PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

INTELIGENCIA COMPETITIVA DE PROMOCIONES APLICANDO ONTOLOGÍAS DE DOMINIO EN FACEBOOK DE EMPRESAS DE TELECOMUNICACIONES DEL PERÚ

TESIS PARA OPTAR EL GRADO ACADÉMICO DE MAGISTER EN CIENCIAS DE LA COMPUTACIÓN

AUTOR

GERALDO COLCHADO RUIZ

ASESOR:

HÉCTOR ANDRÉS MELGAR SASIETA

Abril, 2018

RESUMEN

El mercado de telecomunicaciones en el Perú es muy competitivo y es uno de los sectores económicos que más crecimiento tuvo en los últimos años lo que se refleja en que actualmente existan más celulares que población.

Las 5 empresas de telecomunicaciones (Operadores) ofrecen sus promociones en redes sociales, principalmente en Facebook, para mantener a sus clientes existentes y obtener nuevos clientes. Hay una gran cantidad de datos en Facebook escrita en lenguaje natural sin significado para el computador que los operadores no están utilizando para tener Inteligencia Competitiva. La Inteligencia Competitiva es un proceso que identifica las necesidades de información de los tomadores de decisiones acerca de los competidores, recolecta datos de fuentes públicas y legales, les da significado o semántica y los analiza para dar respuesta a las necesidades de información comunicando los resultados a los tomadores de decisiones.

En esta tesis se propone e implementa un proceso de Inteligencia Competitiva de promociones para los operadores que incluye la recolección de 15,634 *posts* y 1,411,698 comentarios de Facebook como corpus, el proceso de creación manual de una ontología de dominio en telecomunicaciones con 119 palabras, 27 conceptos y 6 relaciones en 5 niveles jerárquicos, la clasificación de los *posts* usando la ontología de telecomunicaciones, el proceso de creación semiautomático de una ontología de dominio en polaridad a partir de *WordNet* en español y *SentiWordNet* con 9,344 palabras, el análisis de polaridad o clasificación de polaridad positiva, negativa o neutra de cada comentario, la implementación de una aplicación web para que los tomadores de decisiones puedan realizar búsquedas de *posts* basadas en la ontología de telecomunicaciones y responder a sus necesidades de información o preguntas relevantes y la implementación de una aplicación web que compara los resultados de los *posts* por operador en un formato de línea de tiempo incluyendo comentarios positivos y negativos logrando la Inteligencia Competitiva.

El proceso de Inteligencia Competitiva como el proceso de creación de la ontología de dominio en telecomunicaciones podrían ser aplicados en empresas de telecomunicaciones de otros países y también en otros contextos donde existan varios competidores que ofrezcan productos o servicios equivalentes que puedan compararse. El proceso de creación de ontología en polaridad puede ser replicado por otros investigadores para análisis de polaridad en otros idiomas distintos al inglés y español dada la disponibilidad de *WordNet* en varios idiomas.

Reconocimiento: Un *paper* del mismo autor y relacionado a la presente tesis fue aceptado y presentado en el *"International Conference on Information Technology & Systems (ICITS 2018)"* y fue publicado en el libro *"Advances in Intelligent Systems and Computing, vol 721. Springer, Cham"* en Enero del 2018 con el siguiente título *"Competitive Intelligence Using Domain Ontologies on Facebook of Telecommunications Companies of Peru"* con DOI https://doi.org/10.1007/978-3-319-73450-7_73.

Palabras clave: Inteligencia Competitiva, Ontología, Tomador de Decisiones, Redes Sociales, Facebook, Empresas de Telecomunicaciones, Análisis de Polaridad, *WordNet*, *SentiWordNet*, Procesamiento de Lenguaje Natural.

Índice General

Índice de Figuras

Índice de Tablas

1. PROBLEMÁTICA Y MARCO CONCEPTUAL

1.1. Problema

El mercado de telecomunicaciones en el Perú es muy competitivo, ya que a fines del 2016, según el Organismo Supervisor de Inversión Privada en Telecomunicaciones (OSIPTEL)[1], se tenían activas aproximadamente 37.7 millones de líneas celulares sobre una población en Perú estimada de 31.5 millones según INEI [2]; es decir, hay más de 1 celular por habitante en el Perú. Respecto a participación de mercado, según OSIPTEL [3], a fines del 2016 existen 5 operadores de telecomunicaciones que son Movistar con 44.51%, Claro con 32.64%, Entel con 12.85%, Bitel con 9.85% y Virgin con 0.15%, quienes constantemente están ofreciendo promociones para captar nuevos clientes y mantener y fidelizar a sus clientes existentes.

Formalmente, los operadores publican en el sistema de consulta de tarifas SIRT [4] de OSIPTEL la información detallada de sus planes tarifarios y promociones tarifarias, el acceso a este sistema es público y pueden consultarlo tanto los operadores como cualquier persona. Adicionalmente, los operadores publican en sus páginas web de forma más resumida sus planes tarifarios y promociones tarifarias. En algunos casos, dependiendo de la importancia y el presupuesto de marketing que tengan, los operadores publican sus promociones en medios masivos tradicionales como televisión, radio, diarios y revistas. En todos los casos anteriores no es posible conocer cuántas personas han visto la información publicada y cuáles son sus opiniones. Además, pueden pasar varias horas o hasta días desde la publicación hasta que la persona lo vea.

En el 2009, con la llegada al Perú de la tecnología celular 3G y los *smartphones* [5], los operadores progresivamente empezaron a publicar sus promociones en redes sociales. En el 2014, con la llegada del 4G y mayores velocidades en datos, creció el número de personas con acceso a redes sociales desde los *smartphones* y que empezaron a seguir a los operadores. Facebook es una de las redes sociales más utilizadas por los operadores como medio masivo para publicar sus promociones y que éstas puedan ser vistas en corto tiempo por sus seguidores. Según sus páginas de Facebook, a la fecha de escrita esta tesis, Movistar [6] y Claro [7] cuentan con más de 4 millones de seguidores, mientras que Entel [8] y Bitel [9] tienen más de 1 millón de seguidores y Virgin [10] tiene más de 100,000 seguidores en Facebook.

La publicación de promociones por los 5 operadores y los comentarios de los seguidores generan una gran cantidad de datos en Facebook que los operadores no están aprovechando para ver cómo le está yendo a la competencia, compararse con ellos y ser más competitivos. Está práctica, denominada inteligencia competitiva, es realizada principalmente en el sector empresarial y diversas disciplinas como administración y dirección de empresas, marketing, la ciencia de la información y la

[1] https://www.osiptel.gob.pe/articulo/24-lineas-en-servicio-por-empresa

[2] https://www.inei.gob.pe/media/MenuRecursivo/Cap03020.xls

[3] https://www.osiptel.gob.pe/articulo/24-lineas-en-servicio-por-empresa

[4] http://serviciosenlinea.osiptel.gob.pe/ConsultaSIRT

[5] Celulares inteligentes con acceso a servicio de datos

[6] https://www.facebook.com/pg/movistarperu/community/?ref=page_internal

[7] https://www.facebook.com/pg/AmericaMovilPeruSAC/community/?ref=page_internal

[8] https://www.facebook.com/pg/EntelPeru/community/?ref=page_internal

[9] https://www.facebook.com/pg/bitelperu/community/?ref=page_internal

[10] https://www.facebook.com/pg/inkacelperu/community/?ref=page_internal anteriormente se llamaba Virgin Mobile Peru pero fue vendida en Setiembre del 2017 a Inkacel (Fuente: https://elcomercio.pe/economia/mercados/virgin-mobile-vendio-operaciones-inkacel-noticia-456278)

ingeniería informática han contribuido a impulsarla (García Alsina y Ortoll Espinet 2012). La inteligencia competitiva es un proceso dinámico, sistemático y recursivo que transforma, empleando técnicas analíticas específicas, la información relevante y legalmente obtenida sobre el entorno competitivo del pasado, presente y futuro, con el propósito de facilitar la toma de decisiones en beneficio de la empresa (Gógova 2015).

Los datos publicados en Facebook por los operadores y los comentarios de sus seguidores son de dominio público, ya que los mismos operadores cuando hacer una publicación eligen que sea pública, es decir que cualquier persona que tenga acceso a Facebook la pueda ver [11] y por consiguiente es legal su uso en la práctica de inteligencia competitiva.

Facebook provee interfaces gratuitas en protocolo HTTP que pueden utilizarse en un lenguaje de programación para extraer estos datos, sin embargo, el problema principal es que los textos están en lenguaje natural, lo que hace difícil para un computador entender su significado o semántica. Ver Figura 1 de una publicación real de un operador con texto escrito en lenguaje natural que las personas pueden entender su significado, pero no un computador.

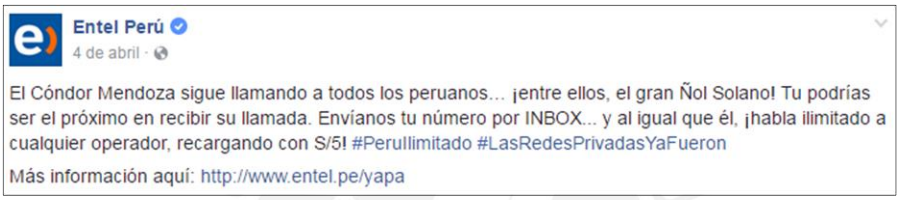


Figura 1: Ejemplo de publicación o *post* con texto en lenguaje natural

Los motores de búsqueda como Google [12] permiten buscar por palabras completas por lo que para ubicar el *post* de la Figura 1 se podría buscar por cualquiera de las palabras que contiene como por ejemplo la palabra "llamada". Sin embargo, esto no es muy relevante para un operador.

Para un operador es relevante saber, por ejemplo, qué promociones hay respecto al servicio de voz que ofrecen los competidores. También le es relevante saber la tarifa que se está cobrando por el servicio, las características del servicio ofrecido y quienes pueden acceder al servicio. Lo cual se expresa en la pregunta ejemplo de Tabla 1.

Tabla 1: Ejemplo de pregunta relevante para operador

Pregunta
¿Qué promociones de servicio de voz ilimitado a cualquier operador se ofrecen para prepago?

Si se le pidiera al computador responder la pregunta de la Tabla 1 tomando como ejemplo la publicación de Figura 1, aún si el computador pudiera buscar publicaciones por palabras no podría responderla por completo porque no se trata sólo de buscar palabras sino de buscar conceptos

[11] https://www.facebook.com/help/120939471321735?helpref=related&ref=related

[12] https://www.google.com.pe

relevantes para el operador en las publicaciones. Un concepto puede estar identificado por varias palabras e incluso expresiones [13] que son un conjunto de palabras sujetas a alguna pauta.

Según la Real Academia Española un concepto [14] es una representación mental asociada a un significante lingüístico, esta representación mental depende del contexto, en el contexto de telecomunicaciones una palabra podría tener un significado distinto que en otro contexto. Por ejemplo, la palabra "operador" significa que es una empresa de telecomunicaciones y en una fábrica podría significar que es una persona que opera una máquina, para ello es importante darles semántica a las palabras dependiendo del contexto.

La semántica [15] es la disciplina que estudia el significado de las unidades lingüísticas y de sus combinaciones. Esta semántica el computador no la tiene a menos que una persona le proporcione los conceptos y las palabras que los identifican para que pueda entender el texto escrito en lenguaje natural.

En la pregunta ejemplo de Tabla 1 se subrayan cuatro conceptos los cuales pueden ser identificados por su significado o semántica con las palabras indicadas en la Tabla 2:

Tabla 2: Ejemplo de conceptos y palabras que los identifican

Conceptos	**Palabras que identifican al concepto**
Servicio: Voz	llama, habla
Tarifa del Servicio: Cero	ilimitado
Operador: Movistar, Claro, Entel, Bitel, Virgin	"cualquier operador"
Plan Tarifario: Prepago	recarga
Aplicación a texto de Figura 1: El Cóndor Mendoza sigue llamando a todos los peruanos… ¡entre ellos, el gran Ñol Solano! Tu podrías ser el próximo en recibir su llamada. Envíanos tu número por INBOX... y al igual que él, ¡habla ilimitado a cualquier operador, recargando con S/5! #PerúIlimitado #LasRedesPrivadasYaFueron Más información aquí: http://www.entel.pe/yapa	

Para que un computador pueda entender el significado o semántica de las publicaciones o *posts* de promociones que hacen los operadores en Facebook, primero hay que definir los conceptos en el contexto de telecomunicaciones que son relevantes para el operador como los servicios, tarifas, operadores y planes tarifarios indicados en el ejemplo de Tabla 2, luego una lista de palabras o términos que los identifican. Lo cual puede ser representado mediante una ontología, en este caso, en el dominio o contexto de las telecomunicaciones. Las ontologías proporcionan vocabularios de palabras y términos sobre conceptos dentro de un dominio y sus relaciones, acerca de las actividades que tienen lugar en ese dominio, y sobre las teorías y principios elementales que rigen ese dominio (Gómez-Pérez et al. 2003).

Al operador le es relevante saber también qué opinaron los seguidores respecto a la publicación o *post* y responder por ejemplo a la pregunta de Tabla 3 que complementa a pregunta de Tabla 1:

[13] dle.rae.es/?w=expresión

[14] dle.rae.es/?w=concepto

[15] dle.rae.es/?w=semántica

Tabla 3: Ejemplo de pregunta relevante para operador con opinión

Pregunta
¿Qué promociones de servicio de voz ilimitado a cualquier operador se ofrecen para prepago y qué opinan (+/-) sus seguidores?

El texto de los comentarios u opiniones acerca de un *post* que escriben los seguidores en Facebook, a pesar de que está en lenguaje natural sin sentido para el computador, puede ser utilizado para determinar el sentimiento positivo o negativo del seguidor y responder a la pregunta de ejemplo de Tabla 3. De igual manera, para que el computador entienda el significado positivo o negativo del comentario podrían definirse palabras o términos que identifiquen la positividad (+) o negatividad (-) del comentario, los cuales pueden ser representados mediante una ontología de dominio en polaridad.

Según la Real Academia Española el sentimiento [16] es el estado afectivo del ánimo, el estado de ánimo [17] es la disposición en que se encuentra alguien, causada por la alegría, la tristeza, etc. Esta disposición de la persona de estar alegre o triste con el servicio que recibe del operador influye en la polaridad de los comentarios que realiza, si está alegre lo más probable es que su comentario sea de polaridad positiva o a favor del operador y si está triste sea de polaridad negativa o en contra del operador.

En la Tabla 4 se muestran ejemplos de preguntas relevantes que a un operador le interesaría saber para tener inteligencia competitiva:

Tabla 4: Ejemplos de preguntas relevantes para operador

Pregunta
¿Qué ofrecen otros operadores en Redes Sociales ilimitadas y qué opinan (+/-) sus seguidores?
¿Cuál es el operador que más opiniones (+) tiene en servicio de Datos?
¿Cuál es el operador qué más promociones de portabilidad ofrece?
¿Qué promociones de música ilimitada se ofrecen y qué opinan (+/-) sus clientes?
¿Qué beneficios a la comunidad dan otros operadores y qué opiniones (+/-) tienen?
¿Qué promociones de entradas a conciertos se ofrecen y cuál tiene la opinión más (+)?

Para poder dar respuesta a las preguntas relevantes para un operador como las del ejemplo de Tabla 3 y 4 y que le permitan tener inteligencia competitiva no es posible lograrlo sólo con los datos disponibles en Facebook que no tienen significado o semántica para un computador y sin tener un proceso definido por lo que en esta tesis se plantea implementar un proceso de inteligencia competitiva de promociones que comprende:

- Creación de ontología de dominio en telecomunicaciones que cubra los conceptos principales que un operador utiliza en promociones e incluya un vocabulario de palabras y términos que identifiquen los conceptos para poder clasificar las publicaciones o *posts.*
- Creación de ontología de dominio en polaridad que incluya palabras y términos que permitan identificar la positividad o negatividad de los comentarios de los seguidores.

[16] dle.rae.es/?w=sentimiento

[17] http://dle.rae.es/?id=GjqhajH

- Extracción de todas las publicaciones en Facebook y sus comentarios de los 5 operadores y la utilización de las dos ontologías para clasificar por conceptos de telecomunicaciones los *posts* y la polaridad de los comentarios.
- Implementación de un motor de búsqueda semántica que le permita al operador responder preguntas relevantes, poder compararse con la competencia y tengan información valiosa para mejorar sus promociones para ser más competitivos.

1.2. Marco Conceptual

Para tener inteligencia competitiva se empieza identificando las necesidades de información que le son relevantes al operador y, que aporta el norte al que se debe ir (Gógova 2015). Para ello, es necesario que se establezcan los conceptos de telecomunicaciones que se usan en las promociones y que servirán a los operadores para buscar en los *posts* de Facebook y tener información relevante para la toma de decisiones y ser más competitivos. Estos conceptos se deben organizar en una ontología de dominio que pueda ser usada por un computador para realizar las búsquedas. En la sección 2.2.1, 2.2.2 y 2.2.3 se muestran los conceptos de telecomunicaciones, de inteligencia competitiva y de ontologías de dominio respectivamente.

1.2.1. Telecomunicaciones en Perú

La competencia en telecomunicaciones móviles se inicia en la década de 1990 la cual finalizó con 3 operadores: Telefónica, BellSouth y Nextel. La primera operadora de telefonía celular en el Perú fue Tele 2000 creada en 1991 por el empresario peruano Genaro Delgado Parker [18]. En 1994 el estado peruano privatiza [19] con Telefónica de España a la Compañía Peruana de Teléfonos (CPT-Perú) que brindaba telefonía en Lima y a la Empresa Nacional de Telecomunicaciones (Entel-Perú) que brindaba telefonía fuera de Lima. En 1997 el operador BellSouth de EEUU compra a Tele 2000 y en 1998 ingresa al mercado el operador Nextel de EEUU [20]. En esta década el servicio ofrecido se concentraba principalmente en Lima.

La década del 2000 finalizó con 3 operadores: Telefónica, Claro y Nextel. En 2001 ingresa al mercado el operador TIM [21] de Italia. En 2004 Telefónica compra [22] a BellSouth. En 2005, el operador América Móvil de México compra a TIM [23] bajo su marca Claro. En esta década el servicio ofrecido se expande a todo el Perú.

En la década de 1990 y 2000 las promociones de los operadores se realizaban en medios de comunicación tradicionales como televisión, radio, diarios y revistas. A partir del 2009, a finales de la década del 2000, los operadores empiezan a publicar sus promociones en redes sociales.

Los 3 operadores Telefónica, Claro y Nextel compitieron durante 9 años desde el 2005 hasta el 2013 [24]. El 2013 el operador Entel de Chile compra [25] a Nextel y cambia la marca por Entel en 2014 [26]. En

[18] http://elcomercio.pe/economia/negocios/historia-revolucion-telefonia-movil-peru-330906

[19] http://elpais.com/diario/1994/03/01/economia/762476407_850215.html

[20] https://www.osiptel.gob.pe/Archivos/Publicaciones/Secci%C3%B3n_III.pdf

[21] https://www.osiptel.gob.pe/Archivos/Sector_telecomunicaciones/Desarrollo_Sector/Hito_Am%C3%A9ricaM%C3%B3vilCompraTIM.pdf

[22] http://elpais.com/diario/2004/03/09/economia/1078786802_850215.html

[23] http://larepublica.pe/11-08-2005/america-movil-compra-tim

[24] https://www.osiptel.gob.pe/articulo/24-lineas-en-servicio-por-empresa

[25] http://archivo.elcomercio.pe/economia/negocios/entel-chile-concreto-compra-nextel-peru-us400-millones-noticia-1559247

2014 inicia sus operaciones [27] el operador Bitel de Vietnam. Finalmente ingresa el operador virtual Virgin [28] en 2016. En 2017 existen 5 operadores compitiendo: Telefónica bajo la marca Movistar, Claro, Entel, Bitel y Virgin.

Para tener inteligencia competitiva es relevante para un operador conocer qué promociones están ofreciendo los operadores respecto a los servicios que brindan. Actualmente los operadores ofrecen a sus clientes estos tres servicios principales [29]:

- **Servicio de voz:** Permite al cliente hacer llamadas de telefonía celular hacia números destino del mismo operador (dentro de la red del operador (en adelante, *on-net*)) o hacia números de otros operadores (fuera de la red del operador (en adelante, *off-net*)). Adicionalmente, se permiten llamadas a números fijos y a números de larga distancia internacional. Usualmente cuando se quiere indicar que se puede llamar a destinos *on-net* y *off-net* se dice "llamadas a cualquier operador" y cuando además incluye destinos fijos y larga distancia internacional se dice "llamadas a todo destino".
- **Servicio de mensajes de texto**: También llamado SMS (*Short Message Service*), permite al cliente enviar mensajes de texto cortos de hasta 140 caracteres hacia destinos *on-net* y *off-net*.
- **Servicio de datos**: Permite al cliente navegar en internet desde el celular y utilizar aplicaciones que usan datos como, por ejemplo:
    - **Correo:** Gmail, Outlook
    - **Redes sociales y Mensajería:** Facebook, Instagram, WhatsApp, Messenger
    - **Música:** Spotify
    - **Video:** Youtube, Netflix

Con el ingreso de las redes 4G en el 2014 [30] que permiten velocidades de datos muy altas se incrementó el uso del servicio de datos por los clientes quienes demandan cada vez más promociones por los operadores. En 2016 se desató una guerra de promociones [31] de datos y en 2017 se ofrece uso ilimitado de aplicaciones [32] (*apps*) en contenidos de entretenimiento como música (Ej.: Spotify) y video (Ej.: Youtube) para capturar nuevos clientes.

Los servicios de voz, mensajes de texto y datos que ofrecen los operadores tienen una tarifa en soles diferente dependiendo de la modalidad de contratación que haga el cliente con el operador. A fines del 2016 se tenían 37.7 millones de líneas celulares activas siendo la modalidad de contratación más común el prepago que representa el 69%, y el postpago con 31% según Osiptel [33].

En la modalidad prepago, el cliente cuando necesita usar los servicios hace una recarga virtual en soles y generalmente recibe un bono o regalo y además con los soles recargados puede comprar bolsas de minutos, SMS o datos (Ej.: Bolsa de 100 MB x 7 días x S/. 5, Bolsa de 20 minutos a cualquier operador x 3 días x S/. 3, Bolsa de WhatsApp ilimitado x 1 día x S/. 1). Cuando se acaba el saldo de los bonos y bolsas y el cliente aún tiene saldo de recarga en soles puede usarlo directamente para realizar llamadas de voz (Ej.: A tarifa de S/. 0.35 x minuto) o para enviar SMS (Ej.: A tarifa de S/. 0.10 x SMS).

---

[26] http://elcomercio.pe/economia/negocios/nextel-llama-entel-apunta-30-mercado-peru-287297

[27] http://elcomercio.pe/economia/negocios/bitel-nuevo-operador-movil-lanzo-oficialmente-peru-288354

[28] http://rpp.pe/economia/economia/virgin-mobile-inicio-hoy-sus-operaciones-en-el-peru-noticia-981465

[29] http://www.movistar.com.pe/movil, http://www.claro.com.pe/personas/movil/, http://www.entel.pe/, http://www.bitel.com.pe, https://www.inkacel.com (Ex Virgin Mobile Perú)

[30] http://elcomercio.pe/tecnologia/actualidad/4g-lte-diez-datos-debes-red-comenzara-funcionar-manana-375491

[31] http://elcomercio.pe/economia/peru/necesitas-acerca-guerra-telefonia-movil-225608

[32] http://elcomercio.pe/economia/negocios/operadoras-mantienen-guerra-ofertas-apps-ilimitadas-418094

[33] http://www.osiptel.gob.pe/articulo/25-lineas-en-servicio-por-modalidad-y-por-empresa

En la modalidad postpago, el cliente paga mensualmente al operador el costo de un plan tarifario (Ej.: S/. 79) que le otorga al inicio del mes una cantidad fija de minutos, SMS y datos vigentes hasta el final del mes (Ej.: 500 minutos a cualquier operador + 800 SMS + 2 GB de datos + WhatsApp y Facebook ilimitado). Generalmente los planes postpago se ofrecen como control y libre. En el caso de control cuando se acaba la cantidad fija entregada, el cliente ya no puede hacer uso del servicio hasta que recargue soles o compre una bolsa o se le otorgue nuevamente los minutos, SMS y datos al inicio del siguiente mes. En el caso de libre cuando se acaba la cantidad fija entregada, el cliente puede seguir usando los servicios y el exceso es tarifado a un costo predefinido (Ej.: S/. 0.20 x minuto, S/. 0.05 x SMS, S/. 0.05 x MB) el cual viene sumado en el recibo de servicios que se le entrega al cliente a fin de mes.

Los operadores constantemente están ofreciendo promociones en la modalidad prepago [34] y postpago [35] también llamados planes tarifarios prepago, control y libre; por ejemplo, dando mejores bonos o mayores cantidades de minutos, SMS, datos por el mismo costo de la bolsa o costo mensual del plan tarifario.

La portabilidad numérica en 1 día, que entró en vigencia en 2014, intensificó la competitividad entre los operadores quienes generalmente otorgan más beneficios a los clientes que se portan con su número telefónico desde otro operador. Por ejemplo, Entel fue el operador cuya estrategia comercial desde el inicio de sus operaciones en 2014 fue captar nuevos clientes por portabilidad lanzando planes que incluían iPhones [36] a S/. 9 y, que según Osiptel [37], a Mayo del 2017 le ha permitido capturar un total neto de más de 1 millón de líneas celulares por portabilidad.

Todos los conceptos indicados como, por ejemplo, servicios, tarifas, recargas, uso ilimitado, modalidad de contratación, prepago, postpago control, postpago libre y portabilidad son relevantes para un operador conocer qué es lo que está ofreciendo la competencia y qué opiniones positivas o negativas tienen sus seguidores, pero no pueden saberlo porque los datos publicados en Facebook son texto en lenguaje natural sin significado o semántica para el computador y además no cuentan con un proceso definido de inteligencia competitiva para implementarlo.

1.2.2. Inteligencia competitiva

Según la Norma UNE 166006 (Norma UNE 166006), inteligencia competitiva es un proceso ético y sistemático de recolección y análisis de información acerca del ambiente de negocios, de los competidores y de la propia organización, y comunicación de su significado e implicaciones destinada a la toma de decisiones.

La inteligencia competitiva es una herramienta estratégica básica para lograr la supervivencia y el éxito en las condiciones de competencia actuales (Arroyo Varela 2005). El proceso para lograr inteligencia parte de la unidad más elemental, que son los datos; una vez que son organizados, los datos se convierten en información; la información, cuando se analiza, se convierte en inteligencia (Arroyo Varela 2005).

Arroyo Varela (2005) plantea el ciclo de la inteligencia competitiva en cuatro fases según se muestra en la Figura 2 y que comprende:

[34] https://elcomercio.pe/economia/negocios/tarifas-moviles-batalla-dos-432997

[35] https://elcomercio.pe/economia/evoluciona-negocio-telefonia-movil-peru-noticia-470898

[36] http://elcomercio.pe/economia/negocios/pagina-entel-saturada-tienes-planes-287490

[37] http://www.osiptel.gob.pe/repositorioaps/data/1/1/1/par/reporte-portabilidad-numerica-mayo2017/Portabilidad_Numerica-may2017.pdf

- **Fase 1 - Identificación de necesidades de información:** Identificar, a lo largo de la organización, las necesidades de inteligencia para la toma de decisiones clave; para ello hay que identificar a los tomadores de decisiones clave y sus necesidades particulares de inteligencia.
- **Fase 2 - Recogida de información:** De fuentes impresas, electrónicas y orales, sobre eventos del entorno de la empresa.
- **Fase 3 - Análisis y síntesis de la información:** Los decisores necesitan un análisis preciso, argumentos y recomendaciones, lo que implica que es necesario aportar valor añadido a la información.
- **Fase 4 - Reparto de la inteligencia a los tomadores de decisiones:** La inteligencia es comunicada a los tomadores de decisiones quienes la utilizarán en la formulación de sus planes.

