

**Tesina**

**Agentes Web Inteligentes para la detección de individuos preponderantes en el albedrio de sus círculos sociales en Facebook**

**Alumno:**

**David Israel García Alcázar**

**Director: Dr. Julio Cesar Ponce Gallegos**

**Asesores:**

**Dr. Alejandro Padilla Díaz**

**Dr. Juan Pedro Cardona Salas**

**Aguascalientes, Ags., Junio 2015**

Tabla de contenido

[1.1. Abstract 4](#_Toc421834085)

[1.2. Introducción 5](#_Toc421834086)

[2. Planteamiento del problema 9](#_Toc421834087)

[1. Justificación 14](#_Toc421834088)

[2.1. Planteamiento del problema 14](#_Toc421834089)

[2. Justificación 19](#_Toc421834090)

[2.2. Preguntas de Investigación 19](#_Toc421834091)

[2.3. Objetivos 19](#_Toc421834092)

[2.3.1. Objetivo General 19](#_Toc421834093)

[2.3.2. Objetivos Específicos 19](#_Toc421834094)

[2.4. Hipótesis 20](#_Toc421834095)

[3. Marco de Referencia 21](#_Toc421834096)

[3.1. Redes Sociales y sus implicaciones psicológicas 21](#_Toc421834097)

[3.2. Minería de Datos 23](#_Toc421834098)

[3.3. Minería Web 26](#_Toc421834099)

[3.4. Minería de Datos Sociales 27](#_Toc421834100)

[3.5 Entendiendo una Red Social 32](#_Toc421834101)

[3.6. Análisis de aspectos psicológicos en base a texto 36](#_Toc421834102)

[3.7. Reduciendo el espacio de búsqueda 38](#_Toc421834103)

[3.8. Métodos de clasificación 38](#_Toc421834104)

[3.9. Recolección de datos 40](#_Toc421834105)

[4. Método de Investigación utilizado 42](#_Toc421834106)

[4.1. Experimentación 43](#_Toc421834107)

[4.1.1. ¿Qué es un Framework Web? 43](#_Toc421834108)

[4.1.2. ¿Por qué Facebook? 45](#_Toc421834109)

[4.1.3. Proceso de desarrollo del Framework PythonSocialNetworkAnalyzer 46](#_Toc421834110)

[4.1.4. Arquitectura del Framework 47](#_Toc421834111)

[4.1.5. Problemática en el desarrollo del Framework y en la experimentación 59](#_Toc421834112)

[4.1.6. Arquitectura del Clúster 65](#_Toc421834113)

[4.1.7. Descripción del Clúster: 65](#_Toc421834114)

[4.1.8. Funcionamiento del Clúster: 66](#_Toc421834115)

[5. Resultados 68](#_Toc421834116)

[6. Conclusiones 74](#_Toc421834117)

[Bibliografía 75](#_Toc421834118)

# 1.1. Abstract

Desde su invención las redes sociales tales como *Facebook* han cobrado fuerza y popularidad, a tal grado que las vemos en muchos aspectos de nuestra vida, cada día consiguen una mayor cantidad de seguidores, más aplicaciones las utilizan como credenciales, se genera una mayor cantidad de información a través de estas, vemos como las personas hacen públicos varios aspectos de su vida, y generan cadenas de influencia entre amigos, al igual que sucede en los círculos sociales de nuestro entorno que no son virtuales. Toda esta información y cadenas de influencia brindan una nueva alternativa para hacer llegar a las personas información y nuevas tendencias de una forma más optimizada al saber utilizar los flujos de información que en ella se mueven, identificando a los nodos raíz a través de los cuales haciendo llegar ciertos datos, puede que los mismos se muevan a través de su red social invirtiendo una menor cantidad de recursos en su propagación, de la misma forma que en nuestro entorno se ubican lugares clave a través de los cuales se puede hacer llegar cierta información a una mayor cantidad de personas, con esto hablamos de optimizar recursos tales como el tiempo, dinero, energía y recursos computacionales. Pero para lograr esto es necesario investigar la forma en la que se generan estas cadenas. Hoy en día existen varios métodos para detectar nodos líderes mediante el uso de teoría de grafos con modelos estadísticos, sin embargo es poco el avance que se ha logrado en el campo de la minería de redes sociales utilizando minería de texto en comparación con el área matemática en la que la mayor cantidad del esfuerzo ha sido invertido, a lo largo de esta investigación se verá el proceso que se siguió para desarrollar un Framework en lenguaje Python, para realizar Minería de Redes Sociales en base a Minería de Texto. Utilizando el conteo de palabras positivas y negativas que contiene un post y sus comentarios, se comparara el grado de pertenencia a un sentido positivo o negativo de acuerdo a la mayoría registrada en el conteo sobre las palabras pertenecientes a un diccionario positivo o negativo. Este resultado será relacionado con la popularidad del post que será medida por la cantidad de likes que este contenga.

