de *Machine Learning* y *NLP*. Para entender por qué ambas metodologías son tan populares, hicimos un repaso de ambas, detallando los pasos que las componen y los desafíos con los que se enfrentan hoy en día para obtener buenos resultados.

El desarrollo de Opinator nos fué útil por distintos motivos. El principal, es que sirvió como herramienta para comparar los resultados de los distintos servicios y para caracterizar a los usuarios autores de los tweets. Sin embargo, hay otros motivos. Inicialmente teníamos una idea de qué queríamos hacer con Opinator, pero en la medida en que fuimos investigando qué recursos queríamos usar, y qué resultados deseábamos obtener, tuvimos que replantear y rediseñar la arquitectura y funcionamiento de Opinator en varias ocasiones. El resultado es la arquitectura Web cuyo funcionamiento se describe en en la sección 3.4.1.

Otro motivo por el que hallamos útil a este desarrollo, es que nos sirvió para profundizar nuestra experiencia con Javascript, Node, MongoDB y Angular. Desde un punto de vista personal, estas tecnologías nos resultaban muy interesantes para hacer desarrollo Web. El desarrollo de Opinator fué una buena oportunidad para comenzar un proyecto de cero que las integre a todas.

Escribiendo estas palabras finales, observamos la bibliografía y notamos que la mayoría de artículos y escritos referenciados son de mediados del 2000 en adelante, si bien hay trabajos de otras décadas. Claramente, el fenómeno de querer conocer qué opina la gente en Internet es algo relativamente contemporáneo. Creemos que esto se debe a que por esos años, el uso de Internet se masificó lo suficiente con la popularidad de las redes sociales.

Tengamos en cuenta que en ese entonces, el acceso a las redes sociales era solo a través de un Navegador Web. Ya con el *input* que se podía generar sólo por ese medio, las empresas y organizaciones tenían a mano mucha información de donde extraer valor para comprender al mercado. Hoy existen otras vías para acceder a las redes sociales, como las aplicaciones móviles nativas, permitiendo estar conectado en casi todo el tiempo del día que uno quiera, mientras el usuario lleve consigo un *smartphone*. Las formas en que un usuario puede generar *input* con estos dispositivos son aún más; no sólo tiene fácil acceso a la Red para escribir lo que quiera, lo que siente o lo que pasa *en ese mismo momento*, sino que también puede aportar datos de geolocalización, puede subir imágenes y video, enviar audio, entre otras cosas. Las cantidad de entrada de datos que tienen las redes sociales es mayor que en sus orígenes. Esto amplía las posibilidades de caracterizar a sus usuarios, y de elaborar perfiles mucho más detallados que antes, si se “cruzan” todos estos datos de manera correcta y eficiente. Desde el punto de vista de los usuarios, los beneficios podrían traducirse en una mejor experiencia general. Si sitios como Facebook, Youtube o Twitter lograran dominar el análisis de texto de manera que evitaran el spam por completo, se lograría una gran avance en pos de la experiencia de usuario. Además, podrían elaborarse los contenidos a medida de los gustos e intereses de cada usuario. Esto, si bien ya lo hacen la mayoría de redes sociales, creemos que tienen aún lugar para mejorar.

Bibliografía

[1] **Charu Aggarwal, ChengXiang Zhai** (2012)**.** *Mining text data.* Springer, Nueva York.

[2] **Kushal Dave, Steve Lawrence, David Pennock** (2003). *Mining the peanut gallery: Opinion extraction and semantic classification. WWW '03 Proceedings of the 12th international conference on World Wide Web.* Páginas 519-528.

[3] **R. Kibble** (2013) *Introduction to Natural Language Processing.* University of London, Londres.

[4] **R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa**. (2011) *Natural Language Processing (Almost) from Scratch. The Journal of Machine Learning Research archive.* Volumen 12, 2/1/2011. Páginas 2493-2537

[5] **Erich Gamma, Richard Helm, Ralph Johnson and John Vlissides** (1994). *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison-Wesley, Boston, Massachusetts.

[6] **Bing Liu** (2012). *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers, San Rafael, California.

[7] **E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne.** (2008) Finding high-quality content in social media. *WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining. Palo Alto, California.* Páginas 183-194.

[8] **W. Cohen, H. Hirsh**, 1998. Joins that generalize: text classification using Whirl. *KDD' 98 Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining.* Nueva York. Páginas 160-173.

[9] **Boyd-Graber, J. and P. Resnik** (2008)**.** Holistic sentiment analysis across languages: multilingual supervised latent dirichlet allocation**.** *EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* Páginas 45-55. Cambridge, Massachusetts.

[10] **B. Sigurbjörnsson and R. Van Zwol**, 2008. *Flickr tag recommendation based on collective knowledge*. *WWW '08 Proceedings of the 17th international conference on World Wide Web.* Páginas 327-336. Beijing, China.

[11] **Steven Pinker** (2003). *The Language Instinct: How the Mind Creates Language.* Penguin Books, Londres.

[12] **Ralf Herbrich, Thore Graepel** (2010). *Handbook of Natural Language Processing*. Chapman & Hall, Londres.

[13] UIT (2016) http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf

[14] Library of congress (2010). http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/

[15] Obar, Wildman (2015). *Social Media Definition and the Governance Challenge: An Introduction to the Special Issue*. Telecommunications Policy, 39(9). Páginas 745-750.