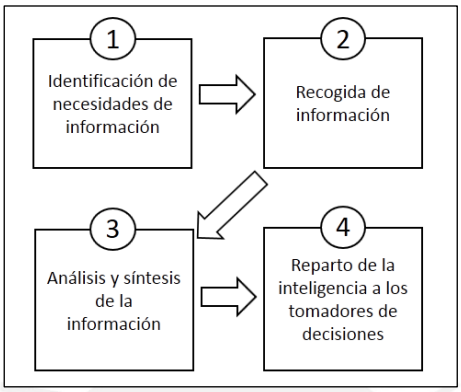
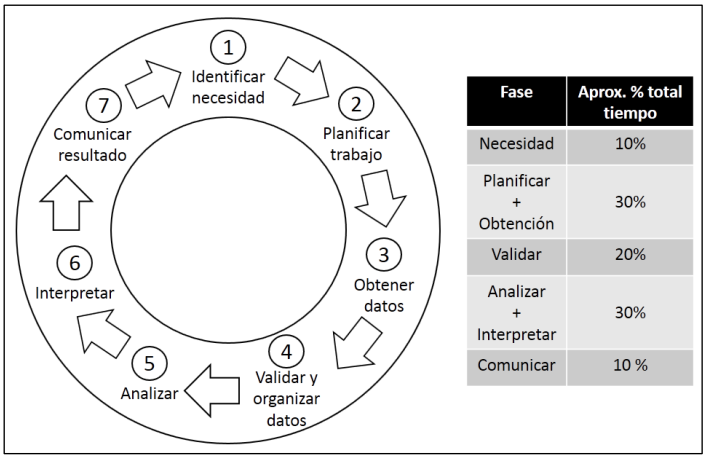


Figura 2: Ciclo de la inteligencia competitiva (Arroyo Varela 2005)

Gógova (2015) plantea el ciclo de la inteligencia competitiva como un proceso con siete actividades según se muestra en la Figura 3 y que comprende:

- **Actividad 1 – Identificar necesidad:** Aporta el norte al que se debe ir; aporta la dirección clara, real y consensuada del posterior trabajo.
- **Actividad 2 – Planificar trabajo:** Qué se debe buscar y dónde; acotar los criterios de relevancia para depurar la información y las técnicas a aplicar en el análisis.
- **Actividad 3 - Obtener datos:** Es la materia prima, la información que se obtenga va a determinar en gran medida los costos posteriores y la calidad del producto final.
- **Actividad 4 – Validar y organizar datos:** Depurar la información obtenida; asegurar su confiabilidad, deshacerse de lo irrelevante/superfluo y estructurarla en bloques relacionados.
- **Actividad 5 – Analizar:** Aporta el máximo valor en el proceso, relaciona de manera congruente y correcta la información relevante, detecta relaciones ocultas y/o emergentes. Es el paso que sustenta el nivel de entendimiento profundo del entorno.
- **Actividad 6 – Interpretar:** Conecta los resultados del análisis con los objetivos/problemática de la empresa; confiere el grado de utilidad real del trabajo realizado para la toma de decisiones. Es el paso que más inteligencia aporta al proceso.
- **Actividad 7 – Comunicar resultado:** Hacer que el valor del trabajo realizado llegue, en plazo y forma, a todos y cada uno de aquellos que lo han pedido y/o lo necesitan.



Fase	Aprox. % total tiempo
Necesidad	10%
Planificar + Obtención	30%
Validar	20%
Analizar + Interpretar	30%
Comunicar	10 %

Figura 3: Ciclo de la inteligencia competitiva (Gógova 2015)

Respecto a la fase 2 de **recogida de información** de la Figura 2 y a la actividad 3 de **obtener datos** de la Figura 3, Muñoz Cañavate (2012) menciona que las redes sociales se han convertido en una nueva fuente de información actual y muy dinámica, las redes sociales pueden ser utilizadas por las empresas para establecer vínculos con el exterior de una forma totalmente distinta, así los clientes pueden aportar información sobre los productos y servicios a través de sus comentarios; además las redes sociales permiten a una empresa realizar campañas de publicidad, promoción de productos y servicios. De forma similar, los operadores utilizan Facebook para publicar sus promociones y recibir la retroalimentación de sus seguidores a través de comentarios por lo que puede usarse como fuente de información para la inteligencia competitiva como se plantea en esta tesis. Gógova (2015) indica que la inteligencia competitiva maneja datos e información del pasado, del presente y previsiones / tendencias / prospectivas del futuro por lo que en la recolección de datos de esta tesis se extraerán todos los *posts* publicados por cada operador y sus comentarios desde que iniciaron en Facebook.

Respecto a la fase 4 de **reparto de la inteligencia a los tomadores de decisiones** de la Figura 2 y a la actividad 7 de **comunicar resultado** de la Figura 3, Gógova (2015) indica que el propósito de la inteligencia competitiva es facilitar la toma de decisiones con menor grado de riesgo e incertidumbre, en beneficio de la empresa.

1.2.3. Ontologías de dominio

Una estrategia cada vez más dominante para la organización de la información en una forma más amigable y entendible para el computador está asociada con el término ontología o, a veces denominada, "ingeniería ontológica" u "ontología aplicada". La ontología se puede entender como un vocabulario controlado para representar los tipos de entidades o conceptos en un *dominio* dado (Arp et al. 2015).

Según Arp et al. (2015), en las ciencias como por ejemplo en biología y biomedicina, hay gran cantidad de información que se genera de forma diaria como resultado de los avances tecnológicos e investigaciones y se publican principalmente en revistas científicas. A pesar de que los investigadores tienen acceso a estas publicaciones tienen obstáculos para la accesibilidad debido a que los miembros de la comunidad de investigadores científicos utilizan palabras, terminologías y

sistemas de codificación diferentes para describir los mismos conceptos en los resultados de sus investigaciones lo que dificulta las búsquedas de investigaciones o *papers* por el computador.

De igual manera, en telecomunicaciones se manejan varios conceptos como servicios, tarifas, recargas, uso ilimitado, modalidad de contratación, planes tarifarios y portabilidad. Para comunicar una promoción en una publicación o *post* de Facebook las áreas de marketing de los operadores, usan diferentes palabras o términos, algunos usan palabras más formales y otros más informales para describir el mismo concepto y que los seguidores entiendan mejor la promoción. Debido al uso de palabras diferentes para un mismo concepto se hace difícil para un computador poder clasificar y encontrar todos los *posts* de un concepto y tener información relevante para el operador. En la Tabla 5 se muestran ejemplos de conceptos de telecomunicaciones y palabras utilizadas en los *posts* publicados en Facebook de los operadores.

Tabla 5: Ejemplo de conceptos de telecomunicaciones y palabras utilizadas en *posts* de Facebook

Conceptos	Palabras utilizadas en *posts* de Facebook que identifican al concepto
Servicio: Voz	habla, llama, minutos, telefonía
Servicio: Datos	4g, chatear, postear, velocidad, sube, megas, navega, internet, gb, facebook, fb
Tarifa del Servicio: Cero	gratis, ilimitado, “sin consumir tu saldo”, “sin costo”, “sin saldo mínimo”
Plan Tarifario: Control y Libre	contrato, deuda, postpago, recibo
Portabilidad	cámbiate, ex, migra

Las palabras y conceptos del ejemplo indicado en la Tabla 5 podrían servir para crear una ontología de dominio en telecomunicaciones, tal como se plantea en esta tesis, que pueda ser utilizada en la búsqueda de *posts* de Facebook de conceptos; el ejemplo es coherente con las definiciones de Gruber (1995) que indica que una ontología es una especificación explícita de una conceptualización y de (Neches et al. 1991) que indica que una ontología define los términos básicos y las relaciones que comprenden el vocabulario de un área temática.

Horridge (2011) indica los siguientes componentes que tiene una ontología:

- **Individuos:** Representan objetos en el dominio en el que estamos interesados. Los individuos también son conocidos como instancias. Los individuos pueden ser referidos como "instancias de clases". En la Figura 4 se muestra un ejemplo de representación de individuos en el dominio de telecomunicaciones. Los individuos están representados como diamantes.

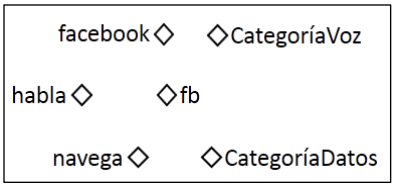


Figura 4: Ejemplo de representación de individuos

- **Propiedades:** Las propiedades son relaciones binarias entre individuos, es decir, las propiedades unen a dos individuos. Por ejemplo, en el dominio de telecomunicaciones, la propiedad `tieneAbreviatura` vincula al individuo facebook con el individuo fb y `tieneCategoría` vincula a facebook con categoríaDatos tal como se muestra en Figura 5.

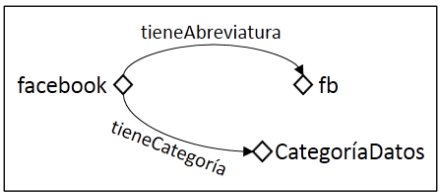


Figura 5: Ejemplo de representación de propiedades

- **Clases:** Las clases se interpretan como conjuntos que contienen individuos. La palabra "concepto" se utiliza a veces en lugar de la clase. Las clases son una representación concreta de conceptos. Las clases pueden organizarse en una jerarquía de superclase-subclase, que también se conoce como taxonomía. Las subclases especializan sus superclases. En la Figura 6 se muestra una representación de dos clases Palabra y Categoría en el dominio de telecomunicaciones que contienen individuos. Las clases pueden representarse como círculos u óvalos similar a los conjuntos en los diagramas de Venn. En la Figura 7 se muestran dos taxonomías o jerarquías de superclase-subclase en el dominio de telecomunicaciones, la primera taxonomía es la superclase Plan con subclases Prepago, Control y Libre y la segunda es la superclase Servicio con subclases Voz, SMS y Datos, se muestra también la relación `tieneServicio` entre la superclase Plan con la superclase Servicio.

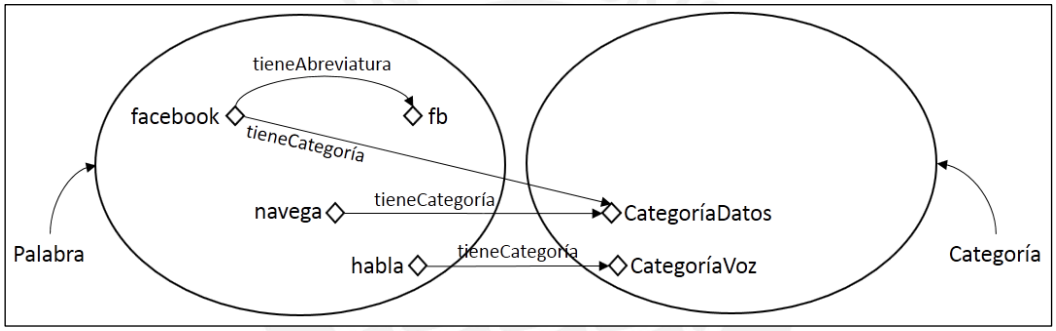


Figura 6: Ejemplo de representación de clases conteniendo individuos

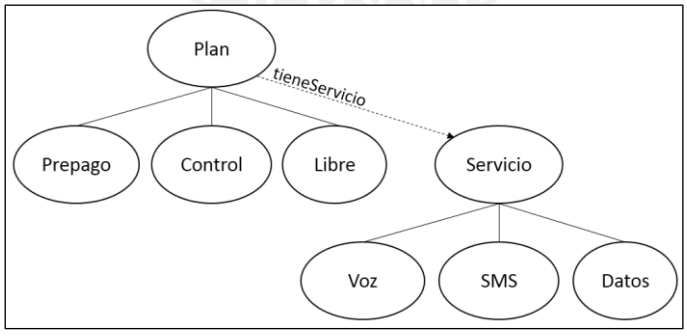


Figura 7: Ejemplo de representación de taxonomías y relaciones entre clases

Los componentes de la ontología que son los individuos, propiedades y clases serán utilizados en la creación de la ontología de dominio en telecomunicaciones que se explicará en detalle en Capítulo 5.

2. GENERALIDADES

2.1. Introducción

Los operadores actualmente no explotan los datos de Facebook de sus competidores para tener información relevante y ser más competitivos, a pesar de que los datos de Facebook como las publicaciones de promociones y sus comentarios son públicos al estar escritos en lenguaje natural no es posible para un computador entender su significado o semántica y poder clasificar las publicaciones en los conceptos del dominio de telecomunicaciones como por ejemplo planes, servicios, tarifas y portabilidad y clasificar los comentarios de los seguidores por polaridad positiva o negativa y permitir al operador compararse con sus competidores en los diferentes conceptos y tener inteligencia competitiva.

2.2. Objetivo general

Diseñar e implementar un proceso de inteligencia competitiva para los operadores, a partir de las publicaciones de promociones y sus comentarios registrados en Facebook por los operadores y sus seguidores, que les permita responder preguntas relevantes y compararse con la competencia en el dominio de las telecomunicaciones para que tengan información relevante en la toma de decisiones y contribuya al diseño de promociones más competitivas.

2.3. Objetivos específicos

OE.1: Extraer y limpiar todas las publicaciones (*posts*) realizadas por los 5 operadores en sus páginas de Facebook, incluyendo los comentarios publicados por sus seguidores, para generar el corpus.

OE.2: Clasificar semánticamente cada publicación (*post*) usando su texto completo e identificando palabras y expresiones en lenguaje natural para darles significado en el dominio de telecomunicaciones.

OE.3: Clasificar semánticamente la polaridad (positiva, negativa, neutra) de cada comentario usando su texto completo e identificando palabras en lenguaje natural que reflejen la polaridad.

OE.4: Implementar un motor de búsqueda semántica de publicaciones (*posts*) en el dominio de telecomunicaciones y un comparador de resultados por operador, y mostrar información relevante para el diseño de promociones más competitivas.

2.4. Resultados esperados

Del **OE.1** se plantean los siguientes resultados esperados e indicadores mostrados en la Tabla 6:

Tabla 6: Resultados esperados e indicadores de OE.1

Objetivo específico	Resultados esperados	Indicadores de verificación
OE.1: Extraer y limpiar todas las publicaciones (*posts*) realizadas por los 5 operadores en sus páginas de Facebook, incluyendo los comentarios publicados por sus seguidores, para generar el corpus.	**RE.1a:** Proceso automático de extracción de *posts* y sus comentarios de páginas de Facebook de los 5 operadores.	• Diagrama de flujo del proceso de extracción. • Software de extracción de *posts* y comentarios. • Estadística de *posts* y comentarios extraídos.
	RE.1b: Proceso automático de limpieza de *posts* y sus comentarios para generar el corpus.	• Diagrama de flujo del proceso de limpieza. • Software de limpieza de *posts* y comentarios. • Estadística de *posts* y comentarios limpiados.

Del **OE.2** se plantean los siguientes resultados esperados e indicadores mostrados en la Tabla 7:

Tabla 7: Resultados esperados e indicadores de OE.2

Objetivo específico	Resultados esperados	Indicadores de verificación
OE.2: Clasificar semánticamente cada publicación (*post*) usando su texto completo e identificando palabras y expresiones en lenguaje natural para darles significado en el dominio de telecomunicaciones.	**RE.2a:** Ontología de dominio, en formato RDF, para representar el conocimiento en telecomunicaciones.	• Diagrama de conceptos y relaciones. • Diagrama de relaciones entre instancias. • Tabla de palabras por concepto.
	RE.2b: Proceso automático para clasificar todas las publicaciones aplicando la ontología en telecomunicaciones.	• Diagrama de flujo del proceso de clasificación de *posts*. • Software de clasificación de *posts*. • Estadística de clasificación de *posts*.

Del **OE.3** se plantean los siguientes resultados esperados e indicadores mostrados en la Tabla 8:

Tabla 8: Resultados esperados e indicadores de OE.3

Objetivo específico	Resultados esperados	Indicadores de verificación
OE.3: Clasificar semánticamente la polaridad (positiva, negativa, neutra) de cada comentario usando su texto completo e identificando palabras en lenguaje natural que reflejen la polaridad.	**RE.3a:** Ontología de dominio, en formato CSV, para representar la polaridad.	• Diagrama del proceso de creación de la ontología. • Algoritmo de cálculo de polaridad de palabra. • Tabla de palabras de la ontología.
	RE.3b: Proceso automático para clasificar todos los comentarios aplicando la ontología en polaridad.	• Diagrama de flujo del proceso de clasificación de comentarios. • Software de clasificación de comentarios. • Estadística de clasificación de comentarios.

Del **OE.4** se plantean los siguientes resultados esperados e indicadores mostrados en la Tabla 9:

Tabla 9: Resultados esperados e indicadores de OE.4

Objetivo específico	Resultados esperados	Indicadores de verificación
OE.4: Implementar un motor de búsqueda semántica de publicaciones (*posts*) en el dominio de telecomunicaciones y un	**RE.4a:** Aplicación Web que permita al usuario realizar búsquedas semánticas de publicaciones en el dominio de telecomunicaciones.	• Diagrama de componentes de la aplicación web de búsquedas. • Interfaz gráfica de usuario para ingresar los criterios de búsqueda de *posts*.

comparador de resultados por operador y mostrar información relevante para el diseño de promociones más competitivas.		• Algoritmo de búsqueda de *posts*. • Interfaz gráfica de usuario de resultados de búsqueda. • Software de aplicación web de búsqueda.
	RE.4b: Aplicación Web que compare los resultados de la búsqueda semántica por operador.	• Diagrama de componentes de la aplicación web de comparación. • Interfaz gráfica de usuario de comparación de *posts* por operador. • Software de aplicación web de comparación.

En la Figura 8 se muestra un resumen e interrelación de los objetivos específicos y resultados esperados, se empieza con un proceso extractor (RE.1a) que recolecta todas las publicaciones o *posts* y comentarios de Facebook de los 5 operadores (Movistar, Claro, Entel, Bitel y Virgin) y genera archivos de texto en formato CSV. Estos archivos ingresan a un proceso limpiador (RE.1b) que principalmente elimina las publicaciones y comentarios con texto vacío y genera archivos de texto en formato CSV limpios. Los archivos limpios de *posts* ingresan al proceso clasificador de *posts* (RE.2b) que utiliza una ontología de dominio de telecomunicaciones (RE.2a) para clasificar en conceptos de telecomunicaciones a todas las publicaciones y generar archivos de texto en formato CSV clasificados. Los archivos limpios de comentarios ingresan al proceso clasificador de comentarios (RE.3b) que utiliza una ontología de dominio de polaridad (RE.3a) para clasificar positivo, negativo o neutro a todos los comentarios y generar archivos de texto en formato CSV clasificados. Ambos grupos de archivos de texto clasificados en formato CSV de *posts* y comentarios son utilizados por la aplicación web de búsqueda semántica de *posts* (RE.4a) para que los tomadores de decisiones del operador puedan responder preguntas relevantes buscando por conceptos y palabras de la ontología de dominio de telecomunicaciones (RE.2a). Los *posts* y comentarios que se obtienen como resultados de la búsqueda se pueden comparar mediante la aplicación web comparador de resultados (RE.4b) que brinda información relevante al operador para la toma de decisiones y tener inteligencia competitiva.

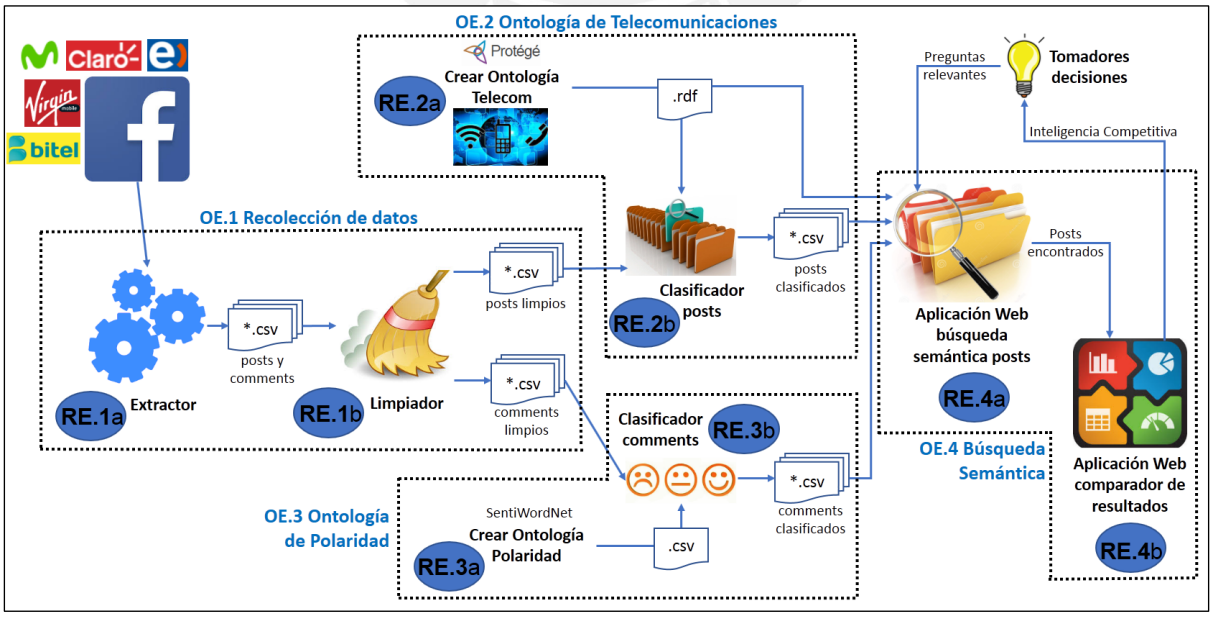


Figura 8: Resumen e interrelación de objetivos específicos y resultados esperados

2.5. Métodos y procedimientos

Para lograr el objetivo general de diseñar e implementar un proceso de inteligencia competitiva, se realizó un mapeo de los objetivos y resultados esperados de esta tesis con el modelo de ciclo de inteligencia competitiva de Arroyo Varela (2005) de Figura 2 y de Gógova (2015) de Figura 3 para asegurar que se cubran todas las fases o actividades del proceso de inteligencia competitiva. Los resultados se muestran en la Tabla 10.

Tabla 10: Mapeo de modelos de ciclo de inteligencia competitiva con objetivos y resultados esperados de tesis

<table>
<tr><th>Arroyo Varela (2005)</th><th>Gógova (2015)</th><th colspan="2">Tesis</th></tr>
<tr><td rowspan="2">Fase 1 - Identificación de necesidades de información</td><td>Actividad 1 – Identificar necesidad</td><td colspan="2" rowspan="2">Objetivo general:
Diseñar e implementar un proceso de inteligencia competitiva para los operadores, a partir de las publicaciones de promociones y sus comentarios registrados en Facebook por los operadores y sus seguidores, que les permita responder preguntas relevantes y compararse con la competencia en el dominio de las telecomunicaciones para que tengan información relevante en la toma de decisiones y contribuya al diseño de promociones más competitivas</td></tr>
<tr><td>Actividad 2 – Planificar trabajo</td></tr>
<tr><td>Fase 2 - Recogida de información</td><td>Actividad 3 - Obtener datos</td><td>OE.1: Extraer y limpiar todas las publicaciones (posts) realizadas por los 5 operadores en sus páginas de Facebook, incluyendo los comentarios publicados por sus seguidores, para generar el corpus.</td><td>RE.1a: Proceso automático de extracción de posts y sus comentarios de páginas de Facebook de los 5 operadores.
RE.1b: Proceso automático de limpieza de posts y sus comentarios para generar el corpus.</td></tr>
<tr><td rowspan="4">Fase 3 - Análisis y síntesis de la información</td><td>Actividad 4 – Validar y organizar datos</td><td>OE.2: Clasificar semánticamente cada publicación (post) usando su texto completo e identificando palabras y expresiones en lenguaje natural para darles significado en el dominio de telecomunicaciones.
OE.3: Clasificar semánticamente la polaridad (positiva, negativa, neutra) de cada comentario usando su texto completo e identificando palabras en lenguaje natural que reflejen la polaridad.</td><td>RE.2a: Ontología de dominio, en formato RDF, para representar el conocimiento en telecomunicaciones.
RE.2b: Proceso automático para clasificar todas las publicaciones aplicando la ontología en telecomunicaciones.
RE.3a: Ontología de dominio, en formato CSV, para representar la polaridad.
RE.3b: Proceso automático para clasificar todos los comentarios aplicando la ontología en polaridad.</td></tr>
<tr><td>Actividad 5 – Analizar</td><td rowspan="3">OE.4: Implementar un motor de búsqueda semántica de publicaciones (posts) en el dominio de telecomunicaciones y un comparador de resultados por operador y mostrar información relevante para el diseño de promociones más competitivas.</td><td>RE.4a: Aplicación Web que permita al usuario realizar búsquedas semánticas de publicaciones en el dominio de telecomunicaciones.</td></tr>
<tr><td>Actividad 6 – Interpretar</td><td rowspan="2">RE.4b: Aplicación Web que compare los resultados de la búsqueda semántica por operador.</td></tr>
<tr><td>Fase 4 - Reparto de la inteligencia a los tomadores de decisiones</td><td>Actividad 7 – Comunicar resultado</td></tr>
</table>

Para poder obtener los resultados esperados se usarán las herramientas de software libre y dominio público indicadas en la tabla 11.

Tabla 11: Herramientas de software libre y dominio público a utilizar

Herramienta	Descripción
Python	Lenguaje de programación Python versión 3.5.2 de 64 bits (Distribución Anaconda 4.2.0 https://www.continuum.io)
Api Graph de Facebook	Api Graph de Facebook versión 2.8. El Api Graph de Facebook https://developers.facebook.com/docs/graph-api, es una interface pública HTTPS que permite leer y escribir datos en cuentas de Facebook. Para utilizar el Api Graph de Facebook es necesario generar un token de acceso el cual puede obtenerse de dos formas:

	• En https://developers.facebook.com/tools/explorer ingresando con un usuario y contraseña válido de Facebook se genera un token temporal vigente por 1 hora y media. • Registrándose como desarrollador de Facebook en https://developers.facebook.com/ se puede generar un token válido por 2 meses.
Librería requests	Librería requests de Python para usar páginas HTTP y HTTPS http://docs.python-requests.org
Protégé	Software Protégé versión 5.2.0 de la universidad de Stanford http://protege.stanford.edu para la elaboración de ontologías en formato RDF.
HermiT	Reasoner HermiT 1.3.8.413, incluido en Protégé versión 5.2.0, para generar las inferencias de la ontología.
Librería ontospy	Librería ontospy de Python https://github.com/lambdamusic/OntoSpy/wiki para leer ontologías en formato RDF.
Librería NLTK	NLTK o *Natural Language Toolkit* http://www.nltk.org/ es una librería de Python para trabajar con texto en lenguaje natural. Proporciona interfaces fáciles de usar para más de 50 corpus y recursos léxicos como WordNet, junto con un conjunto de bibliotecas útiles para procesamiento de texto.
WordNet	WordNet es una gran base de datos léxica o de palabras en inglés de dominio público. Los sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos de conocimiento (*synsets*), cada uno de ellos expresando un concepto distinto. Se utilizó WordNet 3.0 en inglés https://wordnet.princeton.edu/ contenida en la librería nltk de Python que contiene aproximadamente 145,000 synsets.
Open Multilingual WordNet	Open Multilingual WordNet facilita el uso de WordNet 3.0 en inglés en otros 15 idiomas como el español y es de dominio público; contiene 36,681 palabras en español que corresponden a 38,512 *synsets*. Open Multilingual WordNet http://compling.hss.ntu.edu.sg/omw/ y http://adimen.si.ehu.es/web/MCR/ está contenida en la librería nltk de Python y está completamente integrada con WordNet 3.0.
SentiWordNet	Es un recurso léxico o conjunto de palabras para la minería de opinión o análisis de polaridad de dominio público. Se utilizó SentiWordNet 3.0 en inglés http://sentiwordnet.isti.cnr.it/ contenida en la librería nltk de Python que asigna a cada synset de WordNet un puntaje de polaridad de positividad y negatividad.
Librería flask	Librería flask de Python que es un microframework para desarrollo de aplicaciones web http://flask.pocoo.org
Librería Google charts	Librería para hacer gráficos en páginas web, Google charts https://developers.google.com/chart/
Librería bootstrap	Librería de estilos para páginas web, bootstrap versión 3.3.7 http://getbootstrap.com/

En la Tabla 12 se muestran los métodos y procedimientos a utilizar por cada resultado esperado:

Tabla 12: Métodos y procedimientos de resultados esperados

Resultados esperados	Métodos y Procedimientos
RE.1a: Proceso automático de extracción de *posts* y sus comentarios de páginas de Facebook de los 5 operadores.	Se usará Python para la programación del extractor y el cálculo de estadísticas. Para el acceso a Facebook se usará el Api Graph de Facebook con la librería requests de Python. Los *posts* y comentarios extraídos serán almacenados en archivos de texto en formato CSV.
RE.1b: Proceso automático de limpieza de *posts* y sus comentarios para generar el corpus.	Se usará Python para la programación del limpiador y el cálculo de estadísticas. Se generará un solo archivo CSV que contendrá todos los *posts* de los 5 operadores y un archivo CSV de los comentarios por cada *post*.
RE.2a: Ontología de dominio, en formato RDF, para representar el conocimiento en telecomunicaciones.	Se usará Protégé para la elaboración de la ontología de dominio en telecomunicaciones y se generará un archivo en formato RDF. Se usará el reasoner HermiT, incluido en Protégé, para generar las inferencias y se generará un archivo en formato RDF.
RE.2b: Proceso automático para clasificar todas las publicaciones aplicando la ontología en telecomunicaciones.	Se usará Python para la programación del clasificador y el cálculo de estadísticas. Para la lectura de los archivos de la ontología en formato RDF se usará la librería de python ontospy. Se generará un solo archivo CSV que contendrá todos los *posts* clasificados.
RE.3a: Ontología de dominio, en formato CSV, para representar la polaridad.	Se usará WordNet, Open Multilingual WordNet y SentiWordNet contenidas en la librería NLTK de Python, para la elaboración de la ontología de dominio en polaridad y se generará un archivo en formato CSV.
RE.3b: Proceso automático para clasificar todos los comentarios aplicando la ontología en polaridad.	Se usará Python para la programación del clasificador y el cálculo de estadísticas. Para la lectura del archivo en formato CSV de la ontología se usará Python. Se generará un archivo CSV de comentarios clasificados por cada *post*.
RE.4a: Aplicación Web que permita al usuario realizar búsquedas semánticas de publicaciones en el dominio de telecomunicaciones.	Se usará Python para la programación de ambas aplicaciones web junto con el microframework de Python flask. Para la interfaz gráfica de usuario se usará la librería gráfica Google charts y la librería de estilos bootstrap.
RE.4b: Aplicación Web que compare los resultados de la búsqueda semántica por operador.	

2.6. Justificación

El aporte de esta tesis es muy importante porque es una aplicación práctica real de un proceso formal de inteligencia competitiva en el sector de telecomunicaciones del Perú, que es uno de los sectores económicos más dinámicos y competitivos en el mercado. La innovación de esta tesis es que combina recursos de ingeniería informática, para extraer y procesar datos de dominio público, con un proceso de negocio como la inteligencia competitiva que agrega valor en la planificación estratégica y financiera de las empresas operadoras de telecomunicaciones del Perú.

Todos los operadores (Movistar, Claro, Entel, Bitel y Virgin) se pueden beneficiar, utilizando esta investigación, de la siguiente manera:

- **Organización del conocimiento de telecomunicaciones:** Para tener inteligencia competitiva hay que saber qué se necesita monitorear, los operadores tendrán organizados todos los conceptos relevantes de telecomunicaciones y sus relaciones como por ejemplo planes, servicios, tarifas y portabilidad en una ontología de dominio que es un modelo de conocimiento que tendrá un vocabulario de palabras y expresiones que identifiquen los conceptos. Con este modelo se podrá dar significado o semántica y clasificar todas las publicaciones de Facebook y será usado también para que el operador pueda hacer las búsquedas y responder a preguntas relevantes.
- **Comparación de Competitividad:** Al contar con todas las publicaciones y comentarios de Facebook, los operadores podrán realizar búsquedas por uno o varios conceptos y podrán ver en una línea de tiempo la evolución del pasado hasta el presente de la cantidad de publicaciones por operador y la cantidad de comentarios con su polaridad positiva, negativa o neutra pudiendo compararse con los otros operadores y en caso requieran podrán acceder al detalle; esto les permitirá ver tendencias y hacer los ajustes necesarios en sus promociones y estrategias comerciales y de inversión para el futuro tal como lo sugiere Gógova (2015).
- **Diseño de Promociones:** Los operadores podrán saber en qué conceptos están igual, mejor o peor que los otros operadores y revisando los comentarios de los seguidores según la polaridad podrán tomarlos en cuenta en el diseño de nuevas promociones que sean más competitivas y beneficie a sus clientes.

Adicionalmente, en la coyuntura actual donde la participación de mercado de los operadores ha cambiado considerablemente en los últimos dos años es relevante para los operadores contar con un proceso formal de inteligencia competitiva de promociones para tomar las acciones necesarias que les permitan mejorar su participación de mercado.

En la Figura 9 se aprecia el crecimiento en participación de mercado del 2014 al 2016 de Entel y Bitel quienes alcanzaron a fines del 2016, según Osiptel [38], una participación del mercado celular sumada del 22.70% con tendencia al alza en comparación con tendencia a la baja de Movistar y Claro; en este contexto es relevante para los operadores tener inteligencia competitiva de promociones tal como se plantea en esta tesis.

[38] https://www.osiptel.gob.pe/articulo/24-lineas-en-servicio-por-empresa

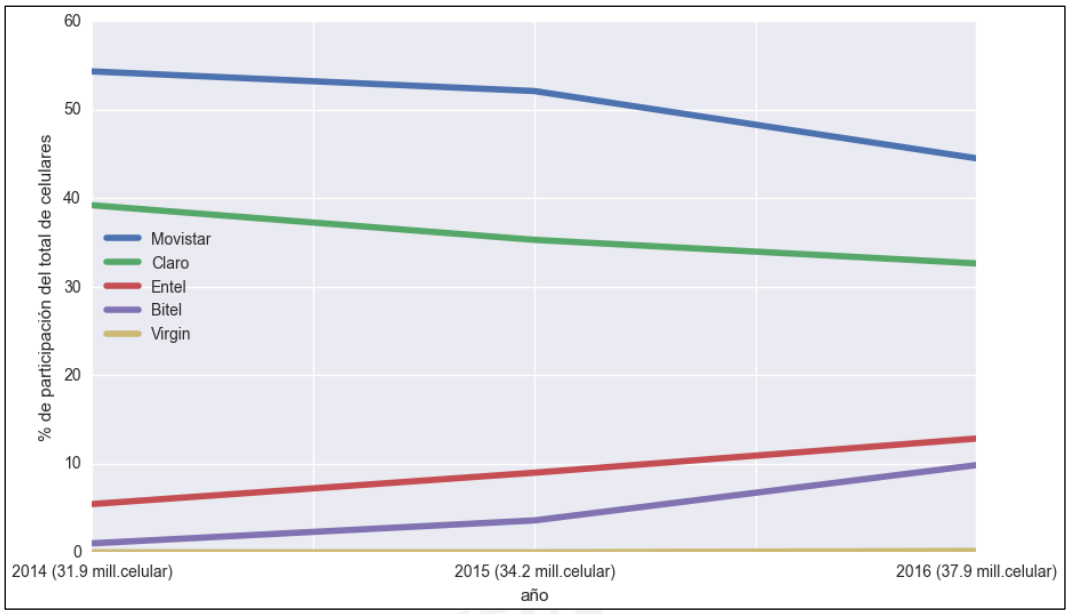


Figura 9: Evolución de participación de mercado por operador desde 2014

2.7. Alcance

Esta tesis es de aplicación práctica y corresponde al área de investigación de ciencias de la computación denominado ingeniería ontológica y al área de investigación de inteligencia competitiva que es una especialización de vigilancia tecnológica que contribuye a la investigación, desarrollo e innovación (I+D+i).

Se implementará un proceso de inteligencia competitiva que inicia con la extracción de todas las publicaciones o *posts* y sus comentarios de Facebook de los 5 operadores desde el primer *post* y comentario publicado hasta la fecha en que se ejecute el proceso de extracción, se creará una ontología de dominio en telecomunicaciones para clasificar por su significado cada *post*, se creará una ontología de dominio en polaridad para clasificar por su polaridad cada comentario, se implementará una aplicación web que permita al operador realizar búsquedas por uno o varios conceptos en el dominio de telecomunicaciones para responder a preguntas relevantes, finalmente se implementará una aplicación web que compare los resultados encontrados de los *posts* y comentarios por operador generando la inteligencia competitiva al operador.

Para la elaboración de la ontología de dominio en telecomunicaciones, se usarán los conceptos más relevantes utilizados en promociones por los operadores desde el punto de vista del autor de esta tesis que está familiarizado por más de 10 años en los conceptos del dominio en telecomunicaciones.

No se incluye el uso de software de base de datos, todo el corpus será almacenado en archivos de texto en formato CSV. La aplicación web a implementar en *microframework* de Python flask no se incluye su despliegue en un servidor web de producción como nginx o apache, podrá ser usada en entorno de desarrollo/test.

3. ESTADO DEL ARTE

Para validar el enfoque dado a esta tesis y verificar su aporte se hizo una revisión del estado del arte mediante una revisión sistemática de la literatura que es un proceso formal y repetible para identificar, evaluar e interpretar toda la investigación disponible relacionada con una pregunta de investigación (Kitchenham y Charters 2007). La revisión sistemática tiene las siguientes fases:

- **Planificación:** Se elabora el protocolo de revisión sistemática que contiene las preguntas de investigación, los términos de búsqueda, selección de fuentes, criterios de inclusión y exclusión, criterios de estimación de calidad y la descripción de cómo se realizará el proceso de selección y extracción de datos.
- **Ejecución:** Se realiza la búsqueda en las fuentes de datos y se hace la selección inicial y final de las investigaciones según lo indicado en el protocolo de revisión sistemática.
- **Resultados:** Se muestran los resultados obtenidos a partir de las investigaciones seleccionadas y se da respuesta a las preguntas de investigación indicadas en el protocolo de revisión sistemática.

3.1. Protocolo de Revisión Sistemática

3.1.1. Preguntas de Investigación

En esta revisión sistemática se busca responder a la siguiente pregunta de investigación principal:

- ¿De qué manera se está aplicando ontologías de dominio en el proceso de inteligencia competitiva de empresas a partir de redes sociales o de datos públicos en internet?

A partir de la pregunta de investigación principal se formularon las siguientes preguntas específicas:

- **PE1:** ¿Qué proceso o fases fueron propuestos para la inteligencia competitiva y qué métodos y procedimientos fueron definidos en cada fase?
- **PE2:** ¿Qué características tenían las ontologías de dominio propuestas?
- **PE3:** ¿Cómo se planteó utilizar las ontologías de dominio para dar semántica a los datos y contribuir en la inteligencia competitiva?
- **PE4:** ¿Qué fuentes de datos fueron propuestas como entrada al proceso de inteligencia competitiva?

3.1.2. Definición de los términos de búsqueda

Se realizaron búsquedas preliminares utilizando varias combinaciones de términos de búsqueda derivadas de la pregunta de investigación principal. Las búsquedas preliminares se hicieron en EBSCO Discovery Service [39], para tener un estimado de la cantidad de investigaciones existentes y revisar el título y resumen de un primer grupo de investigaciones. En la Tabla 13 se muestran los términos de búsqueda principales utilizados y los alternativos o relacionados encontrados.

Tabla 13: Términos de búsqueda

Término de búsqueda principal	Términos de búsqueda alternativos o relacionados
"competitive intelligence"	"intelligence analysis"

[39] http://search.ebscohost.com - EBSCO Discovery Service (EDS) es un índice de recursos de información que permite recuperar, a través de una plataforma de búsqueda unificada, los contenidos de colecciones suscritas por la Biblioteca de la Pontificia Universidad Católica del Perú, así como portales y repositorios gratuitos de acceso abierto.

ontology	ontological, ontologies
"social network"	"social media", "public sources", "public resources", web, internet

3.1.3. Selección de fuentes y Documentación del proceso de búsqueda

Para las búsquedas se seleccionaron las siguientes fuentes a las cuales tiene acceso la Pontificia Universidad Católica del Perú: ACM Digital Library [40], IEEE Xplore [41], Web of science [42], EBSCO Discovery Service [43], Proquest [44] y Springer [45]. Se documentó el proceso de búsqueda para que pueda ser replicable con la siguiente información: Nombre de base de datos, Cadena de búsqueda, Fecha de búsqueda y Número de resultados.

3.1.5. Criterios de inclusión y exclusión

Se incluyeron las investigaciones que cumplieron con todos los siguientes criterios de inclusión:

- Que en la investigación se aplique un proceso de inteligencia competitiva.
- Que en la investigación se use ontologías de dominio para darle significado (semántica) a los datos extraídos.
- Que en la investigación se extraigan los datos de preferencia de redes sociales o en su defecto de fuentes públicas disponibles en la web o internet.
- Que la investigación esté escrita en inglés, español o portugués.

Se excluyeron las investigaciones que cumplieron con los siguientes criterios de exclusión:

- La investigación es menos detallada con respecto a otra investigación de los mismos autores. Se selecciona la investigación con mayor detalle, o si hubiera poca diferencia, se selecciona la primera en haber sido publicada.

3.1.6. Criterios de estimación de calidad

Para evaluar la calidad de las investigaciones se elaboró una lista de preguntas combinando preguntas generales que aplican a cualquier investigación y específicas que aplican a las preguntas de investigación.

Preguntas generales:

1. ¿Se indica claramente el objetivo de la investigación?
2. ¿Se presenta un caso de estudio, experimento o prototipo replicable?
3. ¿Se indican las limitaciones de la investigación?
4. ¿La discusión de resultados y conclusiones son coherentes y guardan relación con el objetivo de la investigación?

Preguntas específicas:

5. ¿Se indica y sustenta claramente el proceso o fases de la inteligencia competitiva?

[40] http://dl.acm.org/

[41] http://ieeexplore.ieee.org

[42] http://isiknowledge.com

[43] http://search.ebscohost.com

[44] https://search.proquest.com/

[45] http://link.springer.com/

6. ¿Se indican claramente las características de las ontologías de dominio utilizadas?
7. ¿Se indica claramente cómo se aplican las ontologías de dominio para darle semántica a los datos?
8. ¿Se indican las fuentes de datos utilizadas como entrada para el proceso de inteligencia competitiva?

A cada pregunta se le asignó tres opciones de respuesta de acuerdo a la siguiente escala: Sí (1 punto), No (0 puntos), Parcialmente (0.5 puntos). Los puntajes totales de calidad pueden estar entre 0 puntos (pésima calidad) y 8 puntos (óptima calidad). El umbral aceptable de calidad definido para seleccionar una investigación es de 4 puntos como mínimo.

3.1.7. Proceso de selección y Proceso de extracción de datos

El proceso de selección se realizó en dos fases:

- En la **fase de selección inicial,** luego de realizadas las búsquedas en las bases de datos y eliminadas las investigaciones duplicadas, se revisaron los títulos y resúmenes de las investigaciones y se eliminaron aquellas que claramente no cumplían con los criterios de inclusión y exclusión.
- En la **fase de selección final** se recuperó y revisó el texto completo de las investigaciones restantes para validar el cumplimiento de los criterios de inclusión y exclusión, además se validaron los criterios de estimación de calidad excluyendo las investigaciones con puntaje menor al umbral aceptable de calidad establecido en el protocolo de revisión sistemática.

Para las investigaciones seleccionadas en la fase de selección final se extrajo y completó la información indicada en la Tabla 14 que comprende datos generales y datos específicos que contribuyan a elaborar la síntesis y poder dar respuesta a las preguntas de investigación.

Tabla 14: Formulario de extracción de datos

Campo	Descripción	Tipo
Id	Identificador de investigación en formato I[número]. Ejemplo: I001	Datos generales
Título	Título de la investigación	Datos generales
Tipo de fuente	Revista, congreso o capítulo de libro	Datos generales
Fuente	Nombre de la revista, congreso o libro	Datos generales
Autor(es)	Autores de publicación	Datos generales
Año	Año de publicación	Datos generales
País	País de procedencia de los autores	Datos generales
Proceso de inteligencia competitiva	¿Qué proceso o fases fueron propuestos para la inteligencia competitiva y qué métodos y procedimientos fueron definidos en cada fase?	Pregunta de investigación 1
Características de ontologías de dominio	¿Qué características tenían las ontologías de dominio propuestas?	Pregunta de investigación 2
Semántica de datos con ontologías de dominio	¿Cómo se planteó utilizar las ontologías de dominio para dar semántica a los datos y contribuir en la inteligencia competitiva?	Pregunta de investigación 3
Fuentes de datos para inteligencia competitiva	¿Qué fuentes de datos fueron propuestas como entrada al proceso de inteligencia competitiva?	Pregunta de investigación 4

3.2. Ejecución de la Revisión Sistemática

3.2.1. Búsqueda

En la Tabla 15 se muestran los resultados obtenidos luego de ejecutar cada una de las búsquedas en las diferentes bases de datos seleccionadas como fuentes. Se obtuvieron en total 366 investigaciones.

Tabla 15: Resultados de búsqueda

Nombre base datos	Cadena de búsqueda	Fecha de búsqueda	Nº de resultados
ACM Digital Library	(acmdlTitle:("competitive intelligence" OR "intelligence analysis") OR recordAbstract:("competitive intelligence" OR "intelligence analysis") OR keywords.author.keyword:("competitive intelligence" OR "intelligence analysis")) **AND** ontolog* **AND** ("social network" OR "social media" OR "public sources" OR "public resources" OR web OR internet)	10/Jul/2017	5
IEEE Xplore	(("Document Title":"competitive intelligence" OR "Abstract":"competitive intelligence" OR "Index Terms":"competitive intelligence" OR "Document Title":"intelligence analysis" OR "Abstract":"intelligence analysis" OR "Index Terms":"intelligence analysis") **AND** ontolog* **AND** ("social network" OR "social media" OR "public sources" OR "public resources" OR web OR internet))	10/Jul/2017	103
Web of science	("competitive intelligence" OR "intelligence analysis") **AND** ontolog* **AND** ("social network" OR "social media" OR "public sources" OR "public resources" OR web OR internet)	11/Jul/2017	13
EBSCO Discovery Service	(TI "competitive intelligence" OR AB "competitive intelligence" OR SU "competitive intelligence" OR TI "intelligence analysis" OR AB "intelligence analysis" OR SU "intelligence analysis") **AND** ontolog* **AND** ("social network" OR "social media" OR "public sources" OR "public resources" OR web OR internet)	10/Jul/2017	101 [46]
Proquest	(ti("competitive intelligence" OR "intelligence analysis") OR ab("competitive intelligence" OR "intelligence analysis") OR if("competitive intelligence" OR "intelligence analysis")) **AND** ontolog* **AND** ("social network" OR "social media" OR "public sources" OR "public resources" OR web OR internet)	10/Jul/2017	31
Springer	("competitive intelligence" OR "intelligence analysis") **AND** ontolog* **AND** ("social network" OR "social media" OR "public sources" OR "public resources" OR web OR internet)	10/Jul/2017	113
		Total	**366**

3.2.2. Selección inicial y Selección final

Se eliminaron en total 62 investigaciones duplicadas quedando 304 investigaciones a las cuales se revisaron los títulos y resúmenes y se eliminaron 254 que claramente no cumplían con los criterios de inclusión y exclusión quedando 50 investigaciones tal como se aprecia en la Figura 10. Se revisó el texto completo de las 50 investigaciones de la selección inicial, validando el cumplimiento de los criterios de inclusión y exclusión y los criterios de estimación de calidad seleccionando las investigaciones con puntaje mayor o igual al umbral aceptable de calidad del protocolo de revisión sistemática quedando 19 investigaciones tal como se aprecia en la Figura 10.

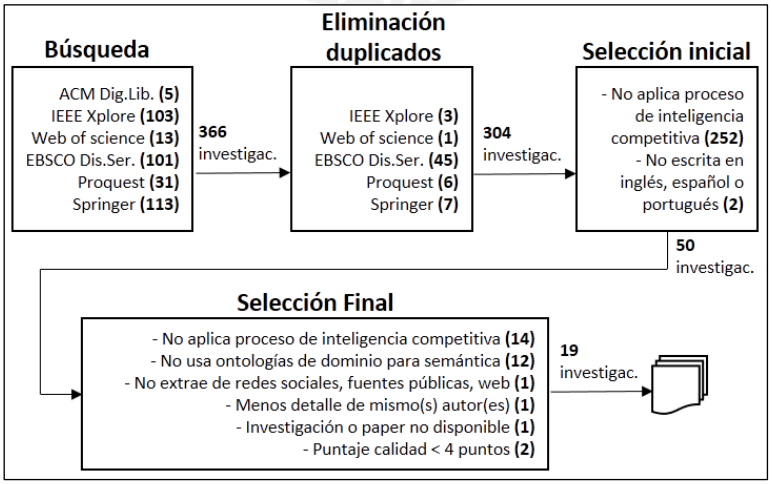


Figura 10: Búsqueda y selección de investigaciones

[46] La búsqueda en EBSCO Discovery Service devolvió 144 resultados, pero sólo exportó 101 porque indica "Nota: Las repeticiones exactas se eliminaron de los resultados"

En la selección final se obtuvieron 19 investigaciones las cuales se listan en la Tabla 16 con su año de publicación, autor(es), puntaje de calidad y país(es) de procedencia ordenado de forma descendente por año de publicación. En la Tabla 17 se muestran las 19 investigaciones seleccionadas por bases de datos.

Tabla 16: Investigaciones seleccionadas ordenadas por año descendente

ID	Año	Autor(es)	Puntaje de calidad	País(es) de procedencia
i01	2017	(Hassan et al. 2017)	7.0	Francia
i02	2015	(Abdellaoui y Nader 2015)	4.5	Argelia
i03	2015	(Spruit y Cepoi 2015)	8.0	Países bajos
i04	2015	(Vasilateanu et al. 2015)	6.0	Rumania
i05	2015	(Wongthongtham y Abu-Salih 2015)	6.5	Australia
i06	2014	(Chakraborti y Dey 2014)	5.0	India
i07	2014	(Chouder y Chalal 2014)	5.0	Argelia
i08	2014	(Olszak 2014)	5.0	Polonia
i09	2011	(Dai et al. 2011)	6.5	Finlandia
i10	2011	(Del-Fresno-García 2011)	4.5	España
i11	2011	(Liu et al. 2011)	4.0	China
i12	2010	(Jin y Yan 2010)	4.5	China
i13	2009	(Liu y He 2009)	5.0	China
i14	2009	(Zhao y Jin 2009)	5.0	China
i15	2008	(Chen et al. 2008)	5.0	China
i16	2008	(Nagano et al. 2008)	6.0	Japón
i17	2008	(Nemrava et al. 2008)	5.0	República Checa
i18	2007	(Li et al. 2007)	8.0	China
i19	2007	(Zhang et al. 2007)	4.5	China

Tabla 17: Investigaciones seleccionadas por base de datos

Nombre base datos	Resultados	Duplicados	Seleccionados
ACM Digital Library	5	0	3
IEEE Xplore	103	3	8
Web of science	13	1	0
EBSCO Discovery Service	101	45	8
Proquest	31	6	0
Springer	113	7	0
		Total	**19**

3.2.4. Extracción de datos

Para las 19 investigaciones se leyó el texto completo y se extrajo la información relevante en el formato de formulario de extracción de datos (Tabla 14) definido en el protocolo de revisión sistemática y se realizó la síntesis de la información para poder dar respuesta a las preguntas de investigación las cuales se detallan en la sección 3.3.

3.3. Resultados de la Revisión Sistemática

3.3.1. Proceso de inteligencia competitiva (PE1)

Para la primera pregunta específica de investigación ¿Qué proceso o fases fueron propuestos para la inteligencia competitiva y qué métodos y procedimientos fueron definidos en cada fase? se tiene que:

Respecto a las fases del proceso de inteligencia competitiva propuestas, estas se indican explícitamente en 6 (i02, i03, i07, i08, i15 y i18) de 19 investigaciones. En la Tabla 18 se muestra un mapeo con las fases del proceso de inteligencia competitiva propuestas por Arroyo Varela (2005) que se tomó como referencia para elaborar esta tesis.

Tabla 18: Fases propuestas para proceso de Inteligencia Competitiva en 6 investigaciones

Fase	Arroyo Varela (2005)	i02 (Abdellaoui y Nader 2015)	i03 (Spruit y Cepoi 2015)	i07 (Chouder y Chalal 2014)	i08 (Olszak 2014)	i15 (Chen et al. 2008)	i18 (Li et al. 2007)
1	Identificación de necesidades de información	Planificación y dirección		Definición de las necesidades	Planificación y enfoque	Dirección	Dirección
2	Recogida de información	Recolección	Recolección y almacenamiento	Recolección de información	Recolección	Recolección	Recolección
3	Análisis y síntesis de la información	Análisis	Análisis	Análisis de la información	Análisis	Análisis	Análisis
4	Reparto de la inteligencia a los tomadores de decisiones	Diseminación	Diseminación de información	Difusión	Comunicación	Diseminación	Diseminación
5		Retroalimentación		Retroalimentación			

En la Tabla 18 se puede observar que las 6 investigaciones (i02, i03, i07, i08, i15 y i18) tienen tres fases en común que son Recolección, Análisis y Diseminación (sombreadas en gris) que son similares a la segunda, tercera y cuarta fases propuestas por Arroyo Varela (2005) y usadas en esta tesis. Respecto a la primera fase tiene diferentes nombres como planificación y dirección (i02), planificación y enfoque (i08) y dirección (i15, i18), en esta tesis usaremos el nombre "Identificación de necesidades de información" tal como lo propone Arroyo Varela (2005) y utiliza un nombre parecido "Definición de las necesidades" (i07), en esta primera fase los tomadores de decisiones del operador pueden identificar sus necesidades de información mediante preguntas relevantes en el dominio de las telecomunicaciones.

En la Tabla 18 también se observa que dos de las investigaciones (i02, i07) tienen una fase de Retroalimentación; Abdellaoui y Nader (2015) indican que la retroalimentación evalúa el impacto del conocimiento obtenido del ciclo de Inteligencia Competitiva para seguir mejorando, mientras que Chouder y Chalal (2014) indican que la Retroalimentación genera un bucle con la Fase 1 de definición de necesidades de información. En esta tesis no se usará una fase 5 sino que se definirá que la fase 4 de "Diseminación" genere como salida nuevas necesidades de información que ingresen a la fase 1 de "Identificación de necesidades de información" como lo propone Gógova (2015). En la Figura 11 se muestra el proceso de inteligencia competitiva de 4 fases a utilizar en esta tesis.

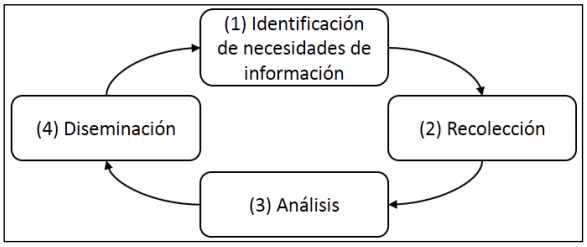


Figura 11: Proceso de inteligencia competitiva

Respecto a los métodos y procedimientos definidos en cada fase, en la Tabla 19 se muestra un resumen de los métodos y procedimientos de la fase de recolección (Figura 11 (2)) que se mencionan explícitamente en 9 (i01, i03, i04, i05, i09, i11, i12, i16 y i18) de 19 investigaciones ordenado descendente por frecuencia donde se aprecia que el 78% utiliza rastreadores de páginas web para la recolección de datos y el 22% usa APIs específicas de la fuente de datos como Twitter.

Tabla 19: Métodos y procedimiento de fase de recolección en 9 investigaciones

Métodos y Procedimientos Fase Recolección	Frecuencia	Porcentaje de frecuencia	Investigaciones
Rastreador de páginas web (Web Crawler, Spider)	7	78%	i01, i03, i04, i11, i12, i16, i18
API de la fuente de datos (Ejemplo: Twitter API)	2	22%	i05, i09
		100%	

En la Tabla 20 se muestra un resumen de los métodos y procedimientos de la fase de análisis (Figura 11 (3)) que se mencionan explícitamente en las 19 investigaciones como mínimo dos veces, ordenado descendente por frecuencia donde se aprecia que los más usados son la ontología (95%), seguido por la identificación de temas (58%), etiquetado gramatical (26%), análisis de polaridad (26%), reconocimiento de entidades nombradas (21%) y Web mining (21%) y el resto con menor porcentaje.

Tabla 20: Métodos y procedimientos de fase de análisis en 19 investigaciones

Métodos y Procedimientos Fase Análisis	Frecuencia	Porcentaje de frecuencia	Investigaciones
Ontología	18	95%	i01, i02, i03, i04, i05, i06, i07, i08, i09, i11, i12, i13, i14, i15, i16, i17, i18, i19
Identificación de temas o conceptos (*Topic detection, Topic tagging, Topic mapping*)	11	58%	i01, i06, i07, i09, i10, i11, i12, i13, i16, i17, i18
Etiquetado gramatical (*Part-of-speech tagging*)	5	26%	i03, i04, i06, i07, i15
Análisis de polaridad (*Sentiment analysis, Opinion mining*)	5	26%	i05, i07, i08, i09, i16
Reconocimiento de entidades nombradas (*Named Entity Recognition – NER*)	4	21%	i03, i05, i06, i09
Web mining, Data mining, Text mining, Semantic mining	4	21%	i07, i08, i10, i19
Herramienta para crear ontología: Protégé	3	16%	i04, i11, i15
Detección de eventos (*Event detection, Event extraction*)	2	11%	i08, i09
Event timeline analysis	2	11%	i07, i09
Bases de datos léxicas (*WordNet, HowNet*)	2	11%	i04, i12
Herramienta para *Named Entity Recognition*: GATE	2	11%	i04, i05
Almacenamiento de datos en data warehouse	2	11%	i02, i05

En la Tabla 21 se muestra un resumen de los métodos y procedimientos de la fase de diseminación (Figura 11 (4)) que se mencionan explícitamente en 10 (i02, i03, i05, i07, i08, i09, i12, i16, i18 y i19) de 19 investigaciones ordenado descendente por frecuencia donde se aprecia que los más usados son los dashboards (40%), las búsquedas (30%) y los reportes (30%) y el resto con menor porcentaje.

Tabla 21: Métodos y procedimientos de fase de diseminación en 10 investigaciones

Métodos y Procedimientos Fase Diseminación	Frecuencia	Porcentaje de frecuencia	Investigaciones
Tablero con información gráfica más importante (*Dashboard*)	4	40%	i02, i05, i07, i08
Búsquedas (*Queries*)	3	30%	i03, i12, i18
Reportes / Excel	3	30%	i02, i05, i07
Estadísticas	2	20%	i02, i18
Gráficos	2	20%	i09, i19
Aplicación móvil / web	2	20%	i07, i16
Monitoreo y alertas	1	10%	i03
SMS / Correo electrónico	1	10%	i07
Interactive visualization tools, Balance scorecard, Service-Oriented Architecture (SOA)	1	10%	i08

3.3.2. Ontologías de dominio (PE2 y PE3)

Para la segunda pregunta específica de investigación ¿Qué características tenían las ontologías de dominio propuestas? y para la tercera pregunta ¿Cómo se planteó utilizar las ontologías de dominio para dar semántica a los datos y contribuir en la inteligencia competitiva? se tiene que:

En la Tabla 22 se muestra un resumen de las características de las ontologías de dominio propuestas en 18 de 19 investigaciones ordenado descendente por frecuencia donde se aprecia que la creación manual de ontología con el apoyo de expertos es la más utilizada con 61%, seguido con el mismo porcentaje de 17% por la creación manual a partir de ontologías existentes y la creación automática de ontología, finalmente se tiene la creación semiautomática de ontología con 5%.