[16] Sonny Ganguly (2015). *Why Social Media Advertising Is Set To Explode In The Next 3 Years* http://marketingland.com/social-media-advertising-set-explode-next-3-years-121691

[17] Ekaterina Stepanova (2011). *The Role of Information Communication Technologies in the "Arab Spring". PONARS Eurasia Policy Memo No. 159*.

[18] James Lewin (2008). Is Social Media Behind Barack Obama's Success? http://www.podcastingnews.com/content/2008/06/is-social-media-behind-barack-obamas-success/

[19] Hernán Gustavo Mirand (2015) El uso de las redes sociales en la campaña presidencial argentina del año 2015. http://www.unsta.edu.ar/wp-content/uploads/2016/05/Hern%C3%A1n-Miranda-El-uso-de-las-redes-sociales-en-la-campa%C3%B1a-presidencial-argentina-del-a%C3%B1o-2015.pdf

[20] Amy Mitchell, Jeffrey Gottfried, Michael Barthel y Elisa Shearer (2016). *The Modern News Consumer*. http://www.journalism.org/2016/07/07/the-modern-news-consumer

[21] Twiplomacy (2016) *World Leaders on Instagram*. http://twiplomacy.com/blog/world-leaders-on-instagram-2016/

[22] Cooper Smith (2014). *Social Big Data: The User Data Collected By Each Of The World's Largest Social Networks — And What It Means* http://www.businessinsider.com/social-big-data-the-type-of-data-collected-by-social-networks-2-2014-1

[23] **Burton, R.** (1977). *Semantic grammar: An engineering technique for constructing natural language understanding systems*. ACM SIGART, Issue 61, February 1977. Página 26. Nueva York.

[24] **Grishman, R., N. T. Nhan, E. Marsh, y L. Hirschman** (1984). *Automated determination of sublanguage syntactic usage*. *ACL '84 Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*. Páginas 96-100. Stanford, California

[25] **Aljoscha Burchardt, Stephan Walter, Alexander Koller, Michael Kohlhase, Patrick Blackburn y Johan Bos**, 2003. *Computational Semantics* http://www.coli.uni-saarland.de/projects/milca/courses/comsem/html/index.html

[26] **Gretchen McCulloch**, 2014. *Scopal Ambiguity: Messing With Words to Make Things Funny* http://www.quickanddirtytips.com/education/grammar/scopal-ambiguity-messing-with-words-to-make-things-funny

[27] **William Hughes, Jonathan Lavery**, 2004. *Critical Thinking: An Introduction to the Basic Skills.* Broadview Press, Canada.

[28] **Schalley, A. C.** 2004. *Cognitive Modeling and Verbal Semantics. A Representational Framework Based on UML. Walter de Gruyter, Berlín.*

[29] **Schalley, A. C,** 2004. *Representing verbal semantics with diagrams. An adaptation of the UML for lexical semantics. COLING '04 Proceedings of the 20th international conference on Computational Linguistics. Artículo No. 785. Ginebra, Suiza.*

[30] Abushihab, I. 2015. *A Pragmatic Stylistic Framework for Text Analysis. International Journal of Education ISSN 1948-5476, Vol. 7, No. 1.*

[31] **Z. Zhang and B. Varadarajan** (2006). *Utility scoring of product reviews. CIKM '06 Proceedings of the 15th ACM international conference on Information and knowledge management. Páginas 51-57. Arlington, Virginia.*

[32] **Jeremy Diamond,** 2016. *Russian hacking and the 2016 election: What you need to know.* http://edition.cnn.com/2016/12/12/politics/russian-hack-donald-trump-2016-election/

[33] **N. Jindal y B. Liu**, 2007. *Review spam detection*. *WWW '07 Proceedings of the 16th international conference on World Wide Web*. Páginas 1189-1190. Banff, Alberta, Canada.

[34] **N. Jindal y B. Liu**, 2008. *Opinion spam and analysis*. *WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining.* Páginas 219-230. Palo Alto, California.

[35] **Twitter, Inc**, 2017. *Streaming APIs.* https://dev.twitter.com/streaming/overview

[36] **Node.js Foundation**, 2017. *About*. https://nodejs.org/en/about/

[37] **StrongLoop, IBM, and other expressjs.com contributors**, 2017. *Express.* http://expressjs.com

[38] **StrongLoop, IBM, and other expressjs.com contributors**, 2017. *Using Express Middleware.* http://expressjs.com/en/guide/using-middleware.html

[39] **Mike Bostock**, 2017. *Data-Driven Documents.* https://d3js.org/

[40] **Ricardo Cabello**, 2017 *three.js - Javascript 3D library.* https://threejs.org/

[41] **Graham Upton and Ian Cook,** 2008. *A Dictionary of Statistics (2 rev. ed) ISBN-13: 9780199541454.*

[42] **David Terr**. Weighted Mean. http://mathworld.wolfram.com/WeightedMean.html

[43] **DATAPLOT**, Weighted standard deviation,
http://www.itl.nist.gov/div898/software/dataplot/refman2/ch2/weightsd.pdf

[44] **Rick Wicklin**, weighted percentiles
http://blogs.sas.com/content/iml/2016/08/29/weighted-percentiles.html