Tabla 22: Características de ontologías de dominio propuestas en 18 investigaciones

Característica	Frecuencia	Porcentaje de frecuencia	Investigaciones
Creación manual de ontología con apoyo de expertos	11	61%	i02, i03, i06, i07, i08, i09, i11, i13, i14, i16, i17
Creación manual de ontología a partir de ontologías existentes	3	17%	i05, i12, i18
Creación automática de ontología	3	17%	i01, i04, i19
Creación semiautomática de ontología	1	5%	i15
		100%	

En la Tabla 23 se muestra un resumen del uso de las ontologías de dominio para dar semántica a los datos en 18 de 19 investigaciones ordenado descendente por frecuencia donde se aprecia que el mayor uso es para la identificación de temas o conceptos (56%) seguido por el reconocimiento de entidades nombradas (33%) y en menor uso para etiquetado gramatical (17%), detección de eventos (11%), análisis de polaridad (11%) y para expandir términos de búsqueda (6%).

Tabla 23: Uso de las ontologías de dominio para dar semántica a los datos en 18 investigaciones

Uso de ontología en textos en lenguaje natural para:	Frecuencia	Porcentaje de frecuencia	Investigaciones
Identificación de temas o conceptos (*Topic detection, Topic tagging*)	10	56%	i01, i02, i04, i06, i11, i13, i15, i17, i18, i19
Reconocimiento de entidades nombradas (*Named Entity*	6	33%	i03, i05, i06, i09, i14,

Recognition – NER)			i16
Etiquetado gramatical (*Part-of-speech tagging*)	3	17%	i04, i06, i15
Detección de eventos (*Event detection*)	2	11%	i08, i09
Análisis de polaridad (*Sentiment analysis, Opinion mining*)	2	11%	i05, i16
Expandir o agregar términos de búsqueda (*Query expansion*)	1	6%	i12
No especificado	1	6%	i07

3.3.3. Fuentes de datos para inteligencia competitiva (PE4)

Para la cuarta pregunta específica de investigación ¿Qué fuentes de datos fueron propuestas como entrada al proceso de inteligencia competitiva? se tiene que:

En la Tabla 24 se muestra un resumen de las fuentes de datos de 19 investigaciones ordenado de forma descendente por frecuencia donde se puede apreciar que las fuentes de datos más utilizadas son las noticias (42%), los *blogs* (37%), las redes sociales (32%) y las páginas web (32%) y en menor frecuencia con 21% a menos el resto de fuentes de datos.

Tabla 24: Frecuencia de uso fuentes de datos para inteligencia competitiva en 19 investigaciones

Fuente de datos	Frecuencia	Porcentaje de frecuencia	Investigaciones
Noticias	8	42%	i01, i02, i03, i05, i06, i08, i10, i19
Blogs	7	37%	i05, i06, i07, i09, i10, i11, i16
Redes sociales	6	32%	i02, i05, i07, i08, i09, i10
Páginas web	6	32%	i07, i10, i13, i14, i15, i18
Internet	4	21%	i02, i03, i07, i12
Websites de opiniones	3	16%	i06, i07, i10
Wikis	3	16%	i02, i09, i14
Fuentes públicas	3	16%	i04, i08, i17
RSS feeds	2	11%	i02, i07
Media	2	11%	i02, i03
Bases de datos de patentes, científicas	2	11%	i02, i18
Intranet	2	11%	i02, i04
Reportes anuales	1	5%	i06

4. RECOLECCIÓN DE DATOS

4.1. Introducción

Para tener inteligencia competitiva se necesita recolectar las promociones publicadas y sus comentarios, que son de dominio público, de las páginas de Facebook de los 5 operadores. En la Figura 12 se muestra en líneas entrecortadas las 4 fases del proceso de inteligencia competitiva que inicia con la fase de identificación de necesidades de información (1), luego por la fase de recolección (2) que está sombreada en gris y se detalla en este capítulo 4 y que genera el corpus que servirá para la fase de análisis (3) que se describe en capítulos 5 y 6 y posteriormente la fase de diseminación (4) que se describe en capítulo 7 para que los tomadores de decisiones del operador tengan información relevante y se genere la inteligencia competitiva.

Este capítulo 4 comprende la descripción y discusión de los resultados alcanzados para el objetivo específico 1 "**OE.1:** Extraer y limpiar todas las publicaciones (*posts*) realizadas por los 5 operadores en sus páginas de Facebook, incluyendo los comentarios publicados por sus seguidores, para generar el corpus".

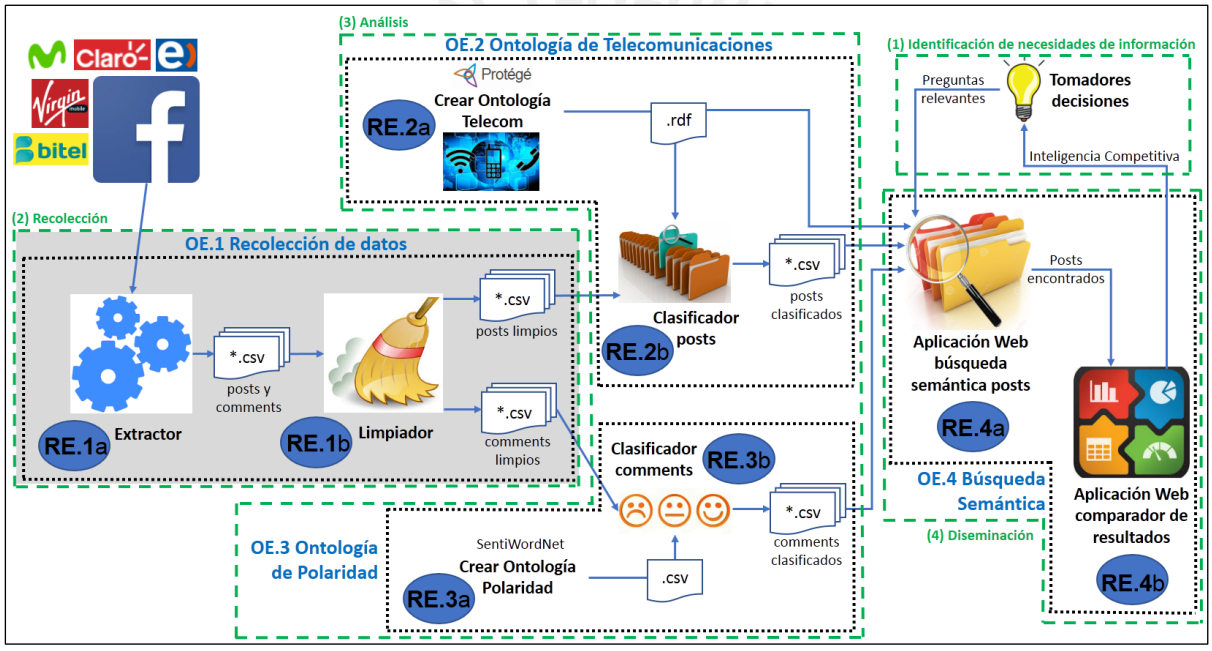


Figura 12: Recolección de datos

4.2. Resultados alcanzados

4.2.1. *Posts*

Respecto al resultado esperado "**RE.1a:** Proceso automático de extracción de *posts* y sus comentarios de páginas de Facebook de los 5 operadores", en la Figura 13 se muestra el diagrama de flujo del proceso de extracción de *posts* para el operador Movistar que permite obtener todos los *posts* de Movistar desde la fecha de extracción hasta el primer *post* que publicó en Facebook, este diagrama también aplica a los demás operadores sólo reemplazando los parámetros de entrada `node_id` y `csv_id` por los datos del operador.

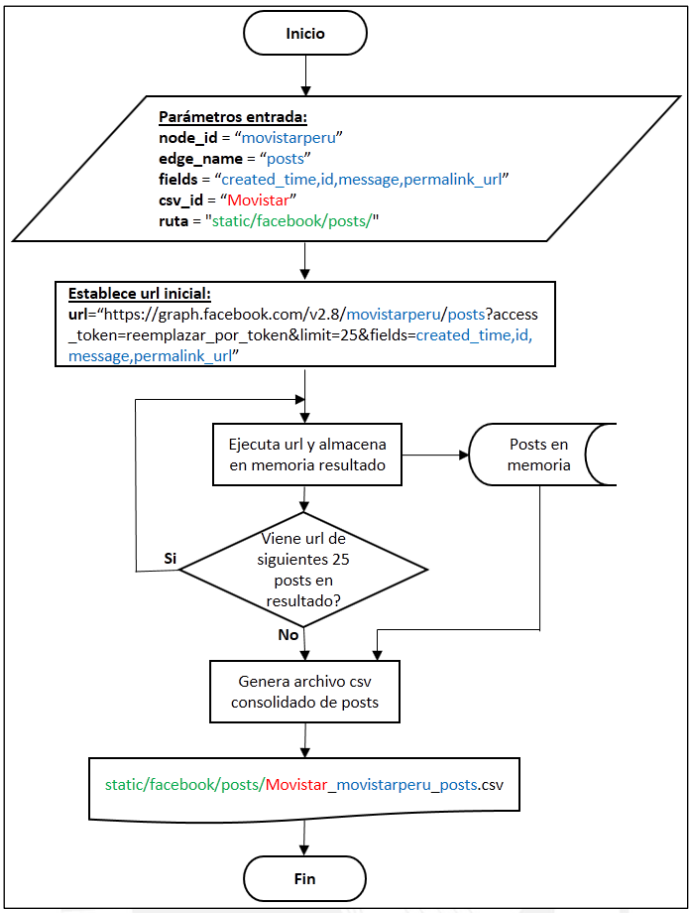


Figura 13: Diagrama de flujo del proceso de extracción de *posts* de un operador

El `node_id` de cada operador, usado en url, se obtuvo de su página de Facebook tal como se aprecia en la Figura 14 para "movistarperu". El `csv_id` es el nombre del operador que aparecerá en el nombre del archivo CSV generado.

Figura 14: Página Facebook de Movistar

A partir del diagrama de flujo de Figura 13 se implementó un software o programa en Python de una librería genérica para extraer datos de Facebook que se muestra en Figura 15.

```python
# Librerías
import requests as rq # Librería para HTTP
import pandas as pd # Librería para DataFrames
import datetime as dt # Librería para fecha y hora
import sys # Para mostrar errores en try except

# Parámetros Globales
token = "EAAF3mZA0TlokBAAhkCzRGrXq2jlPmiOireJcYbdtpodymGFULdYuo7yvEHFQs65ejDHZCx81Ru7w5RkZCZC61JKRc88XcxyVQ..."
url_fb_api = "https://graph.facebook.com/"
version_api = "v2.8" # Esta es la última versión del 5-Octubre-2016
limite_reg = "25" # Límite de registros a devolver en la URL (Recomendable 25)

# Procedimiento para obtener todos los edge-name de un node-id usando Facebook Graph Api y almacena en csv
def facebook_api(node_id, edge_name, fields, csv_id, ruta):
    print("{} - {} - {}".format(csv_id, node_id, edge_name))
    # URL de primera llamada
    url_next = url_fb_api+version_api+"/"+node_id+"/"+edge_name+"?access_token="+token+"&limit="+limite_reg+fields
    # Hacer un loop para obtener todos los edge_name hasta que ya no haya next
    df_edge_name_total = pd.DataFrame() # Inicializa DataFrame vacio
    i = 1
    fecha_hora_ini = dt.datetime.now()
    print("Fecha y hora Inicial          = %s" % fecha_hora_ini)
    while True:
        try:
            url = url_next
            resultado_edge_name = rq.get(url)
            json_edge_name = resultado_edge_name.json() # Formato Diccionario para Python
            df_edge_name = pd.DataFrame(json_edge_name["data"])
            df_edge_name_total = pd.concat([df_edge_name_total,df_edge_name]) # Concateno en un solo DataFrame
            url_next = json_edge_name["paging"]["next"] # Cuando no encuentre next irá por excepción y termina loop
            i = i + 1
        except:
            print("Se llegó al fin de {} con i= {} por {}".format(edge_name,i,sys.exc_info()[0])) # Detalle excepción
            break
    fecha_hora_fin_ext = dt.datetime.now()
    print("Fecha y hora Final extracción = %s" % fecha_hora_fin_ext)
    # Graba en un solo archivo csv. Se usa line_terminator "~" porque hay enter en algunos post
    nombre_csv = csv_id + "_" + node_id + "_" + edge_name + ".csv"
    df_edge_name_total.to_csv(ruta + nombre_csv, index=False, encoding="utf-8", sep="|", line_terminator="~")
    fecha_hora_fin_csv = dt.datetime.now()
    print("Fecha y hora Final csv        = %s\n" % fecha_hora_fin_csv)
```

Figura 15: Librería genérica creada en python para extraer datos de Facebook

Se implementó un programa en Python, mostrado en Figura 16, para automatizar la extracción de todos los *posts* de los 5 operadores que usa la librería genérica de Facebook de Figura 15.

```python
# Librerías
import facebook_api as fb # Librería creada para facebook

# Parámetros Globales
fields_posts = "&fields=created_time,id,message,permalink_url"
ruta_posts = "static/facebook/posts/"

# Obtener posts de Facebook
fb.facebook_api("movistarperu", "posts", fields_posts, "Movistar", ruta_posts)
fb.facebook_api("AmericaMovilPeruSAC", "posts", fields_posts, "Claro", ruta_posts)
fb.facebook_api("EntelPeru", "posts", fields_posts, "Entel", ruta_posts)
fb.facebook_api("bitelperu", "posts", fields_posts, "Bitel", ruta_posts)
fb.facebook_api("VirginMobilePe", "posts", fields_posts, "Virgin", ruta_posts)
```

Figura 16: Programa en python para extraer todos los *posts* de los 5 operadores

Se ejecutó el programa en Python de la Figura 16 el 16/May/2017 en una computadora de escritorio Intel Core i5 con 4GB de RAM con Windows 7 de 64 bits y conexión de internet de banda ancha fija de 8 MB. La duración total de procesamiento fue de aproximadamente 16 minutos obteniendo los resultados mostrados en la Tabla 25.

Tabla 25: *Posts* extraídos por operador

Operador	*Posts* extraídos	Tamaño archivo CSV
Movistar	6,086	1.91 MB
Claro	7,753	2.26 MB
Entel	980	0.27 MB
Bitel	876	0.23 MB

Virgin	168	0.04 MB
Total	**15,863**	**4.71 MB**

Respecto al resultado esperado "**RE.1b:** Proceso automático de limpieza de *posts* y sus comentarios para generar el corpus", en la Figura 17 se muestra el diagrama de flujo del proceso de limpieza de *posts*.

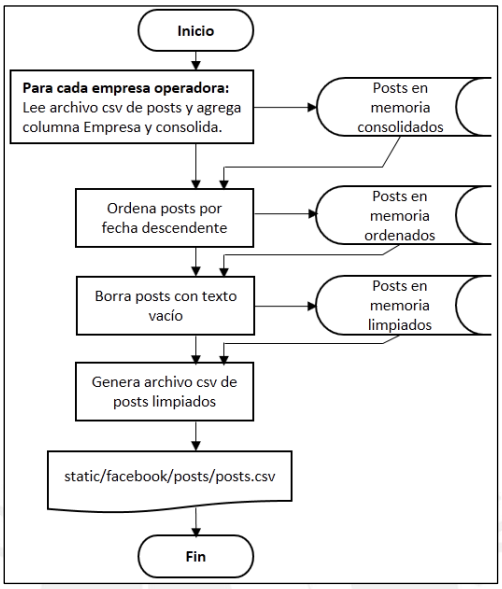


Figura 17: Diagrama de flujo del proceso de limpieza de *posts*

A partir del diagrama de flujo de Figura 17 se implementó un software o programa en Python para automatizar la limpieza de los *posts* de los 5 operadores el cual se muestra en la Figura 18.

```
# Librerías
import pandas as pd # Librería para DataFrames

# Parámetros Globales
ruta_posts = "static/facebook/posts/"

# Procedimiento para limpiar posts de Facebook y genera un sólo archivo csv
def limpia_posts():
    # Inializa DataFrame total
    df_posts_total = pd.DataFrame()
    # Lee archivo csv de posts de cada empresa y almacena en dataframe total
    df_posts = pd.read_csv(ruta_posts + 'Movistar_movistarperu_posts.csv', encoding="utf-8", sep="|", lineterminator="~")
    df_posts["empresa"] = "Movistar" # Adiciona columna para identificar empresa
    df_posts_total = pd.concat([df_posts_total,df_posts])
    df_posts = pd.read_csv(ruta_posts + 'Claro_AmericaMovilPeruSAC_posts.csv', encoding="utf-8", sep="|", lineterminator="~")
    df_posts["empresa"] = "Claro" # Adiciona columna para identificar empresa
    df_posts_total = pd.concat([df_posts_total,df_posts])
    df_posts = pd.read_csv(ruta_posts + 'Entel_EntelPeru_posts.csv', encoding="utf-8", sep="|",lineterminator="~")
    df_posts["empresa"] = "Entel" # Adiciona columna para identificar empresa
    df_posts_total = pd.concat([df_posts_total,df_posts])
    df_posts = pd.read_csv(ruta_posts + 'Bitel_bitelperu_posts.csv', encoding="utf-8", sep="|",lineterminator="~")
    df_posts["empresa"] = "Bitel" # Adiciona columna para identificar empresa
    df_posts_total = pd.concat([df_posts_total,df_posts])
    df_posts = pd.read_csv(ruta_posts + 'Virgin_VirginMobilePe_posts.csv', encoding="utf-8", sep="|",lineterminator="~")
    df_posts["empresa"] = "Virgin" # Adiciona columna para identificar empresa
    df_posts_total = pd.concat([df_posts_total,df_posts])
    # Ordena por Fecha descendente
    df_posts_total.sort_values("created_time", ascending=False, inplace=True)
    # Resetea index
    df_posts_total.reset_index(inplace=True)
    df_posts_total.drop("index", axis=1, inplace=True)
    # Borra posts con texto vacío
    df_posts_total.dropna(subset=["message"], axis=0, inplace=True) # Borra los posts que tienen message NaN
    # Exporta posts
    df_posts_total.to_csv(ruta_posts + "posts.csv", index=False, encoding="utf-8", sep="|", line_terminator="~")

# Limpia posts
limpia_posts()
```

Figura 18: Programa en python para limpiar los *posts* de los 5 operadores

Se ejecutó el programa en Python de la Figura 18 el 16/May/2017 y se obtuvo un solo archivo CSV de tamaño 4.79 MB que contiene todos los *posts* válidos, los resultados se muestran en la Tabla 26.

Tabla 26: *Posts* válidos por operador

Operador	*Posts* extraídos	*Posts* con texto vacío (Limpieza)	*Posts* válidos
Movistar	6,086	82	6,004
Claro	7,753	128	7,625
Entel	980	5	975
Bitel	876	7	869
Virgin	168	7	161
Total	**15,863**	**229**	**15,634**

Respecto a los *posts* válidos se puede apreciar en la Figura 19 que Claro empezó a publicar en Facebook desde el 2009, luego siguió Movistar en 2010, seguido por Entel y Bitel en 2014 y finalmente Virgin en 2016. Entel publicó su primer *post* en Octubre del 2014 y Bitel en Agosto 2014 y se puede apreciar que Claro y Movistar mantuvieron en 2014 una cantidad de *posts* ligeramente inferior al 2013 a pesar del ingreso de 2 nuevos operadores, además se aprecia que 2015 es el primer año donde compiten desde el inicio de año 4 operadores.

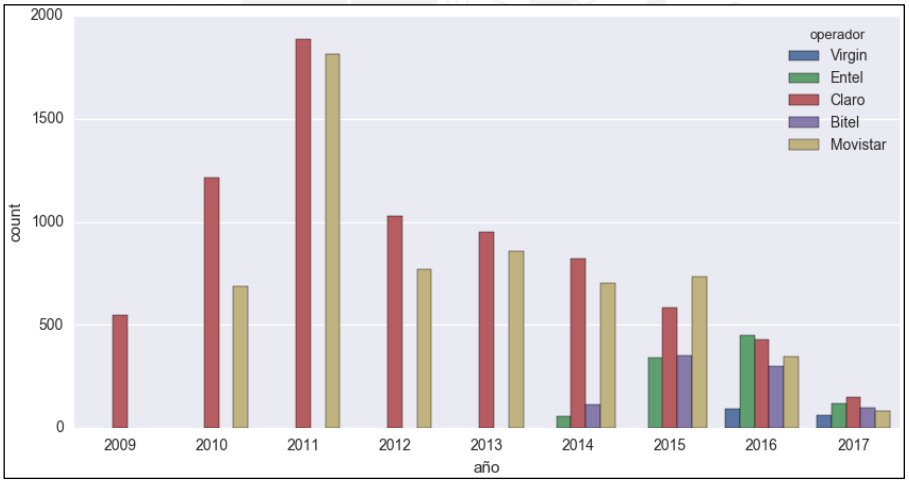


Figura 19: *Posts* válidos publicados en Facebook por operador por año

En la Figura 20 se aprecia que Entel y Bitel publicaron entre 300 a 400 *posts* en 2015, considerando que el año tiene 365 días publicaron en promedio 1 *post* por día mientras que Claro y Movistar prácticamente duplicaron esa cantidad. En 2016 se aprecia una disminución considerable en la cantidad de *posts* publicados por Movistar y Claro y se aprecia una subida considerable de *posts* publicados con promociones por Entel que superan a Movistar y Claro, mientras que Bitel mantiene una cantidad similar de *post* con respecto al año anterior y aparece el quinto operador Virgin que publica su primer *post* en Marzo del 2016.

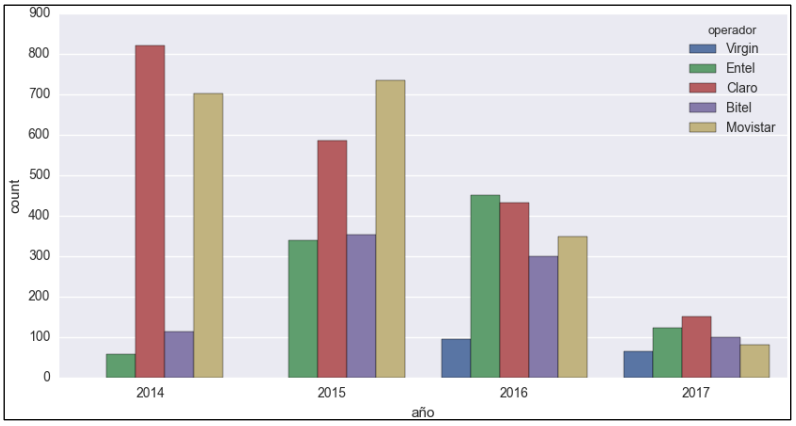


Figura 20: *Posts* válidos publicados en Facebook por operador por año desde 2014

4.2.2. Comentarios

Respecto al resultado esperado "**RE.1a:** Proceso automático de extracción de *posts* y sus comentarios de páginas de Facebook de los 5 operadores", en la Figura 21 se muestra el diagrama de flujo del proceso de extracción de comentarios para un *post* que permite obtener todos los comentarios [47] del *post* desde la fecha de extracción hasta el primer comentario publicado en Facebook, este diagrama también aplica a los demás *posts* sólo reemplazando los parámetros de entrada `node_id` y `csv_id` por el código del *post* y nombre de operador respectivamente.

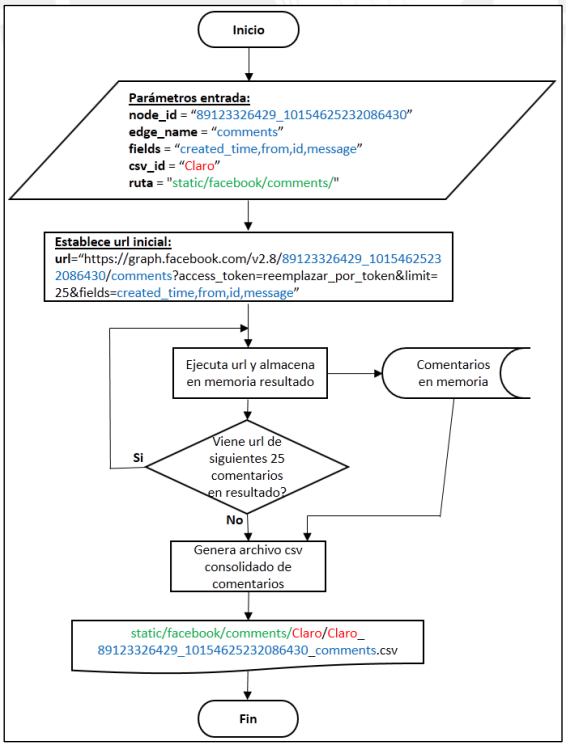


Figura 21: Diagrama de flujo del proceso de extracción de comentarios de un *post*

[47] Proceso extrae todos los comentarios de primer nivel del *post*. No se incluyen los comentarios anidados (respuesta a comentarios).

Para implementar el diagrama de flujo de Figura 21 se reutilizó la librería genérica creada en Python para extraer datos de Facebook mostrada en Figura 15. Luego, se implementó un software o programa en Python para automatizar la extracción de los comentarios de una cantidad configurable de *posts*, a partir del archivo CSV de *posts* válidos, el cual se muestra en la Figura 22.

```
# Librerias
import pandas as pd # Librería para DataFrames
import os.path # Para verificar si existe un archivo
import sys # Para mostrar errores en try except
import facebook_api as fb # Librería creada para facebook

# Parámetros Globales
fields_comments = "&fields=created_time,from,id,message"
ruta_posts = "static/facebook/posts/"
ruta_comments = "static/facebook/comments/"
cantidad_posts = 1000 # posts a procesar para obtener comentarios

# Procedimiento para obtener los comentarios de los posts
def extrae_comments(num_posts):
    # Obtiene los comentarios de los posts
    df_posts = pd.read_csv(ruta_posts + 'posts.csv', encoding="utf-8", sep="|", lineterminator="~")
    lectura = 0
    i = 0
    while lectura < num_posts:
        try:
            empresa = df_posts["empresa"][i]
            id_post = df_posts["id"][i]
            archivo = empresa + "_" + id_post + "_comments.csv"
            ruta = ruta_comments + empresa + "/" + archivo
            if os.path.isfile(ruta) == False:
                lectura = lectura + 1
                print("Leyendo ", lectura, " de ", num_posts, ": ", ruta)
                fb.facebook_api(id_post, "comments", fields_comments, empresa, ruta_comments + empresa + "/")
            i = i + 1
        except:
            print("Se llegó al fin con i= {} por {}".format(i,sys.exc_info()[0])) # Detalle excepción
            break

# Extrae comentarios de la cantidad de posts de variable global
extrae_comments(cantidad_posts)
```

Figura 22: Programa en python para extraer los comentarios de los *posts*

Con el programa de Figura 22 se realizó la extracción de todos los comentarios de los *posts* de los 5 operadores en las fechas 17/May/2017, 18/May/2017, 21/May/2017, 22/May/2017, 23/May/2017, 27/May/2017 y 28/May/2017 con una duración total de aproximadamente 14 horas y media con los siguientes resultados mostrados en Tabla 27.

Tabla 27: Comentarios extraídos por operador

Operador	Comentarios extraídos	Número archivos CSV	Tamaño total
Movistar	450,785	6,004	79.40 MB
Claro	392,992	7,625	87.10 MB
Entel	239,345	975	44.69 MB
Bitel	335,006	869	62.54 MB
Virgin	12,187	161	2.15 MB
Total	**1,430,315**	**15,634**	**275.88 MB**

Respecto al resultado esperado "**RE.1b:** Proceso automático de limpieza de *posts* y sus comentarios para generar el corpus", en la Figura 23 se muestra el diagrama de flujo del proceso de limpieza de comentarios.

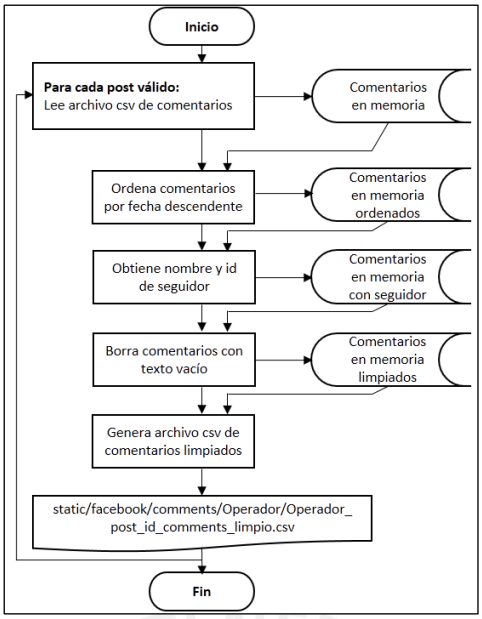


Figura 23: Diagrama de flujo del proceso de limpieza de comentarios

A partir del diagrama de flujo de Figura 23 se implementó un software o programa en Python para automatizar la limpieza de los comentarios de todos los *posts* el cual se muestra en la Figura 24.

```
# Librerias
import pandas as pd
import os.path # Para verificar si existe un archivo
import ast # Para convertir cadena a diccionario
import datetime as dt # Librería para fecha y hora

# rutas
ruta_posts    = "static/facebook/posts/"
ruta_comments = "static/facebook/comments/"

# Obtiene nombre de seguidor de la cadena en formato diccionario
def get_name(cadena):
    try:
        df_from = pd.DataFrame(ast.literal_eval(cadena),index=[0])
        return df_from.ix[0]["name"]
    except:
        return "" # Si recibo NaN devuelvo ""

# Obtiene id de seguidor de la cadena en formato diccionario
def get_name_id(cadena):
    try:
        df_from = pd.DataFrame(ast.literal_eval(cadena),index=[0])
        return df_from.ix[0]["id"]
    except:
        return "" # Si recibo NaN devuelvo ""

# Lee los posts y limpia los comentarios
df_posts = pd.read_csv(ruta_posts + 'posts.csv', encoding="utf-8", sep="|", lineterminator="~")
for i in df_posts.index:
    empresa = df_posts["empresa"][i]
    id_post = df_posts["id"][i]
    archivo_orig   = ruta_comments + empresa + "/" + empresa + "_" + id_post + "_comments.csv"
    archivo_limpio = ruta_comments + empresa + "/" + empresa + "_" + id_post + "_comments_limpio.csv"
    if os.path.isfile(archivo_orig) and not os.path.isfile(archivo_limpio):
        print(str(dt.datetime.now()) + " - Limpiando " + archivo_orig)
        try:
            df_comments = pd.read_csv(archivo_orig, encoding="utf-8", sep="|", lineterminator="~")
            df_comments.sort_values("created_time", ascending=False, inplace=True)
            df_comments.reset_index(inplace=True)
            df_comments.drop("index", axis=1, inplace=True)
            df_comments["from_name"] = df_comments["from"].apply(get_name)
            df_comments["from_name_id"] = df_comments["from"].apply(get_name_id)
            df_res = df_comments[["created_time","from_name","from_name_id","id","message"]].copy(deep=True)
            df_res.dropna(subset=["message"], axis=0, inplace=True) # Borra los comentarios con texto vacío
            df_res.to_csv(archivo_limpio, index=False, encoding="utf-8", sep="|", line_terminator="~")
        except:
            print("Archivo vacío") # Si hay error en read_csv
```

Figura 24: Programa en python para limpiar los comentarios de todos los *posts*

Se ejecutó el programa limpiador de comentarios de la Figura 24 en las fechas 17/May/2017, 18/May/2017, 21/May/2017, 22/May/2017, 23/May/2017, 27/May/2017 y 28/May/2017 obteniendo los resultados mostrados en la Tabla 28.

Tabla 28: Comentarios válidos por operador

Operador	Comentarios extraídos	Comentarios con texto vacío (Limpieza)	Comentarios válidos	Número archivos CSV válidos	Tamaño total válidos
Movistar	450,785	3,911	446,681	5,781	68.28 MB
Claro	392,992	2,067	390,895	6,979	77.73 MB
Entel	239,345	7,387	231,958	974	38.31 MB
Bitel	335,006	4,562	330,444	868	54.15 MB
Virgin	12,187	467	11,720	159	1.82 MB
Total	**1,430,315**	**18,394**	**1,411,698**	**14,761**	**240.30 MB**

Respecto a los comentarios válidos se puede apreciar en la Figura 25 que Movistar y Claro tuvieron un incremento significativo de comentarios hasta el 2013, sin embargo, en el 2014 con el ingreso al mercado de Entel y Bitel redujeron la cantidad de comentarios casi a la mitad. En el 2015 crece significativamente la cantidad de comentarios de Bitel quien lidera y mantiene una cantidad similar en 2016. En el caso de Entel en 2015 llega al mismo nivel de comentarios que Claro y Movistar y en 2016 crece al doble con respecto al año anterior llegando al segundo lugar de comentarios y muy lejos le siguen Movistar, Claro y Virgin. Se evidencia que tanto Entel como Bitel logran en 2016 mayor atención de sus seguidores con los *posts* de promociones que publican en Facebook lo cual se refleja en que lideran la cantidad de comentarios en 2016.

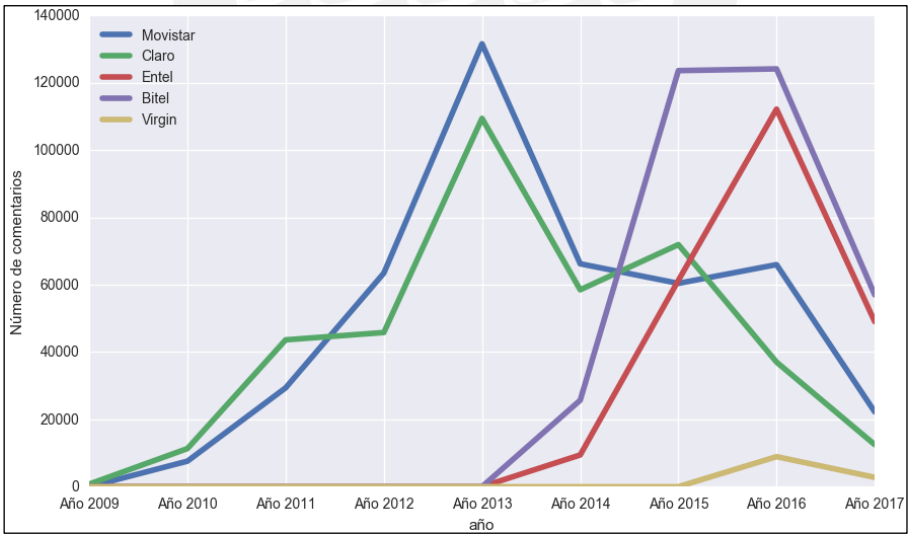


Figura 25: Comentarios válidos publicados en Facebook por operador por año

4.2.3. Corpus

En resumen, luego de la extracción y limpieza de *posts* y comentarios se obtuvo el corpus mostrado en la Tabla 29.

Tabla 29: Corpus

Operador	*Posts* válidos			Comentarios válidos		
	Cantidad	**Archivos CSV**	**Tamaño**	**Cantidad**	**Archivos CSV**	**Tamaño**
Movistar	6,004	1	4.79 MB	446,681	5,781	68.28 MB
Claro	7,625			390,895	6,979	77.73 MB
Entel	975			231,958	974	38.31 MB
Bitel	869			330,444	868	54.15 MB
Virgin	161			11,720	159	1.82 MB
Total	**15,634**	**1**	**4.79 MB**	**1,411,698**	**14,761**	**240.30 MB**

4.3. Discusión

Se recolectaron más de 15,000 *posts* y más de 1,400,000 comentarios que equivalen al total de *posts* y comentarios de los 5 operadores desde el primer *post* y comentario publicados en el año 2009 hasta el mes de Mayo del 2017. Tal como lo indica Gógova (2015), esta información legalmente obtenida sobre el entorno competitivo del pasado y presente permitirá determinar previsiones / tendencias / prospectivas del futuro y facilitar la toma de decisiones en beneficio de la empresa.

La recolección de datos mostrada en este capítulo 4 cubre completamente la segunda fase del proceso de inteligencia competitiva planteada en esta tesis, denominada "Recolección", que está acorde con lo planteado por Arroyo Varela (2005) y lo mencionado en las 6 investigaciones de la revisión sistemática realizada en esta tesis y mostradas en Tabla 18 de capítulo 3 (i02 (Abdellaoui y Nader 2015), i03 (Spruit y Cepoi 2015), i07 (Chouder y Chalal 2014), i08 (Olszak 2014), i15 (Chen et al. 2008), i18 (Li et al. 2007)) que consiste en recolectar datos de fuentes públicas y legales, limpiarla y organizarla.

Respecto a los métodos y procedimientos para la recolección de datos, se utilizó el API de la fuente de datos de Facebook, el API es completamente funcional y no tiene limitaciones de cantidad de registros extraídos en una cantidad de tiempo, la única limitación es que cada token de acceso dura aproximadamente 2 meses que fue tiempo suficiente porque la extracción fue realizada por partes en el transcurso de 2 semanas (Del 16/May/2017 al 28/May/2017). Según la revisión sistemática realizada en esta tesis, en 2 (i05 (Wongthongtham y Abu-Salih 2015), i09 (Dai et al. 2011)) de 9 investigaciones (Tabla 19 de capítulo 3) también usaron el API de la fuente de datos (Twitter) para la recolección de datos, en el resto de investigaciones usaron un rastreador de páginas web (Web Crawler) lo cual no aplica en este caso. Para la programación del extractor se utilizó Python con la librería requests para HTTP/HTTPS que permitió simplificar la programación con el API de la fuente de datos de Facebook. De igual manera se usó Python para realizar la limpieza de datos y para el cálculo de estadísticas tal como se muestra detalladamente en los resultados alcanzados.

Se eligió como fuente de datos a Facebook porque además de que los 5 operadores publican de forma diaria sus promociones en esa red social fue factible extraer los datos mediante el API de Facebook. Según la revisión sistemática realizada en esta tesis (Tabla 24 de capítulo 3), en el 32% de las 19 investigaciones revisadas se menciona a las redes sociales como fuente de datos para la inteligencia competitiva que junto con las noticias (42%) y los blogs (37%) son las 3 fuentes de datos más usadas.

5. ONTOLOGÍA DE TELECOMUNICACIONES

5.1. Introducción

Luego de realizar la recolección de datos en el capítulo 4 y obtener los *posts* limpios se necesita darles semántica o significado a los datos. En la Figura 26 se muestra en líneas entrecortadas las 4 fases del proceso de inteligencia competitiva que inicia con la fase de identificación de necesidades de información (1), luego por la fase de recolección (2) ya realizada, continúa con la fase de análisis (3) que contiene a la Ontología de Telecomunicaciones que está sombreada en gris y se detalla en este capítulo 5 y que genera los *post* clasificados que servirán para la fase de diseminación (4) que se describe en capítulo 7 para que los tomadores de decisiones del operador tengan información relevante y se genere la inteligencia competitiva.

Este capítulo 5 comprende la descripción y discusión de los resultados alcanzados para el objetivo específico 2 "**OE.2**: Clasificar semánticamente cada publicación (*post*) usando su texto completo e identificando palabras y expresiones en lenguaje natural para darles significado en el dominio de telecomunicaciones".

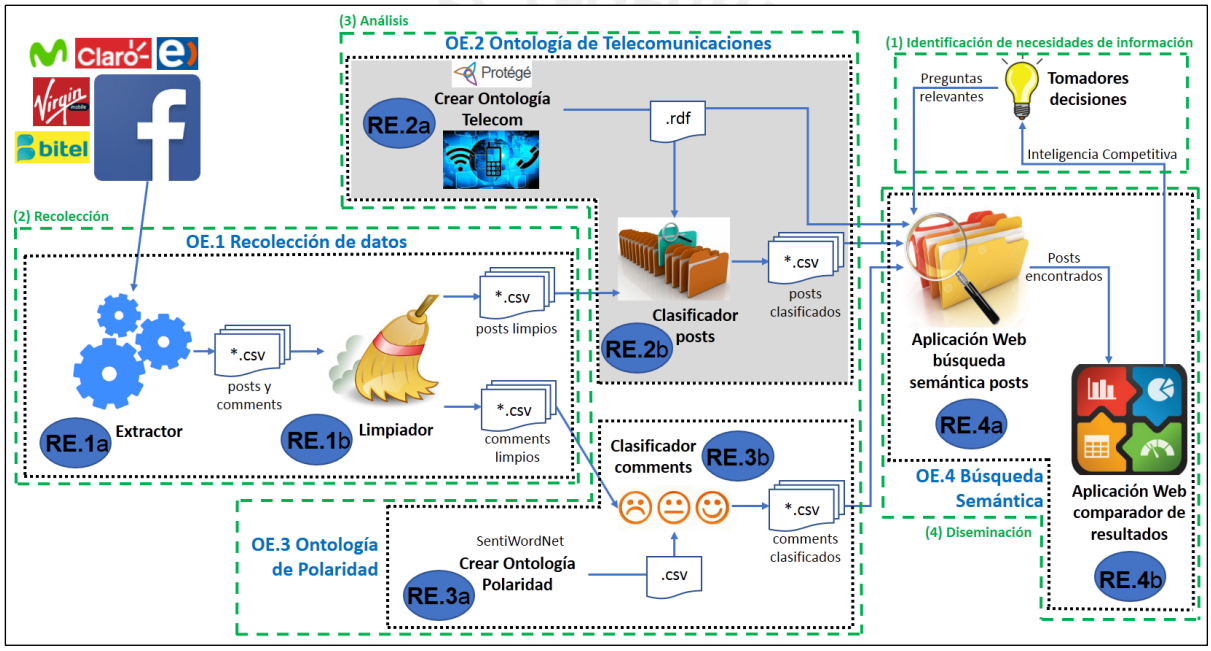


Figura 26: Ontología de Telecomunicaciones

5.2. Resultados alcanzados

5.2.1. Ontología de Telecomunicaciones

Respecto al resultado esperado "**RE.2a**: Ontología de dominio, en formato RDF, para representar el conocimiento en telecomunicaciones"; para el desarrollo de la ontología en telecomunicaciones, se empezó definiendo los siguientes conceptos relevantes desde el punto de vista del autor de esta tesis que trabaja más de 10 años en el dominio de las telecomunicaciones:

- **Servicios:** Que el cliente puede usar desde su celular.

- **Voz:** Permite al cliente llamar a celulares del mismo operador (*on-net*) o de otros operadores (*off-net*). También llamar a fijos nacionales y larga distancia internacional.
- **Mensaje de Texto:** Permite al cliente enviar mensajes de texto *on-net* y *off-net*.
- **Datos:** Permite al cliente usar internet y aplicaciones (*apps*) que usan datos como:
  - Redes Sociales: Facebook, Instagram, WhatsApp, Messenger.
  - Música: Spotify, Apple music.
  - Video: Youtube, Netflix.

- **Planes:** Ofrece los servicios al cliente bajo ciertas condiciones.
  - **Prepago:** Cliente compra recargas de dinero y luego compra bolsas de minutos para voz, cantidad de mensajes de texto o megabytes de datos para usar los servicios hasta una fecha de vencimiento.
  - **Postpago:** Cliente paga al operador una renta fija de dinero mensual para tener cantidades fijas de minutos de voz, mensajes de texto y megabytes de datos válidos por un mes hasta el final de su ciclo de facturación. Cuando las cantidades fijas se terminan existen dos modalidades de contratación:
    - **Control:** Se comporta como Prepago para usar los servicios. En su factura o recibo de servicios mensual sólo se cobra la renta fija.
    - **Libre:** Cliente puede continuar usando los servicios y el exceso es tarifado al cliente en su factura o recibo de servicios mensual junto a la renta fija.

Una vez definidos los principales conceptos relevantes, se tomó 110 *posts* limpios de los más recientes (0.7 % de un total de 15,634 *posts* limpios) y para cada uno se leyó el texto completo y se extrajo manualmente las palabras que identifican a cada concepto relevante previamente definido. En el proceso de extracción se identificaron conceptos nuevos y también palabras repetidas que habían sido previamente extraídas. En la Tabla 30 se muestra un ejemplo de extracción de palabras de los *posts* las cuales están subrayadas y sombreadas.

Tabla 30: Ejemplo de extracción de palabras que identifican conceptos

Operador	Texto completo del *Post*	Palabras extraídas	Conceptos identificados
Entel	¡Con Entel cierra tu verano y disfruta de Facebook para ver fotos, videos, chatear y postear gratis durante el mes de marzo! *De 11pm a 6am. Qué esperas para migrar :) #MigraAEntel Más info aquí: http://goo.gl/RmKqmo	facebook	Servicio:Datos:Redes Sociales
		foto, video, chatear, postear	Servicio:Datos
		gratis	Servicio:Tarifa:Cero **(Nuevo)**
		migra	Portabilidad **(Nuevo)**
Bitel	Este verano ☀ conéctate desde cualquier lugar con el Nuevo B-MIFI y navega con la Velocidad 4G LTE de Bitel. Además, podrás conectar hasta 10 dispositivos al mismo tiempo y vivir una experiencia increíble #Bitel #Bitel4G Promo B-MIFI con Internet Ilimitado 4G: https://goo.gl/HgGTGE Planes B-MIFI: https://goo.gl/TBgYcB	conect, navega, velocidad, 4g, internet	Servicio:Datos
		ilimitado	Servicio:Tarifa:Cero
Claro	Cambiándote gratis al recibo por e-mail Claro, estarás ayudando a los niños de ANIQUEM a rehabilitarse y hacer realidad sus sueños. Únete tú también a los #RecibosDeFelicidad aquí -> http://bit.ly/1TK1RCy	gratis **(Ya existe)**	Servicio:Tarifa:Cero
		recibo	Plan:Postpago:Control Plan:Postpago:Libre
		ayuda, niño, sueño, felicidad	Beneficio:Comunidad **(Nuevo)**
Bitel	¡ATENTOS! Porque el Tío #Bitel sorteará en cualquier momento entradas dobles para el concierto #DiaDeRockPeruano	sortea, entrada, concierto, rock	Beneficio:Cliente **(Nuevo)**

Para la extracción de palabras se trató en lo posible de tomar la palabra en singular, sin género y tomar la raíz de los verbos y convertirlas en minúsculas tal como se muestra en la Tabla 30. Se extrajeron 119 nuevas palabras para la ontología de telecomunicaciones de 61 *posts* limpios y para los 49 *posts* limpios restantes las palabras extraídas ya existían en la ontología, se encontró además nuevos conceptos relevantes que se agregaron a la ontología de telecomunicaciones. En la Figura 27 se muestra la ontología de telecomunicaciones que tiene 27 conceptos y 6 relaciones en 5 niveles jerárquicos.

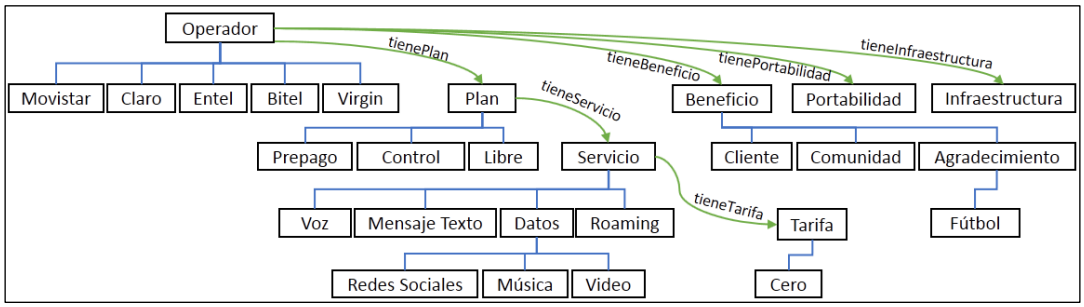


Figura 27: Ontología en Telecomunicaciones – Diagrama de Conceptos y Relaciones

Se utilizó Protégé para la elaboración de la Ontología de dominio en Telecomunicaciones generando un archivo en formato RDF. En la Figura 28 se puede apreciar un ejemplo de las palabras (a) del concepto Servicio:Datos (b) y las relaciones que tiene el concepto Servicio:Datos con los conceptos Tarifa y Plan (c).

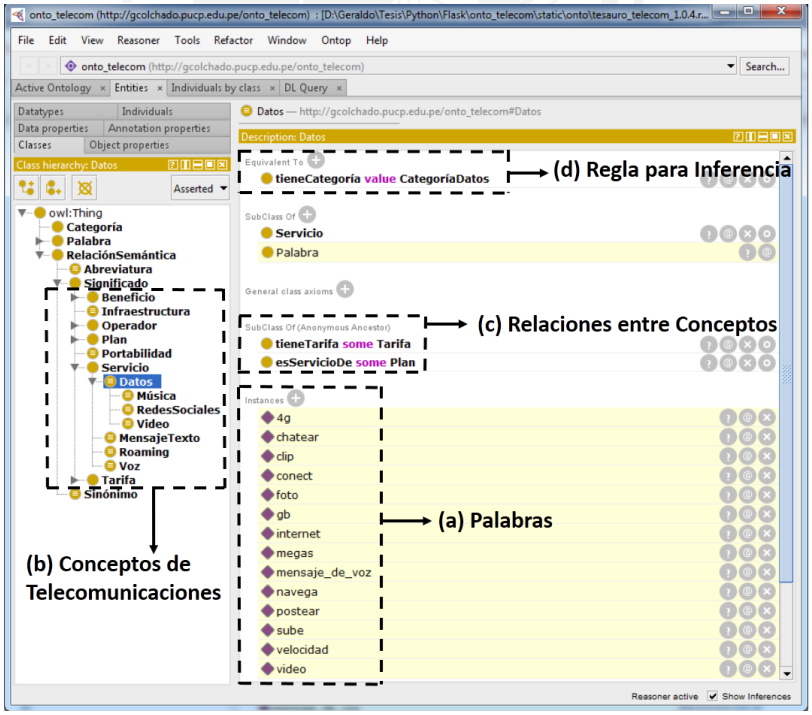


Figura 28: Ontología en Telecomunicaciones en Protégé

En la Figura 29 se puede apreciar un concepto o clase “Palabra” que contiene todas las palabras de la ontología de telecomunicaciones, cada palabra fue relacionada con una o varias instancias de la

clase “Categoría”, a su vez cada concepto o clase de telecomunicaciones fue relacionado con una instancia de la clase “Categoría” como la clase “Datos” que se relaciona con la instancia “CategoríaDatos” para realizar la inferencia (Figura 28 (d)) y obtener las palabras (Figura 28 (a)) mediante el reasoner HermiT.

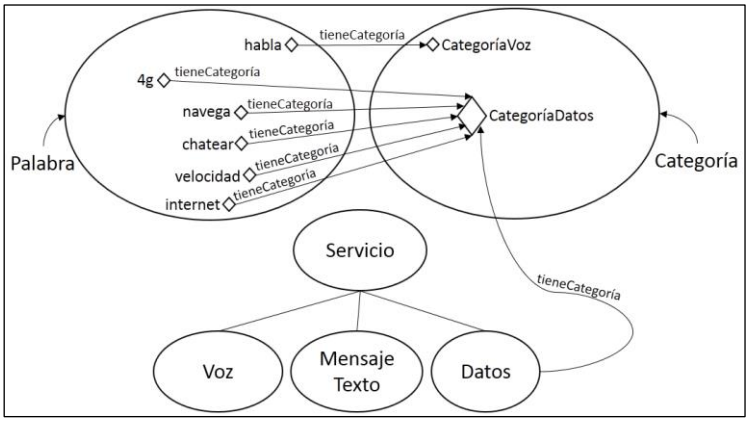


Figura 29: Ontología en Telecomunicaciones - Diagrama de relaciones entre instancias

En la Tabla 31 se muestran todas las palabras de la ontología de telecomunicaciones por cada concepto identificado, se puede apreciar que una palabra puede identificar más de un concepto, como la palabra “recibo” que identifica los conceptos “Operador:Plan:Control” y “Operador:Plan:Libre” y que además de palabras existen expresiones como “sin consumir tu saldo” o “habla con todos”.

Tabla 31: Palabras por concepto de Ontología en Telecomunicaciones

Concepto	**Palabras**
Operador:Plan:Servicio:Voz	habla, llama, minutos, telefonía
Operador:Plan:Servicio:MensajeTexto	sms
Operador:Plan:Servicio:Datos	4g, chatear, clip, conect, foto, gb, internet, megas, “mensaje de voz”, navega, postear, sube, velocidad, video
Operador:Plan:Servicio:Datos:Música	canción, música, orquesta, playlist, spotify
Operador:Plan:Servicio:Datos:RedesSociales	facebook, fb, Instagram, messenger, whatsapp
Operador:Plan:Servicio:Datos:Video	netflix, tv, youtube
Operador:Plan:Servicio:Roaming	frontera
Operador:Plan:Servicio:Tarifa:Cero	“costo alguno”, gratis, gratuit, ilimitad, “no tendrán costo”, “sin consumir tu saldo”, “sin costo”, “sin saldo mínimo”
Operador:Plan:Prepago	prepago, recarga
Operador:Plan:Control	contrato, deuda, postpago, recarga, recibo
Operador:Plan:Libre	contrato, deuda, postpago, recibo
Operador:Beneficio:Comunidad	acopio, afectad, alimento, ayuda, carretera, desastre, donaci, emergencia, felicidad, fenómeno, hermano, huaico, huayco, inundaci, lluvia, mano, niño, puente, sueño, unidos
Operador:Beneficio:Cliente	bono, carrera, cena, club, concierto, cuponera, entrada, ganador, ganaron, invita, pequeño, “por la compra de”, postula, promo, regalo, rock, sorpresa, sortea, teatro, viaj
Operador:Beneficio:Agradecimiento	agradecido, agua, conciencia, “feliz día”, gracias, mujer, sociedad
Operador:Beneficio:Agradecimiento:Fútbol	alenta, blanquirroja, clasificatoria, empate, jueg, partido, puntos
Operador:Infraestructura	cables, continuidad, fibra, normali, operatividad, red, restablec, técnico
Operador:Portabilidad	camb, cámbiate, ex, migra, portabilidad
Operador:Movistar	“cualquier operador”, “habla con todos”, “red privada”, rpm
Operador:Claro	“cualquier operador”, “habla con todos”, “red privada”, rpc
Operador:Entel	“cualquier operador”, “habla con todos”
Operador:Bitel	“cualquier operador”, “habla con todos”
Operador:Virgin	“cualquier operador”, “habla con todos”

5.2.2. Clasificador de *Posts*

Respecto al resultado esperado "**RE.2b:** Proceso automático para clasificar todas las publicaciones aplicando la ontología en telecomunicaciones"; una vez completada la ontología en telecomunicaciones en Protégé se generaron dos archivos en formato RDF, uno conteniendo las clases, relaciones e instancias y otro con el resultado de las inferencias conteniendo las palabras de cada concepto de telecomunicaciones. Se programó en Python un lector de archivos RDF utilizando la librería ontospy para obtener en archivo CSV las 119 palabras de la ontología con los conceptos que identifican, en la Figura 30 se muestra un segmento del código fuente que lee el archivo RDF original e inferido.

```python
import ontospy as os # Librería para trabajar con ontologías en formato RDF/XML
import pandas as pd # Librería para DataFrames

# Lee ontologías
onto_telecom          = os.Ontospy("static/onto/tesauro_telecom_1.0.4.rdf") # original
onto_telecom_inferido = os.Ontospy("static/onto/tesauro_telecom_1.0.4_inferido.rdf") # inferido

# Variables globales
URI_raiz = "http://gcolchado.pucp.edu.pe/onto_telecom#"
jerarquia_significado_raiz = "/RelaciónSemántica/Significado"

# Función que devuelve la jerarquía de una clase
def jerarquia_clase(nombre_clase):
    clase = onto_telecom.getClass(nombre_clase)[0]
    df_jerarquia = pd.DataFrame(clase.ancestors())
    jerarquia = ""
    for i in reversed(df_jerarquia.index):
        jerarquia = jerarquia + "/" + df_jerarquia[0][i].bestLabel()
    return jerarquia + "/"
```

Figura 30: Segmento de código fuente en Python de lector de archivos RDF

En Figura 31 se muestra el diagrama de flujo del proceso de clasificación de *posts* donde se puede apreciar que para cada texto de *post* limpio se buscan las palabras de la ontología para determinar los conceptos a clasificar, los resultados de los *posts* clasificados se almacenan en archivo CSV. En la Figura 32 se muestra un segmento de código fuente del clasificador de *posts*.

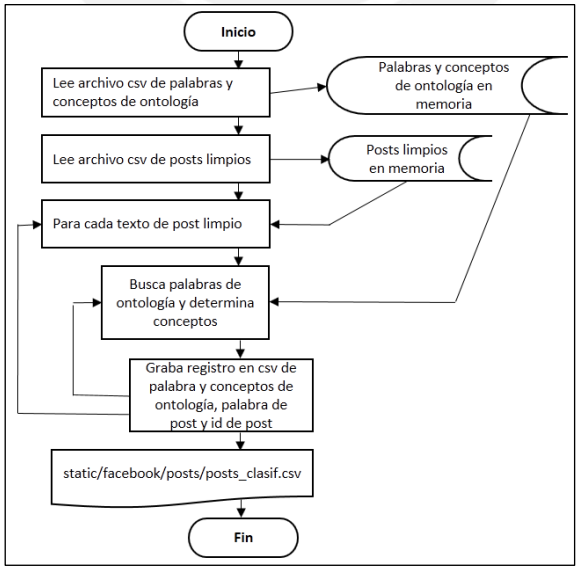


Figura 31: Diagrama de flujo del proceso de clasificación de *posts*

```
def clasifica_todos_posts():
    global df_posts_clasif # Declara como global para cambiar valor dentro del procedimiento
    df_posts_clasif = pd.DataFrame() # Inicializa
    total_post = df_posts["id"].count()
    for i in df_posts.index:
        print("Clasificando post {0:,}".format(i), " de ", "{0:,}".format(total_post))
        id_post = df_posts["id"][i]
        post_fb = df_posts["message"][i]
        clasificador(post_fb) # Devuelve resultados en DataFrames globales df_clasifica_filtro y df_clasifica_comp
        df_resultado = pd.concat([df_clasifica_filtro, df_clasifica_comp])
        df_resultado["id"] = id_post
        df_posts_clasif = pd.concat([df_posts_clasif, df_resultado])
    df_posts_clasif.to_csv("static/facebook/posts/posts_clasif.csv", index=False, encoding="utf-8", sep="|")
```

Figura 32: Segmento de código fuente en Python de clasificador de *posts*

Se ejecutó el clasificador de *posts* el 17/May/2017 para el total de 15,634 *posts* limpios, el proceso demoró aproximadamente 12 minutos y clasificó 10,489 *post* limpios con al menos un concepto por *post* que equivale al 67% del total de *posts* limpios.

5.3. Discusión

Una ontología bien definida se convierte en la columna vertebral del sistema (Li et al. 2007). En esta tesis, el principal componente del proceso de inteligencia competitiva es la ontología de telecomunicaciones, a pesar de que fue creada manualmente, sólo requirió menos del 1% de *posts* (0.7% ó 110 *posts* de un total 15,634 *posts* limpios) que fueron revisados y se extrajo manualmente las palabras que identificaban los conceptos para crear la ontología de telecomunicaciones que permitió clasificar al 67% del total de *posts* limpios. Debido a su importancia fue necesario el apoyo de un experto en telecomunicaciones para su creación manual y asegurar el éxito. Según la revisión sistemática realizada en esta tesis (Tabla 22 de capítulo 3), en el 61% de las 18 investigaciones revisadas también se planteó crear manualmente la ontología con apoyo de expertos ya que es la forma más usada en comparación con la creación automática y semiautomática de ontología que no requieren expertos o sólo parcialmente para validar los resultados.

Las palabras en el idioma español tienen muchas variantes de forma, para minimizar este impacto se extrajo del texto de los *posts*, donde fue posible, las palabras en singular, sin género (Ejemplo: gratuit) y la raíz de los verbos (Ejemplos: conect, viaj) y se almacenaron de esa forma en la ontología para que cuando se busquen esas palabras en los textos para la clasificación cubran la mayor cantidad de variantes de forma de las palabras y por consiguiente se clasifiquen más *posts*. Lo anterior permite cubrir algunas variantes de forma de las palabras, pero no cubre todas las variantes por lo que se propone como trabajo futuro que la ontología almacene las formas base de las palabras (*lemmas*) y que cada palabra del texto de los posts sea convertida a su forma base (*lemma*) mediante un *lemmatizer* y pueda ser comparada con las palabras de la ontología, de esta manera se cubrirían todas las variantes de forma de las palabras de la ontología para la clasificación.

Según la Real Academia Española [48], desambiguar es efectuar las operaciones necesarias para que una palabra, frase o texto pierdan su ambigüedad; en la ontología de telecomunicaciones se tomaron algunas frases o expresiones completas que son un grupo de palabras que al estar juntas tienen un significado que es diferente al significado de cada palabra individualmente. Por ejemplo, en la expresión “sin consumir tu saldo” que es opuesta en significado a “consumir tu saldo”, la palabra “sin” le cambia el significado, sin embargo, el experto tuvo que desambiguar la expresión y

[48] http://dle.rae.es/srv/search?m=30&w=desambiguar

darle un significado en el contexto de las telecomunicaciones que este caso incluye la palabra “sin” y significa que el servicio se ofrece a tarifa cero o gratis. De igual manera el experto desambiguó otras expresiones como “no tendrán costo” y “sin saldo mínimo” que en el contexto de telecomunicaciones también significan que el servicio se ofrece a tarifa cero o gratis.

El principal uso de la ontología de telecomunicaciones creada es realizar la identificación de temas o conceptos en el texto de los *posts* lo cual se denomina en inglés como *Topic detection* o *Topic tagging* y que según la revisión sistemática realizada en esta tesis (Tabla 23 de capítulo 3) es el mayor uso que se le da a las ontologías de dominio con un 56% de 18 investigaciones revisadas seguido por el *Named Entity Recognition* (33%) que en este caso no fue necesario utilizarlo.

Para la creación de la ontología de dominio en telecomunicaciones se utilizó Protégé que fue la herramienta más utilizada para crear ontologías mencionada explícitamente en 3 de 19 investigaciones de la revisión sistemática realizada en esta tesis (Tabla 20 de capítulo 3).

6. ONTOLOGÍA DE POLARIDAD

6.1. Introducción

Luego de realizar la recolección de datos en el capítulo 4 y obtener los comentarios limpios se necesita darles semántica o significado a los datos. En la Figura 33 se muestra en líneas entrecortadas las 4 fases del proceso de inteligencia competitiva que inicia con la fase de identificación de necesidades de información (1), luego por la fase de recolección (2) ya realizada, continúa con la fase de análisis (3) que contiene a la Ontología de Polaridad que está sombreada en gris y se detalla en este capítulo 6 y que genera los comentarios clasificados que servirán para la fase de diseminación (4) que se describe en capítulo 7 para que los tomadores de decisiones del operador tengan información relevante y se genere la inteligencia competitiva.

Este capítulo 6 comprende la descripción y discusión de los resultados alcanzados para el objetivo específico 3 "**OE.3**: Clasificar semánticamente la polaridad (positiva, negativa, neutra) de cada comentario usando su texto completo e identificando palabras en lenguaje natural que reflejen la polaridad".

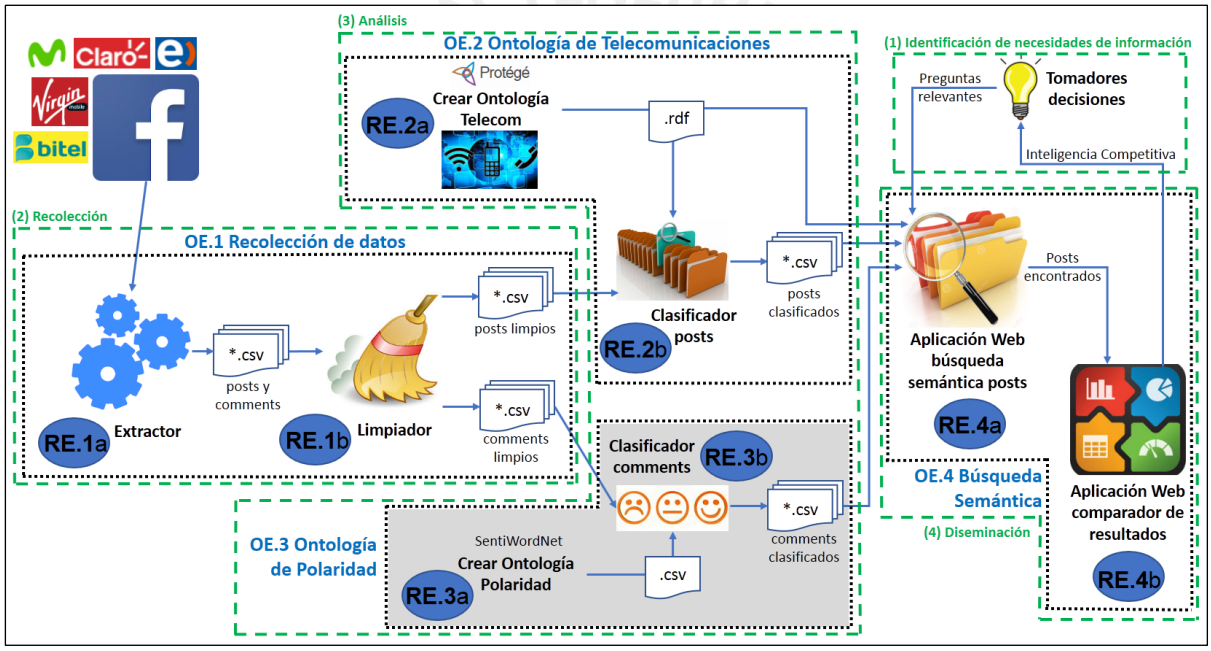


Figura 33: Ontología de Polaridad

6.2. Resultados alcanzados

6.2.1. Ontología de Polaridad

Respecto al resultado esperado "**RE.3a:** Ontología de dominio, en formato CSV, para representar la polaridad", en la Figura 34 se puede apreciar el proceso de creación de la ontología en polaridad, implementado en Python con la librería NLTK, que inició con el Corpus (A) que contiene 1,411,698 textos de comentarios limpios el cual fue convertido a minúsculas (1a), se separó el texto en palabras (1b), se eliminaron las palabras vacías (1c), se contó la frecuencia de palabras obteniendo

891,973 palabras diferentes, luego se obtuvieron los *synsets* [49] de cada palabra (2a) utilizando *WordNet* en español obteniendo *synsets* sólo para 8,916 palabras que equivalen al 1% del total de palabras diferentes, luego para cada *synset* se buscó su polaridad positiva y negativa (2b) en *SentiWordNet* obteniendo el puntaje positivo y el puntaje negativo, luego por cada palabra se revisaron todos sus *synsets* y se tomó el puntaje más positivo y el más negativo y para calcular la polaridad de la palabra se restaron ambos puntajes (2c) obteniendo el archivo de la Ontología (B) en formato CSV con las columnas palabra, frecuencia, *synsets* y polaridad con 8,916 palabras.

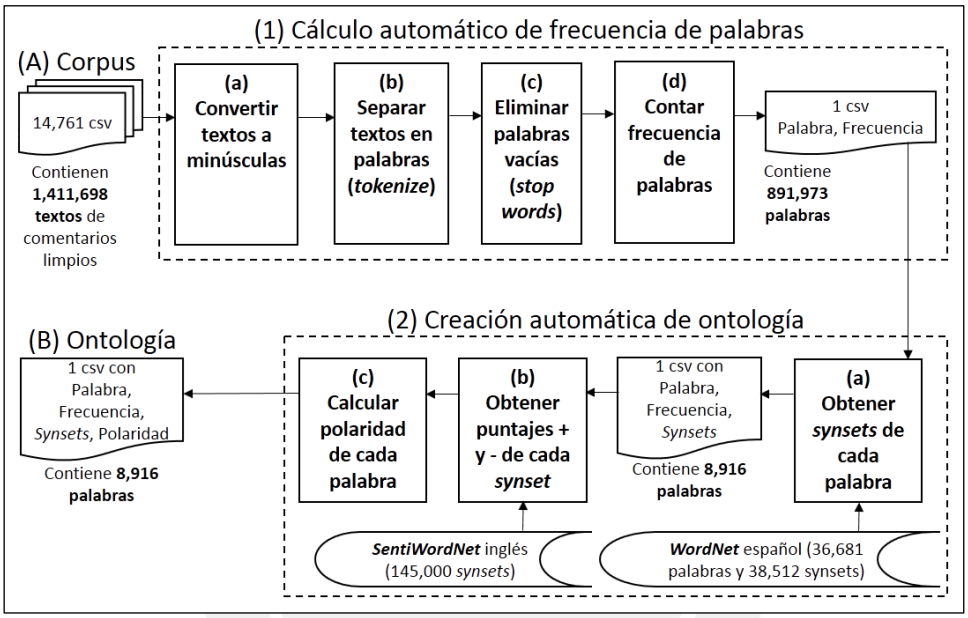


Figura 34: Diagrama del proceso de creación de la ontología en polaridad

En la Tabla 32 se muestran ejemplos en detalle del algoritmo del cálculo de polaridad para la creación automática de la ontología de polaridad (Figura 34 (2)), donde se busca el puntaje más positivo y se resta con el más negativo de los *synsets* de cada palabra.

Tabla 32: Ejemplos del algoritmo de cálculo de polaridad usando *WordNet* y *SentiWordNet*

Palabra	Frecuencia	*WordNet* español	*SentiWordNet* inglés		Polaridad = Mayor_Positivo - Mayor_Negativo
		Synsets	Puntaje Positivo	Puntaje Negativo	
excelente	3262	*Synset('divine.s.06')*	0.875	0.0	**1.0** = 1.0 - 0.0
		Synset('ace.s.01')	0.625	0.0	
		Synset('choice.s.01')	0.625	0.0	
		Synset('excellent.s.01')	1.0	0.0	
		Synset('top-flight.s.01')	1.0	0.0	
bonito	1483	*Synset('beautiful.a.01')*	0.75	0.0	**0.75** = 0.875 - 0.125
		Synset('pretty.s.01')	0.875	0.125	
		Synset('brave.s.03')	0.125	0.125	
		Synset('nice.a.01')	0.875	0.0	

[49] *Synset* o *Synonym Set* es un conjunto de palabras sinónimas que tienen el mismo significado.

		Synset('skipjack.n.02')	0.0	0.0	
		Synset('bonito.n.01')	0.0	0.0	
horrible	701	*Synset('atrocious.s.03')*	0.0	0.625	- **0.625** = 0.25 - 0.875
		Synset('awful.s.02')	0.0	0.625	
		Synset('ghastly.s.01')	0.25	0.625	
		Synset('atrocious.s.02')	0.0	0.875	
		Synset('horrid.s.01')	0.0	0.875	
		Synset('nasty.a.01')	0.0	0.875	
		Synset('nauseating.s.01')	0.0	0.5	

En la Figura 35 se muestra el segmento de código fuente en Python que calcula la polaridad de una palabra a partir de los puntajes de sus *synsets*.

```
from nltk.corpus import sentiwordnet as swn
import ast # Para convertir cadena a diccionario

# Calcula la polaridad
def get_polaridad(synsets):
    mayor_pos = 0
    mayor_neg = 0
    for synset in ast.literal_eval(synsets):
        pos_score = swn.senti_synset(synset).pos_score()
        neg_score = swn.senti_synset(synset).neg_score()
        if pos_score > mayor_pos:
            mayor_pos = pos_score
        if neg_score > mayor_neg:
            mayor_neg = neg_score
    mayores = mayor_pos - mayor_neg
    return mayores
```

Figura 35: Segmento de código fuente en Python para calcular polaridad

En la Tabla 33 se muestra un sub conjunto de palabras de la ontología de polaridad de 8,916 palabras tal cual se almacena en el archivo CSV.

Tabla 33: Subconjunto de palabras de la Ontología de polaridad

Palabra	**Frecuencia**	***Synsets***	**Polaridad**
maravilla	485	*[Synset('admirability.n.01'), Synset('miracle.n.01'), Synset('wonder.n.02'), Synset('wonder.n.01'), Synset('humdinger.n.01'), Synset('prodigy.n.01'), Synset('european_beggar-ticks.n.01'), Synset('marigold.n.01')]*	1.0
genial	3051	*[Synset('divine.s.06'), Synset('bang-up.s.01'), Synset('fantastic.s.02'), Synset('consummate.s.01'), Synset('mean.s.04')]*	1.0
atractivo	73	*[Synset('attractive.a.01'), Synset('engaging.s.01'), Synset('fetching.s.01'), Synset('personable.s.01'), Synset('prepossessing.s.01'), Synset('inviting.a.01'), Synset('charming.s.01'), Synset('good.s.15'), Synset('pleasingness.n.02'), Synset('glamor.n.01'), Synset('attractiveness.n.02'), Synset('appeal.n.02'), Synset('lure.n.01'), Synset('attraction.n.04')]*	0.875
odio	292	*[Synset('abhorrence.n.01'), Synset('hate.n.01')]*	-0.25
peor	7258	*[Synset('worst.r.01'), Synset('even.r.03'), Synset('worst.a.01'), Synset('worse.a.01'), Synset('worse.a.02'), Synset('worse.n.01')]*	-0.3189
pena	3124	*[Synset('grief.n.02'), Synset('pity.n.02'), Synset('woe.n.02'), Synset('misery.n.02'), Synset('sorrow.n.01'), Synset('forfeit.n.02'), Synset('sadness.n.02'), Synset('distress.n.02')]*	-0.375

6.2.2. Clasificador de Comentarios

Respecto al resultado esperado "**RE.3b:** Proceso automático para clasificar todos los comentarios aplicando la ontología en polaridad", para clasificar los comentarios se implementó en Python el proceso mostrado en la Figura 36 donde se aprecia que para cada archivo CSV del Corpus (A), a cada texto de comentario se le convirtió a minúsculas (a), se separó el texto en palabras (b), se eliminó las palabras vacías (c), se buscó cada palabra del texto en la ontología (B) para obtener su polaridad (d) y se sumó la polaridad de las palabras encontradas para determinar la polaridad total del comentario (e); la duración total del proceso de clasificación demoró aprox. 1.9 horas obteniendo 1,109,321 comentarios clasificados con polaridad que equivalen al 78.58% del total de comentarios (C).

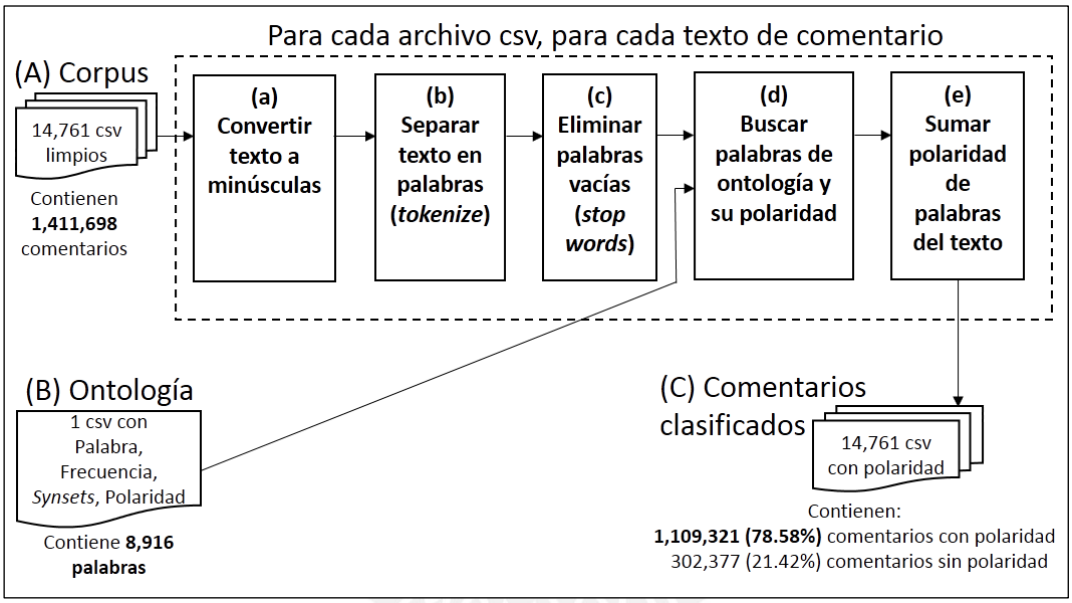


Figura 36: Diagrama de flujo del proceso de clasificación de comentarios usando ontología de polaridad

Para poder aumentar el porcentaje de comentarios clasificados de 78.58% a más, en la Figura 37 se muestra el proceso realizado de adición de palabras a la ontología que a partir de la frecuencia de palabras diferentes de comentarios (A) y las palabras de la ontología (B) obtuvo las palabras no encontradas en la ontología (1) y las ordenó por frecuencia descendente (C), luego se revisó manualmente las primeras 1,000 palabras (2) y se encontró 453 palabras que algunas estaban en plural, a otras les faltaba la tilde, otras estaban en diferentes tiempos verbales y otras estaban en género femenino, se realizó un proceso manual de *lemmatization* y se les transformó a singular, se agregó la tilde, se convirtió a verbo en infinitivo y se convirtió a género masculino (D), luego se buscó las palabras transformadas en la ontología (3) y se encontró 428 palabras (E) que fueron agregadas (4) a la nueva ontología de polaridad (F) que finalmente aumentó de 8,916 palabras a 9,344 palabras.

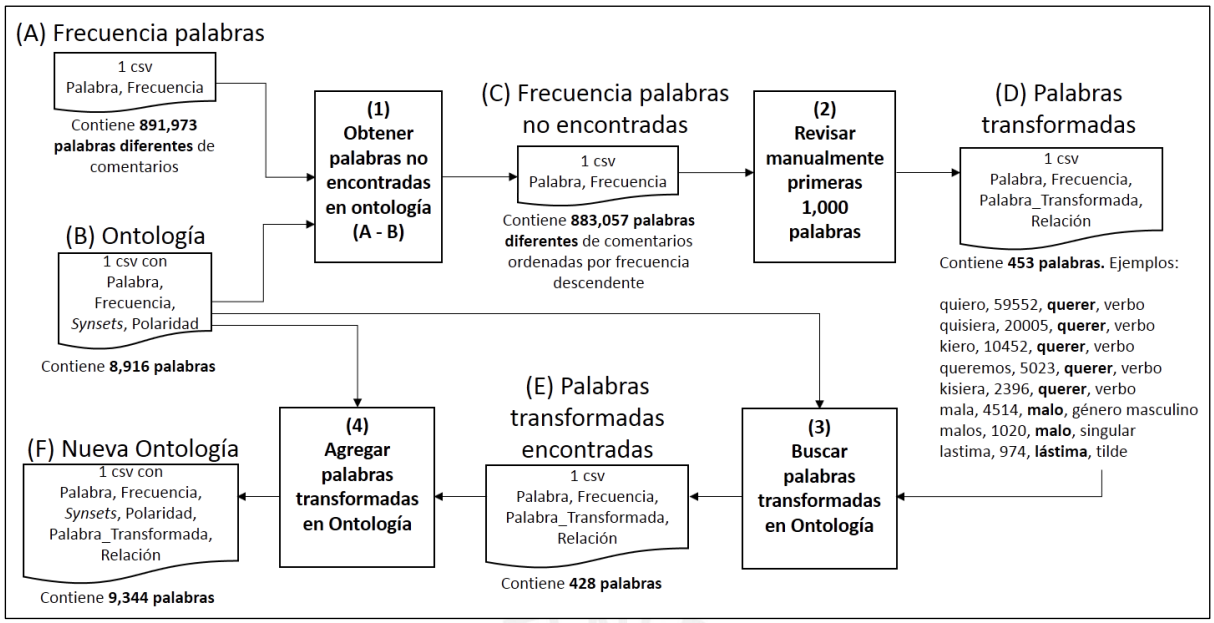


Figura 37: Proceso de adición de palabras a la ontología de polaridad

Se volvió a ejecutar el proceso de clasificación de comentarios utilizando la nueva ontología (Figura 37 (F)) con 9,344 palabras, con una duración total de aproximadamente 2.1 horas obteniendo 1,132,853 comentarios clasificados con polaridad que equivalen al 80.25% del total de comentarios. En la Tabla 34 se muestra un comparativo de la clasificación de comentarios con las dos ontologías.

Tabla 34: Comparativo de clasificación de dos ontologías de polaridad

	Palabras	**Tiempo Proceso**	**Comentarios Clasificados**	**Porcentaje del total**
Ontología	8,916	1.9 horas	1,109,321	78.58%
Nueva Ontología	9,344	2.1 horas	1,132,853	80.25%

En la Figura 38 se muestra un segmento de código fuente en Python que realiza la clasificación de todos los archivos de comentarios CSV.

```
# Lee los posts y clasifica los comentarios
df_posts = pd.read_csv(ruta_posts + 'posts.csv', encoding="utf-8", sep="|", lineterminator="~")
j = 0
for i in df_posts.index:
    if j < num_archivos:
        empresa = df_posts["empresa"][i]
        id_post = df_posts["id"][i]
        archivo_limpio     = ruta_comments + empresa + "/" + empresa + "_" + id_post + "_comments_limpio.csv"
        archivo_polaridad = ruta_comments + empresa + "/polaridad/" + empresa + "_" + id_post + "_comments_polaridad.csv"
        if os.path.isfile(archivo_limpio) and not os.path.isfile(archivo_polaridad):
            print(i)
            print("Inicio: " + str(j+1) + " de " + str(num_archivos) + " " + str(dt.datetime.now()) + " " + archivo_limpio)
            try:
                df_comments = pd.read_csv(archivo_limpio, encoding="utf-8", sep="|", lineterminator="~")
                df_comments["polaridad_palabras"] = df_comments["message"].apply(clasifica_comentario)
                df_comments["polaridad_puntaje"] = df_comments["polaridad_palabras"].apply(calcula_polaridad)
                df_comments.to_csv(archivo_polaridad, index=False, encoding="utf-8", sep="|", line_terminator="~")
                print("Fin: " + str(dt.datetime.now()) + " " + archivo_polaridad)
                j = j + 1
            except:
                print("Fin: " + str(dt.datetime.now()) + " {}".format(sys.exc_info()[0])) # Ejemplo: Si archivo_limpio está vacío
    else:
        break # Ya no continúa clasificando y sale del for
```

Figura 38: Segmento de código fuente en Python para clasificación de comentarios

En la Tabla 35 se puede apreciar algunos comentarios clasificados tal cual se almacenan en el archivo CSV de comentarios con polaridad, los que tienen puntaje de polaridad mayor que cero son comentarios positivos y los menores que cero son negativos.

Tabla 35: Ejemplo de comentarios clasificados con ontología de polaridad

Fecha	Comentario	Polaridad Puntaje	Polaridad Palabras
2017-05-17	esa promoción de susy diaz esta recontra horrible	-0.625	**-0.625** = promoción(swn 0.0) horrible(swn -0.625)
2017-05-14	El servicio de interned pesimo x ica .lento y ni te comunican si tienen falla.q mal	-0.5	**-0.5** = si(swn 0.0) servicio(swn 0.0) mal(swn -0.5)
2017-05-14	Bitel lo máximo en cobertura.pero dime hay telefonia fija para chimbote	0.375	**0.375** = máximo(swn 0.25) dime(swn 0.125)
2017-05-12	Tío bitel tengo mi computadora pero no tengo internet por que dicen que los cables están saturados en mi barrio y telefónica es una wevada como hago para poner internet bitel	0.625	**0.625** = barrio(swn 0.0) computadora(swn 0.0) dicen(swn 0.125) hago(swn 0.125) internet(swn 0.0) internet(swn 0.0) poner(swn 0.375) tío(swn 0.0)
2017-05-04	Buenos días soy de otro operador quisiera cambiarme a entel como lo hago??	0.75	**0.75** = operador(swn 0.0) buenos(swn 0.375) cambiarme(swn -0.25) días(swn 0.125) hago(swn 0.125) quisiera(swn 0.375)

6.3. Discusión

A diferencia del proceso de creación de la ontología de telecomunicaciones explicado en capítulo 5 y que fue manual con el apoyo de un experto, para la ontología de polaridad se creó un proceso automático para poblar en la ontología a 8,916 palabras con su polaridad a partir de *WordNet* y *SentiWordNet*, luego se realizó un proceso manual de *lemmatization* para agregar 428 palabras que son variantes de forma de palabras ya existentes en la ontología resultando un proceso semiautomático que permitió aumentar de 78.58% a 80.25% el porcentaje total de comentarios clasificados. Según la revisión sistemática realizada en esta tesis (Tabla 22 de capítulo 3), el 22% de las 18 investigaciones revisadas también realizaron la creación automática o semiautomática de

ontología donde no requieren expertos o sólo parcialmente para validar los resultados como en este caso.

En el proceso de creación de la ontología de telecomunicaciones explicado en capítulo 5 se revisó manualmente sólo el 0.7% ó 110 *posts* de un total 15,634 *posts* limpios que permitió clasificar al 67% del total de *posts*; debido a la enorme cantidad de comentarios limpios (1,411,698) el 0.7% representaba a 9,882 comentarios lo cual hubiera tomado mucho tiempo revisar manualmente, por eso fue implementando un proceso automático para poblar palabras en la ontología utilizando una base de datos ya existente de lexicones con polaridad como *SentiWordNet* que es accesible desde Python.

Cada palabra en *WordNet* usualmente forma parte de varios *Synsets*, cada *Synonym Set* tiene un significado y un valor de polaridad que puede ser diferente para la misma palabra, es decir una palabra tiene diferentes significados dependiendo del contexto en que se encuentre; debido a la complejidad para desambiguar a una palabra que implica analizar las otras palabras cercanas en el texto o incluso todo el texto del comentario para determinar a qué *Synset* pertenece se decidió asumir el peor escenario, es decir buscar el valor de polaridad de cada uno de los *Synsets* de la palabra y tomar el valor más positivo y restarle el valor más negativo asignando el valor resultante como polaridad de la palabra en la ontología de polaridad, se pensó también tomar el promedio pero en palabras con muchos *Synsets* el valor promedio se reducía mucho.

A cada palabra de la ontología se le asignó un valor de polaridad, este valor de polaridad fue utilizado para ser sumado junto con los valores de polaridad de las otras palabras del comentario y obtener la polaridad total del comentario. No se realizó ningún análisis de casos donde se invierte la polaridad en una expresión como cuando se antepone la palabra “no” o “sin” lo cual se propone como trabajo futuro.

En esta tesis se planteó el uso de una ontología de dominio en polaridad para poder determinar la positividad, negatividad y neutralidad de los comentarios de Facebook como también fue planteado en el 11% de 18 investigaciones de la revisión sistemática (Tabla 23 de capítulo 3).

Con la ontología de telecomunicaciones, la ontología de polaridad y la clasificación de *posts* y comentarios implementados y mostrados en el capítulo 5 y 6 se cubre completamente la tercera fase del proceso de inteligencia competitiva planteada en esta tesis, denominada “Análisis”, que está acorde con lo planteado por Arroyo Varela (2005) y lo mencionado en las 6 investigaciones de la revisión sistemática realizada en esta tesis y mostradas en Tabla 18 de capítulo 3 (i02 (Abdellaoui y Nader 2015), i03 (Spruit y Cepoi 2015), i07 (Chouder y Chalal 2014), i08 (Olszak 2014), i15 (Chen et al. 2008), i18 (Li et al. 2007)) que consiste en clasificar y dar significado a los datos para que puedan utilizarse en la cuarta fase de “Diseminación”, explicada en el capítulo 7, donde el tomador de decisiones podrá obtener respuestas a sus preguntas relevantes y tener inteligencia competitiva.

7. BÚSQUEDA SEMÁNTICA

7.1. Introducción

Luego de dar semántica a los datos en el capítulo 5 y 6 y obtener los *posts* y comentarios clasificados, estos quedan listos para ser buscados. En la Figura 39 se muestra en líneas entrecortadas las 4 fases del proceso de inteligencia competitiva que inicia con la fase de identificación de necesidades de información (1), luego la fase de recolección (2) ya realizada, seguido por la fase de análisis (3) ya realizada y finalmente la fase de diseminación (4) sombreada en gris que se describe en este capítulo 7 para que los tomadores de decisiones del operador tengan respuesta a sus preguntas relevantes y se genere la inteligencia competitiva.

Este capítulo 7 comprende la descripción y discusión de los resultados alcanzados para el objetivo específico 4 "**OE.4**: Implementar un motor de búsqueda semántica de publicaciones (*posts*) en el dominio de telecomunicaciones y un comparador de resultados por operador y mostrar información relevante para el diseño de promociones más competitivas".

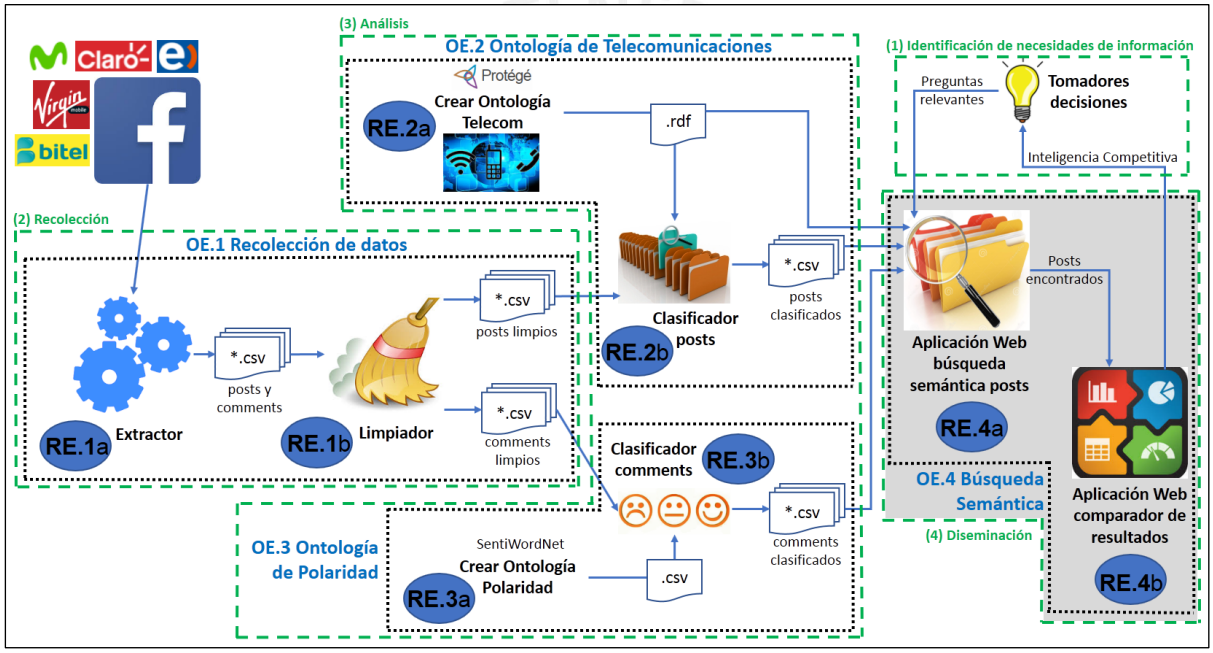


Figura 39: Búsqueda semántica

7.2. Resultados alcanzados

Respecto a los resultados esperados "**RE.4a:** Aplicación Web que permita al usuario realizar búsquedas semánticas de publicaciones en el dominio de telecomunicaciones" y "**RE.4b:** Aplicación Web que compare los resultados de la búsqueda semántica por operador"; para que los tomadores de decisiones puedan tener respuesta a sus preguntas relevantes respecto a la competencia se implementó una aplicación web en Python que les permite realizar búsquedas por conceptos de telecomunicaciones y encontrar todos los *posts* relacionados y poder compararlos. En la Figura 40 se puede apreciar un diagrama de componentes de la aplicación web donde se cuenta con una página inicial (a) donde se puede seleccionar tres opciones en el menú:

1. **Ontología Telecom (b):** Se solicita cuantas palabras se desea buscar y se muestra un árbol de palabras o WordTree de la ontología de dominio en telecomunicaciones donde se puede seleccionar las palabras a incluir en los criterios de búsqueda (b) que son enviados a una página que realiza la búsqueda de *posts* y muestra los resultados (c) y desde donde se puede ejecutar en otra página la comparación de los resultados (d).
2. **Todos los *Posts* (c):** Se aplica el criterio de búsqueda "TODOS" que es enviado a una página que realiza la búsqueda de *posts* y muestra los resultados (c) y desde donde se puede ejecutar en otra página la comparación de los resultados (d).
3. **Clasificador *Post* (e):** Se muestra una página donde se puede escribir el texto de un *post* para probar el clasificador de *posts* (e) usando la ontología de dominio en telecomunicaciones.

Figura 40: Diagrama de componentes de aplicación web

En la Figura 41 se muestra un segmento de código fuente en Python de la aplicación web implementada usando el *microframework flask* que está relacionada con los componentes de la Figura 40.

```python
# Microframework en Python para Desarrollo Web
from flask import Flask, render_template, request
# Librería para DataFrames
import pandas as pd

# Aplicación Web con Microframework Flask
app = Flask(__name__)

@app.route('/')
def onto_telecom():
    return render_template('onto_telecom.html')

@app.route('/onto/')
def onto():
    titulo_html = "Ontología Telecom"
    return html_onto(titulo_html)

@app.route('/buscador/')
def buscador():
    titulo_html = "Buscador"
    return html_busqueda(titulo_html, 0, "TODOS")

@app.route('/buscador_pag/<pag_pal>')
def buscador_pag(pag_pal):
    df_param = pd.DataFrame(pag_pal.split("-"))
    pag = int(df_param[0][0])
    pal = df_param[0][1]
    titulo_html = "Buscador Paginado"
    return html_busqueda(titulo_html, pag, pal)

@app.route('/comparador/<pal>')
def comparador(pal):
    titulo_html = "Comparador"
    return html_comparador(titulo_html, pal)

if __name__ == '__main__':
    app.run("0.0.0.0", 5000)
```

Figura 41: Segmento de código fuente en Python de aplicación web

7.2.1. Criterios de búsqueda de *posts*

Se implementó una interfaz de búsqueda amigable para el usuario, en formato de árbol de palabras o *WordTree*, donde se muestran todos los conceptos, relaciones y palabras de la ontología de dominio en telecomunicaciones y el tomador de decisiones o usuario tenga conocimiento del total de conceptos y palabras con los que puede realizar búsquedas. En la Figura 42 se muestra una parte del árbol de palabras donde el usuario al ingresar a la página se le solicita la cantidad de palabras a buscar, luego para realizar la búsqueda debe seleccionar mediante un *click* un concepto o una palabra que desea agregar al criterio de búsqueda y así sucesivamente hasta completar la cantidad de palabras elegida al inicio.

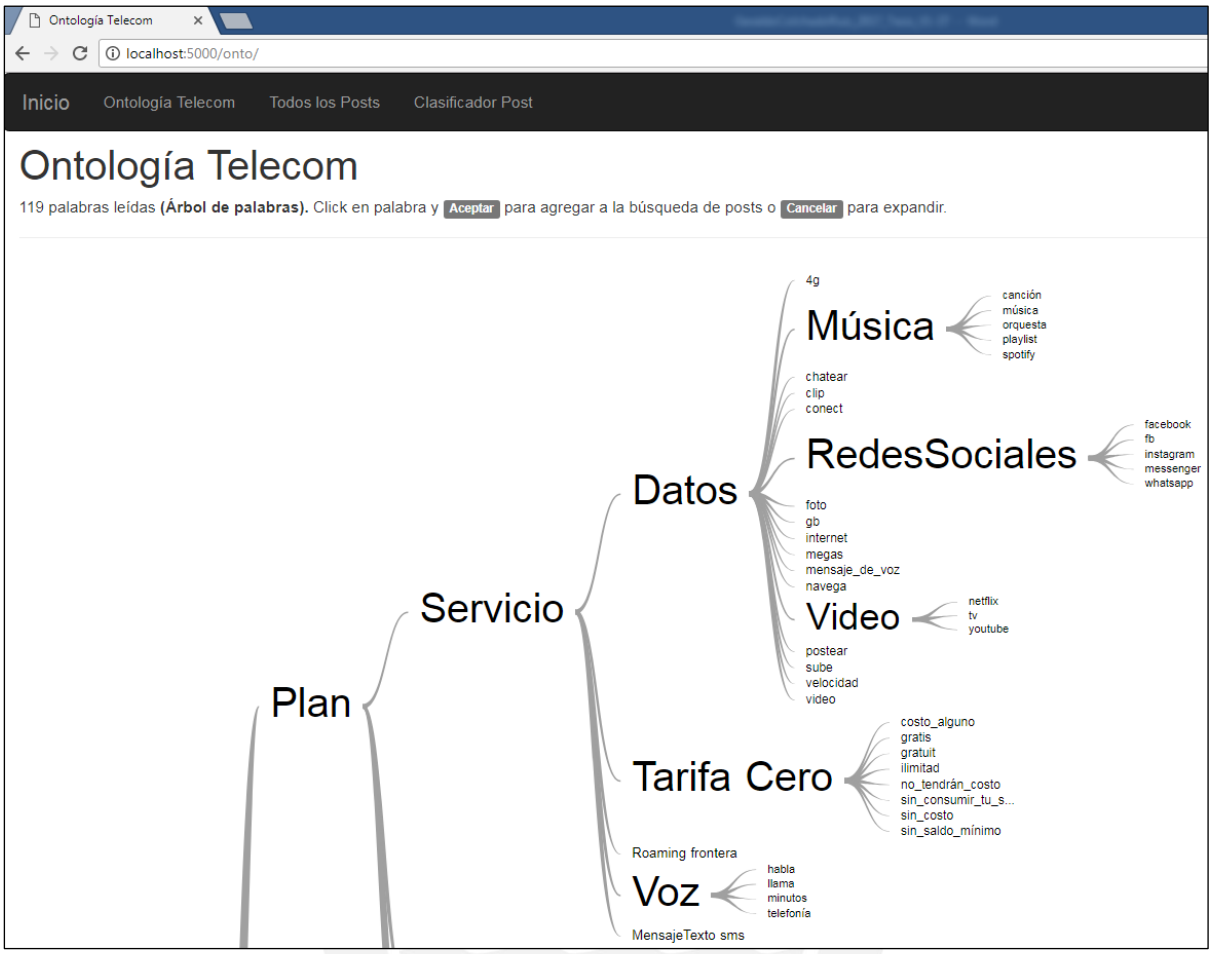


Figura 42: Árbol de palabras o *WordTree* de ontología en telecomunicaciones (Interfaz de búsqueda)

En el árbol de palabras de la Figura 42, todos los conceptos empiezan con la primera letra en mayúsculas (Ejemplo: Datos, Voz, Servicio) y se relacionan con otros conceptos por la jerarquía del árbol empezando de izquierda a derecha, en cambio todas las palabras están en letras minúsculas y están siempre en el último nivel jerárquico del árbol a la derecha; si el usuario elige una palabra sólo se incluirá esa palabra en el criterio de búsqueda con el operador Y (AND), en cambio si elige un concepto se incluirán en el criterio de búsqueda, como un grupo entre paréntesis, todas las palabras que incluye el concepto y que están a la derecha con el operador OR (O) y luego todo el grupo con el operador Y (AND); en la Tabla 36 se muestran dos ejemplos para construir automáticamente el criterio o cadena de búsqueda final por la aplicación web.

Tabla 36: Ejemplos de construcción de criterio de búsqueda de *posts*

Pregunta Relevante	**Palabra elegida**	**Tipo**	**Cadena de búsqueda**	**Cadena de búsqueda final**
Buscar *posts* sobre whatsapp ilimitado	whatsapp	palabra	whatsapp	whatsapp **AND** (“costo alguno” OR gratis OR gratuit OR ilimitad OR “no tendrán costo” OR “sin consumir tu saldo” OR “sin costo” OR “sin saldo mínimo”)
	Cero	concepto	(“costo alguno” OR gratis OR gratuit OR ilimitad OR “no tendrán costo” OR “sin consumir tu saldo” OR “sin costo” OR “sin saldo mínimo”)	

Buscar *posts* sobre Redes Sociales ilimitadas	RedesSociales	concepto	(facebook OR fb OR Instagram OR messenger OR whatsapp)	(facebook OR fb OR Instagram OR messenger OR whatsapp) **AND** (“costo alguno” OR gratis OR gratuit OR ilimitad OR “no tendrán costo” OR “sin consumir tu saldo” OR “sin costo” OR “sin saldo mínimo”)
	Cero	concepto	(“costo alguno” OR gratis OR gratuit OR ilimitad OR “no tendrán costo” OR “sin consumir tu saldo” OR “sin costo” OR “sin saldo mínimo”)	

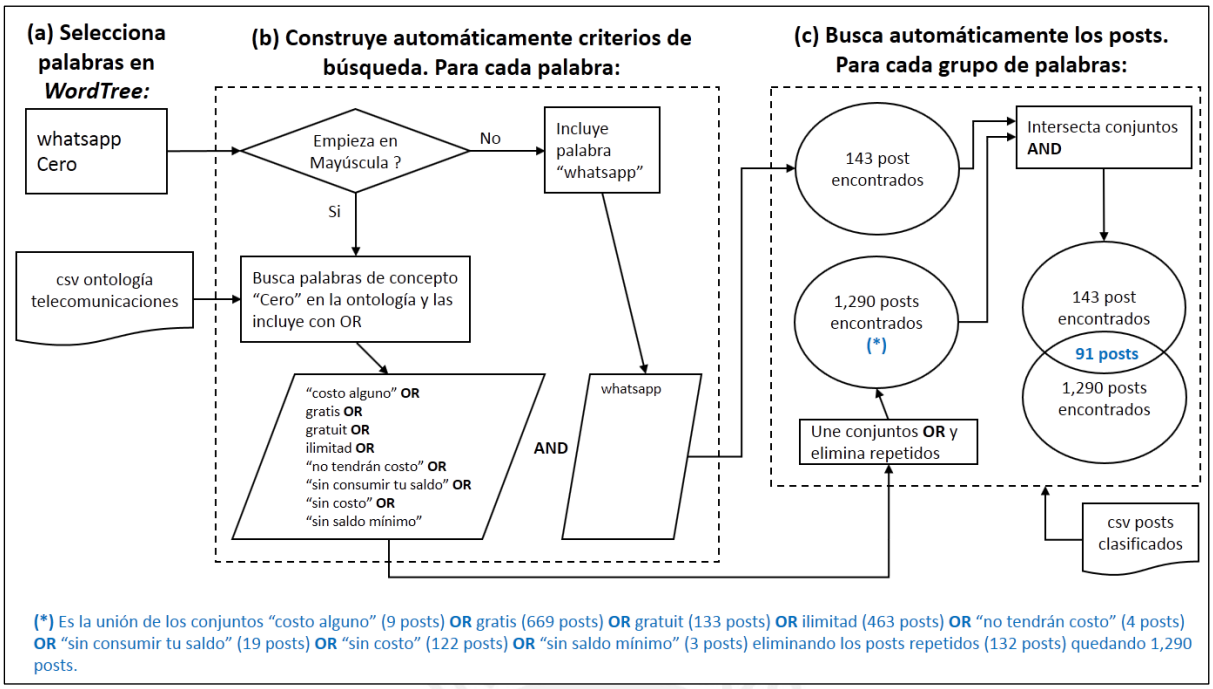
7.2.2. Búsqueda de *posts*

Una vez construida la cadena de búsqueda final se utilizó el archivo CSV de *posts* clasificados para realizar la búsqueda de *posts*, en la Tabla 37 se muestran los registros de dos *posts* en el archivo de *posts* clasificados, en la columna (c) se muestra el identificador único del *post* de Facebook, en la columna (a) se muestra la palabra de la ontología de dominio en telecomunicaciones que es la última de la derecha (Ejemplo: video) y la cadena del lado izquierdo corresponde a la jerarquía de conceptos (Ejemplo: /Operador/Plan/Servicio/Datos/), finalmente en la columna (b) se muestra la palabra original del *post* de Facebook que sirvió para realizar la clasificación.

Tabla 37: Ejemplo de registros de archivo de *posts* clasificados

Palabra Ontología (a)	**Palabra *Post* (b)**	**Identificador de *Post* (c)**
/Operador/Plan/Servicio/Datos/video	videollama	188657494839544_417177178654240
/Operador/Plan/Servicio/Datos/foto	fotos	188657494839544_417177178654240
/Operador/Plan/Servicio/Datos/video	videos	188657494839544_417177178654240
/Operador/Plan/Servicio/Datos/RedesSociales/whatsapp	WHATSAPP	188657494839544_417177178654240
/Operador/Plan/Servicio/Tarifa/Cero/ilimitad	ILIMITADO	188657494839544_417177178654240
/Operador/Plan/Servicio/Tarifa/Cero/gratis	GRATIS	188657494839544_417177178654240
/Operador/Plan/Control/recarga /Operador/Plan/Prepago/recarga	recargar	188657494839544_417177178654240
/Operador/Beneficio/Agradecimiento/feliz_día	feliz día	188657494839544_417177178654240
/Operador/Portabilidad/migra	Migra	198975483478832_1416015081774860
/Operador/Plan/Servicio/Datos/Video/youtube	YouTube	198975483478832_1416015081774860
/Operador/Plan/Servicio/Datos/RedesSociales/instagram	Instagram	198975483478832_1416015081774860
/Operador/Plan/Servicio/Datos/Música/spotify	Spotify	198975483478832_1416015081774860
/Operador/Plan/Servicio/Tarifa/Cero/ilimitad	ilimitados	198975483478832_1416015081774860
/Operador/Plan/Servicio/Datos/conect	conectados	198975483478832_1416015081774860

En la Figura 43 se muestra el algoritmo de búsqueda mediante un ejemplo donde el usuario selecciona las palabras whatsapp y Cero del WordTree (a), luego se construye automáticamente los criterios de búsqueda con una lógica diferente para palabras y para conceptos (b) y finalmente se aplican las operaciones de unión e intersección de conjuntos para procesar los criterios de búsqueda OR y AND respectivamente obteniendo 91 *posts* como resultado (c). En la Figura 44 se muestra un segmento de código fuente en Python que implementa el algoritmo de búsqueda de *posts* de Figura 43.



(*) Es la unión de los conjuntos "costo alguno" (9 posts) **OR** gratis (669 posts) **OR** gratuit (133 posts) **OR** ilimitad (463 posts) **OR** "no tendrán costo" (4 posts) **OR** "sin consumir tu saldo" (19 posts) **OR** "sin costo" (122 posts) **OR** "sin saldo mínimo" (3 posts) eliminando los posts repetidos (132 posts) quedando 1,290 posts.

Figura 43: Algoritmo de búsqueda de *posts*

```
def html_busqueda(titulo_html, pag, palabra_word_tree):
    try:
        if palabra_word_tree == "TODOS":
            ...
        else:
            # Lógica de búsquedas con AND
            df_items = pd.DataFrame(palabra_word_tree.split("."))
            regs_items = df_items[0].count()
            if regs_items > 1:
                regs_items = regs_items - 1
            j = 0
            ...
            df_res = pd.DataFrame()
            df_ids_res = pd.DataFrame()
            df_ids_fin = pd.DataFrame()
            while j < regs_items:
                ...
                # Busca los posts que tengan la palabra del word tree y genera dataframe
                for i in df_posts_clasif.index:
                    if str(df_posts_clasif["Palabra"][i]).find("/" + df_items[0][j]) >= 0:
                        cadena_ids = cadena_ids + df_posts_clasif["id"][i] + " "
                df_ids_res = pd.DataFrame(cadena_ids.split())
                df_ids_res.columns = ["id"]
                df_ids_res.drop_duplicates(inplace = True)
                df_ids_res.set_index("id", inplace=True)
                if j == 0:
                    df_ids_fin = df_ids_res.copy(deep=True)
                else:
                    df_ids_fin = df_ids_fin.join(df_ids_res, how="inner").copy(deep=True)
                j = j + 1
            df_res = df_ids_fin.join(df_posts_con_index, how="inner").copy(deep=True)
            df_res.sort_values(by="created_time",ascending=False,inplace=True)
            df_res.reset_index(inplace=True)
            total_post = df_res["id"].count()
            # Fin Lógica de búsquedas con AND
    except:
        ...
    return cadena_html
```

Figura 44: Segmento código fuente en Python para algoritmo de búsqueda de *posts*

En la Figura 45 se muestran los 91 *posts* encontrados (a), luego de ejecutado en Python el algoritmo de búsqueda de *posts*, donde se muestra sombreado en un color diferente cada palabra o concepto que seleccionó previamente el usuario en el *WordTree* para una rápida identificación visual, adicionalmente se muestran paginados los resultados cada 7 *posts* y se proporcionan botones para navegar entre páginas (b), para cada *post* encontrado se muestra su texto completo (Columna "*Post*"), datos de publicación (Columnas "Fecha" y "Operador") e información respecto a sus comentarios totales, positivos, negativos y neutros (Columnas "N° Coment …" y "% Coment …") y también se incluye un enlace al *post* original (Columna "Facebook").

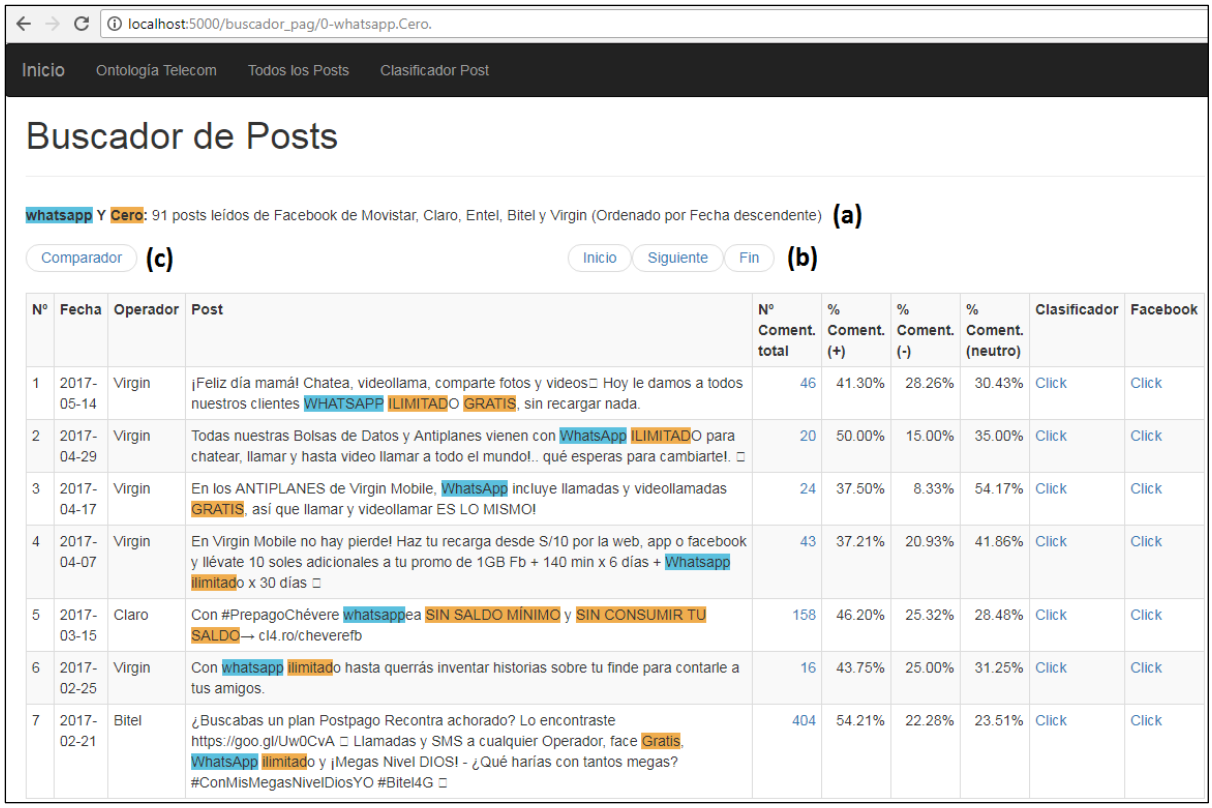


Figura 45: Resultados de la búsqueda de *posts*

En la Figura 45, si el usuario desea ver la clasificación completa de un *post* puede seleccionar un enlace (Columna "Clasificador") y obtener los resultados en formato *WordTree* tal como se aprecia en la Figura 46 donde se seleccionó el *post* N° 7 de Bitel; si el usuario desea ver los comentarios de un *post* con su polaridad puede seleccionar un enlace (Figura 45 Columna "N° Coment. total") y obtener los resultados mostrados en la Figura 47 donde se seleccionó el *post* N° 5 de Claro (Figura 45) y se aprecia sombreado en color verde claro los comentarios positivos (Columna Polaridad Puntaje > 0), en color rojo claro los comentarios negativos (Columna Polaridad Puntaje < 0) y en gris los comentarios neutros (Columna Polaridad Puntaje = 0).

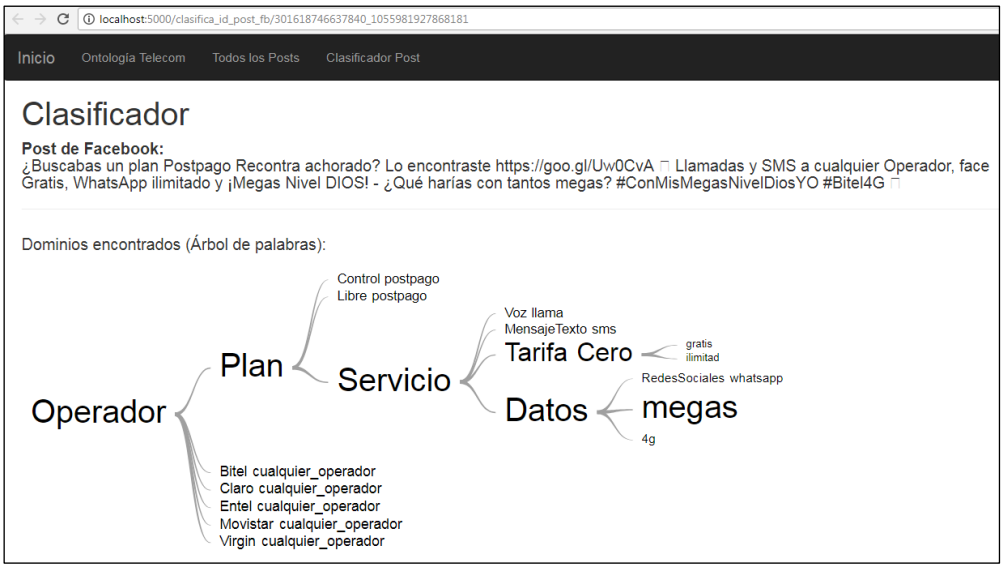
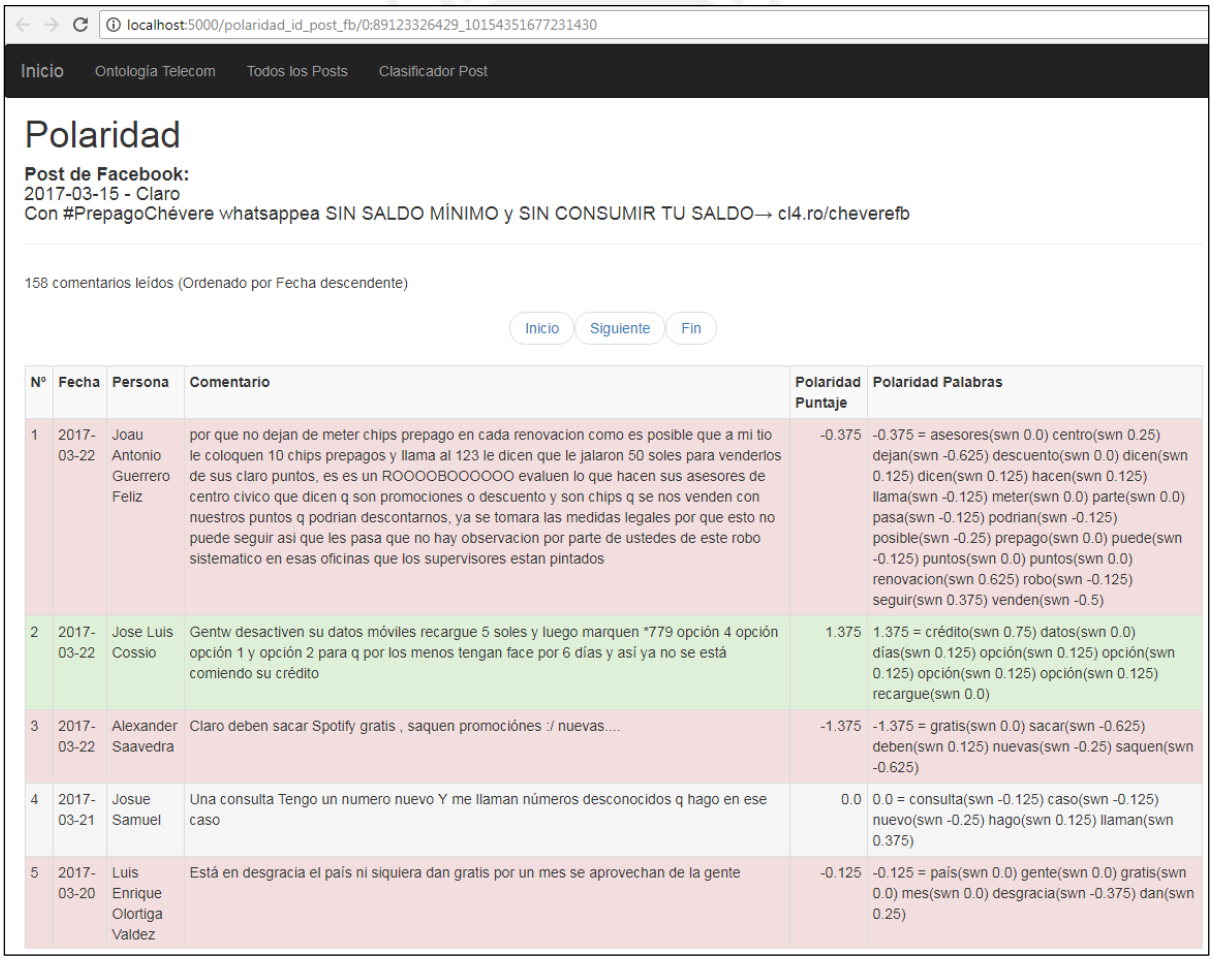


Figura 46: Clasificador de *post*



Nº	Fecha	Persona	Comentario	Polaridad Puntaje	Polaridad Palabras
1	2017-03-22	Joau Antonio Guerrero Feliz	por que no dejan de meter chips prepago en cada renovacion como es posible que a mi tio le coloquen 10 chips prepagos y llama al 123 le dicen que le jalaron 50 soles para venderlos de sus claro puntos, es es un ROOOOBOOOOOO evaluen lo que hacen sus asesores de centro civico que dicen q son promociones o descuento y son chips q se nos venden con nuestros puntos q podrian descontarnos, ya se tomara las medidas legales por que esto no puede seguir asi que les pasa que no hay observacion por parte de ustedes de este robo sistematico en esas oficinas que los supervisores estan pintados	-0.375	-0.375 = asesores(swn 0.0) centro(swn 0.25) dejan(swn -0.625) descuento(swn 0.0) dicen(swn 0.125) dicen(swn 0.125) hacen(swn 0.125) llama(swn -0.125) meter(swn 0.0) parte(swn 0.0) pasa(swn -0.125) podrian(swn -0.125) posible(swn -0.25) prepago(swn 0.0) puede(swn -0.125) puntos(swn 0.0) puntos(swn 0.0) renovacion(swn 0.625) robo(swn -0.125) seguir(swn 0.375) venden(swn -0.5)
2	2017-03-22	Jose Luis Cossio	Gentw desactiven su datos móviles recargue 5 soles y luego marquen *779 opción 4 opción opción 1 y opción 2 para q por los menos tengan face por 6 días y así ya no se está comiendo su crédito	1.375	1.375 = crédito(swn 0.75) datos(swn 0.0) días(swn 0.125) opción(swn 0.125) opción(swn 0.125) opción(swn 0.125) opción(swn 0.125) recargue(swn 0.0)
3	2017-03-22	Alexander Saavedra	Claro deben sacar Spotify gratis , saquen promociónes :/ nuevas....	-1.375	-1.375 = gratis(swn 0.0) sacar(swn -0.625) deben(swn 0.125) nuevas(swn -0.25) saquen(swn -0.625)
4	2017-03-21	Josue Samuel	Una consulta Tengo un numero nuevo Y me llaman números desconocidos q hago en ese caso	0.0	0.0 = consulta(swn -0.125) caso(swn -0.125) nuevo(swn -0.25) hago(swn 0.125) llaman(swn 0.375)
5	2017-03-20	Luis Enrique Olortiga Valdez	Está en desgracia el país ni siquiera dan gratis por un mes se aprovechan de la gente	-0.125	-0.125 = país(swn 0.0) gente(swn 0.0) gratis(swn 0.0) mes(swn 0.0) desgracia(swn -0.375) dan(swn 0.25)

Figura 47: Polaridad de comentarios de *post*

7.2.3. Comparación de *posts*

Si el usuario desea comparar los *posts* encontrados, en la pantalla de resultados de búsqueda de *posts* puede seleccionar el botón "Comparador" (Figura 45 (c)) y los *posts* encontrados son procesados por un comparador de *posts* según Figura 48 donde se generan los tipos de gráficos *pie* por operador y de barras por año y operador por el total de *posts* (a), por total de comentarios (b.1), por total de comentarios + (b.2), por total de comentarios - (b.3) y finalmente tres listas *Top* 10 de *posts* por mayor número de comentarios (c.1), por mayor número de comentarios + (c.2) y por mayor número de comentarios - (c.3). En la Figura 49 se muestra un ejemplo con el total de *posts* por operador para el criterio de búsqueda whatsapp Y Cero.

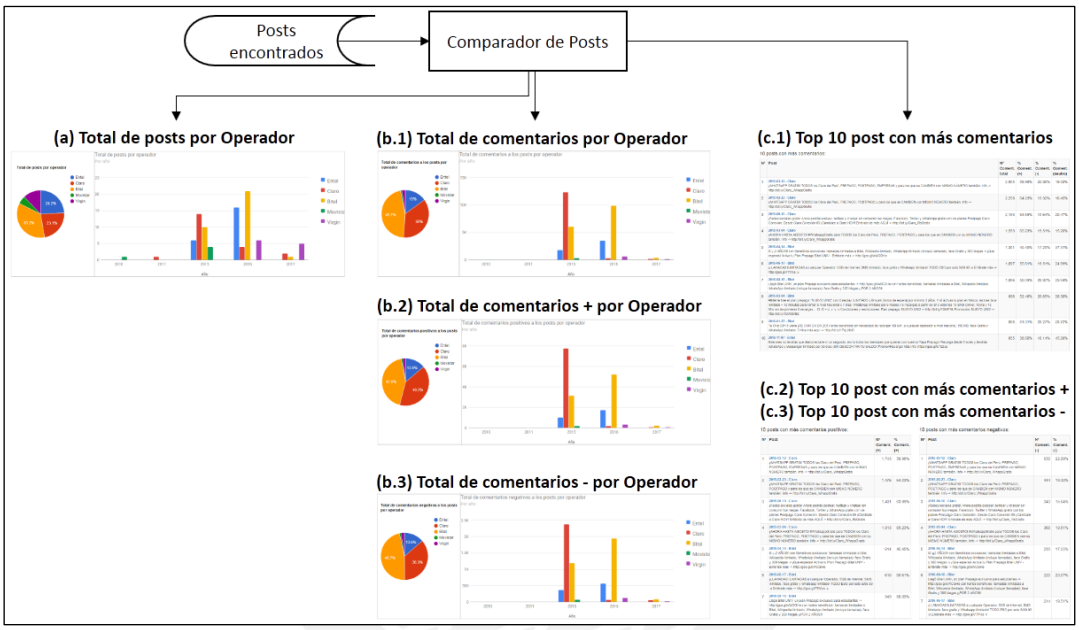


Figura 48: Diagrama de componentes de aplicación web de comparación de *posts*

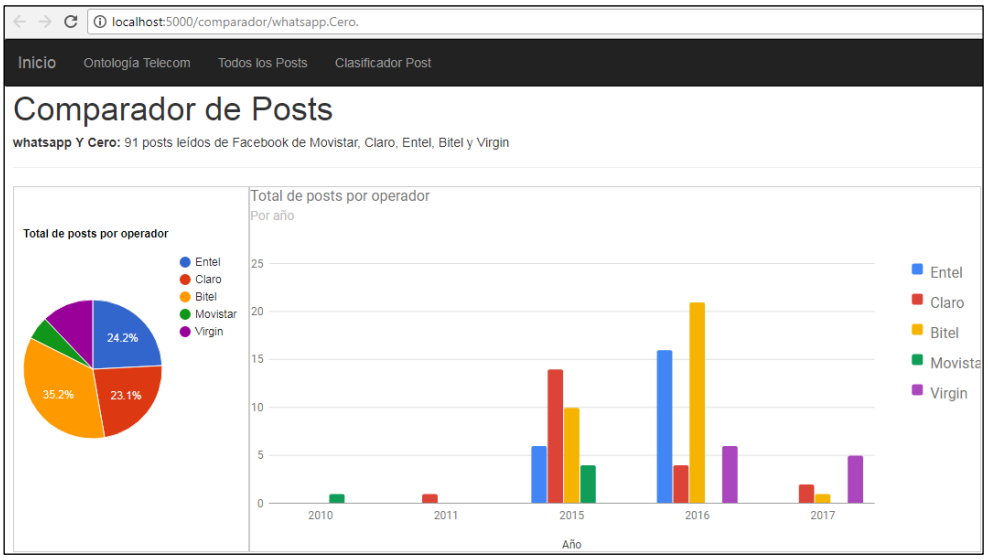


Figura 49: Total de *posts* por operador

En la Figura 50 se muestra un ejemplo con el Top 10 de *posts* con mayor número de comentarios para el criterio de búsqueda whatsapp y Cero.

localhost:5000/comparador/whatsapp.Cero.

Inicio | Ontología Telecom | Todos los Posts | Clasificador Post

Comparador de Posts

whatsapp Y Cero: 91 posts leídos de Facebook de Movistar, Claro, Entel, Bitel y Virgin

10 posts con más comentarios:

Nº	Post	Nº Coment. total	% Coment. (+)	% Coment. (-)	% Coment. (neutro)
1	2015-03-12 - Claro ¡WHATSAPP GRATIS! TODOS los Claro del Perú. PREPAGO, POSTPAGO, EMPRESAS y para los que se CAMBIEN con MISMO NÚMERO también. Info -> http://bit.ly/Claro_WhappGratis	2,955	58.98%	22.00%	19.02%
2	2015-02-23 - Claro ¡WHATSAPP GRATIS! TODOS los Claro del Perú. PREPAGO, POSTPAGO y para los que se CAMBIEN con MISMO NÚMERO también. Info -> http://bit.ly/Claro_WhappGratis	2,298	64.23%	19.32%	16.45%
3	2015-06-10 - Claro ¡Redes sociales gratis! Ahora podrás postear, twittear y chatear sin consumir tus megas. Facebook, Twitter y WhatsApp gratis con los planes Postpago Claro Conexión. Desde Claro Conexión 69 ¡Cámbiate a Claro HOY! Entérate de más AQUÍ -> http://bit.ly/Claro_RsGratis	2,193	63.89%	15.64%	20.47%

Figura 50: *Top* 10 de *posts* con mayor número de comentarios

El comparador de *posts* fue programado en Python utilizando la librería gráfica Google charts y la librería de estilos bootstrap, en la Figura 51 se muestra un segmento de código fuente del comparador de *posts*.

```
def html_comparador(titulo_html, palabra_word_tree):
    ...
    graficos = \
        "<script type='text/javascript' src='/static/js/GoogleCharts.js'></script>\n" \
        "<script type='text/javascript'>\n" \
        ...
        "  function drawChart1() {\n" \
        ...
        "    var chart = new google.visualization.PieChart(document.getElementById('piechart1'));\n" \
        ...
        "  }\n" \
        ...
        "</script>\n"
    ...
    top10_1 = ""
    ...
    if total_post > 0:
        df_top10_1 = df_res.sort_values("comentarios_pos", ascending=False).head(10)
        j = 1
        for i in df_top10_1.index:
            top10_1 = top10_1 + "<tr>" + "" \
                    + "<td>" + "{0:,}".format(j) + "</td>" \
                    + "<td style='font-size:12px;'><b>" + "<a href='/polaridad_id_post_fb/0:" \
                    + str(df_top10_1["id"][i]) + "' target='_blank' title='Ver comentarios en otra pestaña'>" \
                    + str(df_top10_1["created_time"][i])[0:10] + " - " \
                    + str(df_top10_1["empresa"][i]) + "</a></b><br>" + str(df_top10_1["message"][i]) + "</td>" \
                    + "<td align='right'>" + "{0:,}".format(df_top10_1["comentarios_pos"][i]) + "</td>" \
                    + "<td align='right'>" \
                    + "{0:.2f}%".format(df_top10_1["comentarios_pos"][i] / df_top10_1["comentarios"][i] * 100) + "</td>" \
                    + "</tr>\n"
            j = j + 1
        ...
    return cadena_html
```

Figura 51: Segmento de código fuente de comparador de *posts*

7.3. Discusión

Según la revisión sistemática realizada en esta tesis (Tabla 21 de capítulo 3), los dos métodos y procedimientos más utilizados para la fase de diseminación del proceso de inteligencia competitiva son los tableros con información gráfica más importante (*Dashboards*) con 40% de 10 investigaciones revisadas y las búsquedas con 30%; en esta tesis se implementaron ambos empezando por la búsqueda donde se creó una interfaz gráfica de árbol de palabras o *WordTree* dado que su estructura podía contener jerarquías, conceptos y palabras que identifican a los conceptos definidos en la ontología de telecomunicaciones, el otro motivo de elección del *WordTree* fue que la cantidad de palabras y conceptos de la ontología es reducida (119 palabras con 27 conceptos y 6 relaciones en 5 niveles jerárquicos) lo que permite visualizarse en pantalla en un tamaño adecuado para la lectura y amigable para el usuario; luego de realizada la búsqueda y mostrado los *posts* encontrados, el usuario puede elegir la opción de comparación de *posts* que es mostrada en una pantalla web tipo tablero (*dashboard*) con información gráfica más importante y relevante que le permite comparar por año y operador la cantidad de *posts*, comentarios, comentarios positivos, comentario negativos y un *Top 10* de *posts* con más comentarios, con más comentarios positivos y con más comentarios negativos.

En el *dashboard* se utilizó gráficos con línea de tiempo (por año) donde el usuario puede ver la evolución y tendencias en el tiempo por operador de cantidad de *posts*, comentarios, comentarios positivos y comentarios negativos y realizar un análisis de eventos en el tiempo (*Event Timeline Analysis*) para intentar explicar los valores en el tiempo, los eventos en este caso son los *posts* que están identificados por una fecha de publicación, el operador que lo publicó y el texto de la promoción. *Event Timeline Analysis* es uno de los métodos y procedimientos utilizados en la fase de análisis del proceso de inteligencia competitiva y fue usado en el 11% de 19 investigaciones de la revisión sistemática realizada en esta tesis (Tabla 20 de capítulo 3).

Con la herramienta de búsqueda y el *dashboard* implementados y mostrados en este capítulo 7 se cubre completamente la cuarta fase del proceso de inteligencia competitiva planteada en esta tesis, denominada “Diseminación”, que está acorde con lo planteado por Arroyo Varela (2005) y lo mencionado en las 6 investigaciones de la revisión sistemática realizada en esta tesis y mostradas en Tabla 18 de capítulo 3 (i02 (Abdellaoui y Nader 2015), i03 (Spruit y Cepoi 2015), i07 (Chouder y Chalal 2014), i08 (Olszak 2014), i15 (Chen et al. 2008), i18 (Li et al. 2007)) que consiste en comunicar a los tomadores de decisiones información relevante que dé respuesta a sus necesidades de información para tener inteligencia competitiva, de la información obtenida se pueden generar otras necesidades de información que el tomador de decisiones puede absolver haciendo nuevas búsquedas con otros criterios volviendo cíclico el proceso de inteligencia competitiva.

8. CONCLUSIONES Y TRABAJOS FUTUROS

8.1. Conclusiones

Respecto al objetivo general, en esta tesis se diseñó e implementó un proceso cíclico de inteligencia competitiva en 4 fases tomando como referencia la revisión sistemática realizada en el estado del arte y la revisión de literatura del marco conceptual.

En la fase 1 "Identificación de necesidades de información" el tomador de decisiones del operador definió sus preguntas relevantes respecto a las promociones en Facebook de los 5 operadores.

En la fase 2 "Recolección" se extrajeron y limpiaron todos los *posts* y comentarios de Facebook de los 5 operadores obteniendo el corpus.

En la fase 3 "Análisis" se creó la ontología de dominio de telecomunicaciones y se usó para clasificar con conceptos cada *post* y se creó la ontología de dominio de polaridad y se usó para calcular la polaridad positiva, negativa o neutra de cada comentario.

En la fase 4 "Diseminación" se creó una aplicación web para que el tomador de decisiones realice búsquedas de *posts* y se creó otra aplicación web para comparar los *posts* encontrados y tener respuesta a las preguntas relevantes para tener inteligencia competitiva pudiéndose generar nuevas necesidades de información que pueden absolverse realizando nuevas búsquedas con otros criterios haciendo cíclico el proceso.

El proceso cíclico de inteligencia competitiva diseñado puede ser usado en otros contextos donde se tengan varios competidores que ofrezcan productos o servicios equivalentes y publiquen en redes sociales para que puedan compararse, también podría aplicarse en empresas de telecomunicaciones de otros países.

Respecto al objetivo específico 1, en esta tesis se creó un programa extractor y limpiador de *posts* y comentarios y se extrajeron y limpiaron todos los *posts* y comentarios de Facebook de los 5 operadores generando un corpus de 15,634 *posts* y 1,411,698 comentarios que se almacenaron en archivos CSV. El programa extractor y limpiador de datos utiliza el API de Facebook y es genérico por lo que puede ser usado por otros investigadores que requieran extraer datos de Facebook, en la tesis se publicó el 100% del código fuente.

Respecto al objetivo específico 2, en esta tesis se creó y documentó todo el proceso manual de creación de la ontología de dominio en telecomunicaciones que contiene 119 palabras, 27 conceptos y 6 relaciones en 5 niveles jerárquicos siendo necesario sólo revisar el texto completo del 0.7% de los *posts* (110 *posts*) que permitió clasificar con conceptos al 67% del total de *posts*. El proceso manual de creación de la ontología puede ser replicado por otros investigadores para crear ontologías de dominio en otros contextos donde se pueda identificar palabras relacionadas a conceptos.

Respecto al objetivo específico 3, en esta tesis se creó y documentó todo el proceso semiautomático de creación de la ontología de dominio en polaridad que utilizó las bases de datos léxicas *WordNet* en español y *SentiWordNet* en inglés para poblar 9,344 palabras con su polaridad en la ontología y que permitió clasificar con polaridad positiva, negativa o neutra al 80.25% del total de comentarios. Este proceso semiautomático puede ser replicado por otros investigadores especialmente en otros idiomas diferentes al inglés y español ya que *WordNet* está disponible en varios idiomas y está

completamente integrado a *SentiWordNet* en inglés mediante *SynSets* usando la librería NLTK de Python.

Respecto al objetivo específico 4, en esta tesis se creó una aplicación web para realizar búsquedas de *posts* que cuenta con una interfaz de búsqueda de árbol de palabras o WordTree que muestra todos los conceptos y palabras de la ontología de telecomunicaciones y el tomador de decisiones puede seleccionar visualmente las palabras y conceptos a incluir en el criterio de búsqueda de *posts* que correspondan a su pregunta relevante, los resultados obtenidos pueden ser comparados mediante una aplicación web que fue creada para tal fin y que muestra un *dashboard* con la información más relevante donde el tomador de decisiones podría realizar un análisis de eventos en el tiempo (*Event Timeline Analysis*) para interpretar las tendencias. Los *dashboards* y las búsquedas son las herramientas más utilizadas para comunicar información relevante a los tomadores de decisiones según la revisión sistemática realizada en esta tesis.

Considerando lo expuesto, se concluye que se cumplió con los objetivos específicos que contribuyeron a cumplir con el objetivo general de la tesis de "Diseñar e implementar un proceso de inteligencia competitiva para los operadores, a partir de las publicaciones de promociones y sus comentarios registrados en Facebook por los operadores y sus seguidores, que les permita responder preguntas relevantes y compararse con la competencia en el dominio de las telecomunicaciones para que tengan información relevante en la toma de decisiones y contribuya al diseño de promociones más competitivas". Este proceso de inteligencia competitiva puede beneficiar, en principio, a todos los operadores de Telecomunicaciones del Perú.

8.2. Trabajos futuros

Respecto a la ontología de dominio en telecomunicaciones, para aumentar el porcentaje de *posts* clasificados de 67% a más, se podría revisar manualmente como mínimo el 0.7% de los *posts* no clasificados por la ontología (aprox. 36 *posts*) en orden descendente por fecha de publicación y agregar nuevas palabras, conceptos y relaciones a la ontología. También se puede almacenar en la ontología las formas base de las palabras (*lemmas*) para que puedan cubrirse todas las variantes de forma de las palabras en la clasificación de los posts mediante el uso de un *lemmatizer*. Adicionalmente se podría validar la relevancia de los conceptos de la ontología con tomadores de decisiones reales de los operadores para refinar la ontología.

Respecto a la ontología de dominio en polaridad, para mejorar la precisión en el cálculo de polaridad de los comentarios publicados por los seguidores se podría incluir algoritmos para desambiguar una palabra, es decir determinar el significado exacto de la palabra en el texto incluyendo análisis de casos de inversión de polaridad; debido a la amplitud de este tema los algoritmos no fueron incluidos. De igual manera se podrían incluir algoritmos para *lemmatization* para almacenar los *lemmas* en la ontología y mediante *lemmatization* en la clasificación aumentar el porcentaje de comentarios clasificados. También pueden considerarse jergas, abreviaturas y emoticones los cuales no fueron incluidos en la ontología.

Respecto a la aplicación web de comparación de resultados, para mejorar el análisis de eventos en el tiempo (*Event Timeline Analysis*), se podría desagregar los gráficos de línea de tiempo en el eje X de año a mes y también se podría cambiar de formato de barras a líneas donde se visualizan mejor las tendencias.

9. REFERENCIAS BIBLIOGRÁFICAS

Abdellaoui, Sabrina; Nader, Fahima (2015): Semantic data warehouse at the heart of competitive intelligence systems: Design approach. In: Information Systems and Economic Intelligence (SIIE), 2015 6th International Conference, pp. 141–145. Hammamet, Tunisia. IEEE. DOI: 10.1109/ISEI.2015.7358736.

Arp, Robert; Smith, Barry; Spear, Andrew D. (2015): Building Ontologies with Basic Formal Ontology. United States of America: The MIT Press.

Arroyo Varela, Silvia R. (2005): Inteligencia competitiva: Una herramienta clave en la estrategia empresarial. Madrid: Ediciones Pirámide.

Chakraborti, S.; Dey, S. (2014): Multi-document Text Summarization for Competitor Intelligence: A Methodology. In: Computational and Business Intelligence (ISCBI), 2014 2nd International Symposium, pp. 97-100. New Delhi, India. IEEE. DOI: 10.1109/ISCBI.2014.28.

Chen, Y.; Jin, P.; Yue, L. (2008): Ontology-Driven Extraction of Enterprise Competitive Intelligence in the Internet. In: Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference, vol. 2, pp. 35-38. Sanya, China: IEEE. DOI: 10.1109/FGCNS.2008.72.

Chouder, Mohamed Lamine; Chalal, Rachid (2014): Models and tools support to the Competitive Intelligence process. In: ISKO-Maghreb: Concepts and Tools for knowledge Management (ISKO-Maghreb), 2014 4th International Symposium, pp. 1-7. Algiers, Algeria: IEEE. DOI: 10.1109/ISKO-Maghreb.2014.7033466.

Dai, Yue; Kakkonen, Tuomo; Sutinen, Erkki (2011): SoMEST - A Model for Detecting Competitive Intelligence from Social Media. In: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, pp. 241-248. Tampere, Finland. ACM. DOI: 10.1145/2181037.2181078.

Del-Fresno-García, Miguel (2011): INFOSOCIABILIDAD: MONITORIZACIÓN E INVESTIGACIÓN EN LA WEB 2.0 PARA LA TOMA DE DECISIONES. In: *Infosociability: Monitoring and research on the web 2.0 for decision making, vol.* 20 (5), pp. 548–554. DOI: 10.3145/epi.2011.sep.09.

García Alsina, Monserrat; Ortoll Espinet, Eva (2012): La inteligencia competitiva: evolución histórica y fundamentos teóricos. Gijón, Asturias: Ediciones Trea.

Norma UNE 166006, 16/03/2011: Gestión de la I+D+i: Sistema de vigilancia tecnológica e inteligencia competitiva. Asociación Española de Normalización y Certificación (AENOR). España.

Gógova, Sonia (2015): Inteligencia Competitiva: ¿Espías? ¿Oráculos? ¿Estrategas? España: Ediciones Díaz de Santos.

Gómez-Pérez, Asunción; Fernández-López, Mariano; Corcho, Oscar (2003): Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. London: Springer (Advanced information and knowledge processing).

Gruber, Thomas R. (1995): Toward Principles for the Design of Ontologies Used for Knowledge Sharing. In: *International Journal of Human-Computer Studies, vol.* 43 (5-6), pp. 907–928. DOI: 10.1006/ijhc.1995.1081.

Hassan, Thomas; Cruz, Christophe; Bertaux, Aurélie (2017): Ontology-based Approach for Unsupervised and Adaptive Focused Crawling. In: SBD '17, Proceedings of The International Workshop on Semantic Big Data, pp. 2:1–2:6. Chicago, Illinois. ACM. DOI: 10.1145/3066911.3066912.

Horridge, Matthew (2011): A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools. En colaboración con Sebastian Brandt. 1.3 ed.: The University Of Manchester.

Jin, L.; Yan, D. (2010): Post-controlled vocabulary compiling in competitive intelligence system. In: Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference, pp. 560-563. Chengdu, China. IEEE. DOI: 10.1109/ICIME.2010.5478217.

Kitchenham, Barbara; Charters, Stuart (2007): Guidelines for performing Systematic Literature Reviews in Software Engineering. 2.3 ed. UK: Keele University, University of Durham.

Li, Jiao; Huang, Minlie; Zhu, Xiaoyan (2007): An Ontology-Based Mining System for Competitive Intelligence in Neuroscience. In: Web Intelligence Meets Brain Informatics, First WICI International Workshop WImBI 2006, LNAI 4845, pp. 291–304. Berlin Heidelberg: Springer.

Liu, C.; He, J. (2009): Application of Domain Ontology-Based on Semantic Web Technology. In: Computational Intelligence and Natural Computing, 2009. CINC'09. International Conference, vol. 1, pp. 133-136. Wuhan, China. IEEE. DOI: 10.1109/CINC.2009.125.

Liu, Chunnian; Yang, Dehui; Wang, Yonglong (2011): Domain ontology and semantic web applications for study of web competitive intelligence analysis system. In: *International Journal of Web Science, vol.* 1 (1-2), pp. 99–113. DOI: 10.1504/IJWS.2011.044083.

Muñoz Cañavate, Antonio (2012): Recursos de información para la inteligencia competitiva: Una guía para la toma de decisiones. Gijón, Asturias: Ediciones Trea.

Nagano, S.; Inaba, M.; Mizoguchi, Y.; Iida, T.; Kawamura, T. (2008): Ontology-Based Topic Extraction Service from Weblogs. In: Semantic Computing, 2008 IEEE International Conference, pp. 468-475. Santa Monica, CA, USA. IEEE. DOI: 10.1109/ICSC.2008.80.

Neches, Robert; Fikes, Richard; Finin, Tim; Gruber, Thomas; Patil, Ramesh; Senator, Ted; Swartout, William R. (1991): Enabling technology for knowledge sharing. In: *AI Magazine, vol.* 12 (3), pp. 36–56. DOI: 10.1609/aimag.v12i3.902.

Nemrava, Jan; Kliegr, Tomáš; Svátek, Vojtěch; Ralbovský, Martin; Šplíchal, Jiˇrí; Vejlupek, Tomáš (2008): Semantic Annotation and Linking of Competitive Intelligence Reports for Business Clusters. In: Proceedings of the first international workshop on Ontology-supported Business Intelligence, pp. 9:1–9:5. Karlsruhe, Germany. ACM. DOI: 10.1145/1452567.1452576.

Olszak, Celina M. (2014): An Overview of Information Tools and Technologies for Competitive Intelligence Building: Theoretical Approach. In: *Issues in Informing Science & Information Technology, vol.* 11, pp. 139–153.

Spruit, Marco; Cepoi, Alex (2015): CIRA: A Competitive Intelligence reference Architecture for dynamic solutions. In: Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference, vol. 1, pp. 249-258. Lisbon, Portugal. IEEE.

Vasilateanu, A.; Goga, N.; Tanase, E. A.; Marin, I. (2015): Enterprise domain ontology learning from web-based corpus. In: Computing, Communication and Networking Technologies (ICCCNT), 2015 6th International Conference, pp. 1-6. Denton, TX, USA. IEEE. DOI: 10.1109/ICCCNT.2015.7395227.

Wongthongtham, P.; Abu-Salih, B. (2015): Ontology and trust based data warehouse in new generation of business intelligence: State-of-the-art, challenges, and opportunities. In: Industrial Informatics (INDIN), 2015 IEEE 13th International Conference, pp. 476-483. Cambridge, UK. IEEE. DOI: 10.1109/INDIN.2015.7281780.

Zhang, Y.; Wu, J.; Wang, C. (2007): Automatic Competitive Intelligence Collection Based on Semantic Web Mining. In: Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference, pp. 3701-3704. Shanghai, China. IEEE. DOI: 10.1109/WICOM.2007.915.

Zhao, Jie; Jin, Peiquan (2009): Ontological Foundation for Enterprise Competitive Intelligence in the Web. In: Proceedings of the 2009 International Symposium on Information Processing (ISIP'09), pp. 445–448. Huangshan, P. R. China. ACADEMY PUBLISHER.