# 1.2. Introducción

Los orígenes de un enfoque a la estructura social explícitamente usando ideas de una red social son difíciles de discernir. (Scott, 2011) Es por esto que posteriormente veremos que a través de la historia el hecho de concebir la estructura social y las relaciones que hay entre los seres humanos se pueden representar a través de una red social donde los nodos representan personas o grupos sociales, mientras que las aristas representan las relaciones entre las personas o los grupos, a través de esta estructura social es que podemos representar las relaciones a través de un grafo, en el que podemos representar o no, el flujo de información al representarlo como un grafo dirigido, si no se desea representar el flujo de la información se mantiene como un grafo no dirigido. Puede llegar a ofrecer una gran utilidad el hecho de tener representado el flujo de la información ya que este flujo no siempre es bidireccional y puede llegar a ofrecer otros enfoques y cambios de perspectivas el pensar sobre este hecho. A partir de estas ideas es que pudieron surgir las redes sociales en las que los usuarios pueden ser representados como nodos, y sus vínculos son representados por aristas que unen el nodo de un usuario con un amigo, y el flujo de la información puede ser truncado para evitar que dos usuarios sigan en contacto aquí es donde se ve la utilidad de tener representado el flujo de la información, gracias a esta abstracción de la estructura social se puede llevar a una base de datos la misma, que posteriormente puede ser utilizada por una aplicación para ofrecerla a un público, tal es el caso de *Facebook.*

Un creciente número de personas se están uniendo a las redes sociales donde ellos cultivan sus amistades, comparten intereses comunes y discuten diferentes tópicos. A propósito de estas interacciones miembros de redes sociales se afectan entre sí en su comportamiento y opinión. (Carolin, Johannes, & Freimut, 2013) Esta nueva representación de la estructura social ha llegado a tener tal éxito que actualmente *Facebook*  cuenta con más de mil millones de usuarios alrededor del mundo, y cuyo negocio es la publicidad y el análisis de información de los usuarios. Cuando un usuario líder en su círculo social tiende a esparcirse de una forma más rápida ya que el hecho de que alguno de sus amigos le dé un *like* o haga un comentario hace que el mecanismo de *Facebook* le dé un mejor ranking dentro de la sección de noticias de su círculo social.

Las redes sociales proveen un buen escenario para el análisis de la conducta humana y sus relaciones, por lo que el reconocimiento de líderes, personas influyentes e influenciables es viable sobre este tipo de plataformas. Además de que las redes sociales en estos días representan un elemento clave en el esparcimiento de la información, ideas e innovaciones. Podemos llamar a un grupo de miembros influenciables objetivo, y provocar un efecto de influencia cascada por aquellos amigos que recomiendan un producto, publican una innovación o un tópico. (Kimura, Saito, Nakano, & Motoda, 2010)

Hoy en día, investigaciones han mostrado que muchas redes sociales están hechas de múltiples comunidades, que son definidas como grupos con los que las interacciones sociales entre actores son muy intensivas, pero entre ellos son muy débiles. Entonces, siendo capaces de descubrir dichas comunidades de redes es de mucha ayuda para ayudarnos a entender las redes sociales a profundidad y entonces descubrir algunos de los patrones escondidos más útiles detrás de dichas redes. (Yang, Huang, Liu, & Liu, 2009) Al ser una rama de la ciencia con poco tiempo de empezar investigaciones sobre ella, se viene a atacar la problemática sobre la mejora, nuevos enfoques y datos obtenidos por algoritmos con fines de obtener nuevos resultados y motivos sobre las relaciones personales entre los usuarios de las redes sociales, y la relación de estas con el comportamiento humano en la interacción frente a frente, y como es que esta relación y plataformas pueden llegar a afectar las relaciones personales en la vida personal y trato persona a persona.

Una comunidad es un grupo, el cual puede ser creado ya sea por Facebook o sus usuarios. Un usuario puede volverse miembro de una comunidad. (Safaei, Sahan, & Ilkan, 2009)

La minería de redes sociales es un área de investigación activa en sociología, psicología social, antropología para el siglo pasado. La minería de redes sociales tiene una larga historia en las ciencias sociales. Hoy hay una convergencia de redes sociales y tecnológicas y sistemas de información con estructuras sociales intrínsecas. (Safaei, Sahan, & Ilkan, 2009)

Los líderes de grupos sociales son definidos como aquellas personas que inician demasiadas cadenas de influencia de acuerdo al número de cadenas a las que pertenecen. En las ciencias sociales, la influencia se define como el poder de negociación, control sobre la información o persuasión. Burt describe que la influencia social ocurre cuando las emociones de alguien, opiniones o comportamientos son afectados por otros, y la influencia viene de ser un puente sobre un hoyo en la estructura. (Tsai, Tzeng, Lin, & Chen, 2014)

Las redes sociales permiten el esparcimiento de varios tipos de información, incluyendo tópicos, ideas e incluso virus de computadora. La proliferación de e-mails, blogs y servicios de redes sociales en la World Wide Web acelera la creación de grandes redes sociales. Por lo que recientemente la atención hacia la forma en la que esta información se esparce, es decir el flujo de la información en las redes sociales, y otros aspectos tales como los análisis de los sitios de redes sociales, evolución de tópicos y problemas de privacidad. Encontrar los nodos influenciables es uno de los problemas centrales en el análisis de las redes sociales. Entonces, el desarrollo eficientes y prácticos métodos de hacer esto, basado en la difusión de la información es un importante tema de investigación. Los modelos probabilísticos básicos de difusión de la información, tales como el modelo de cascada independiente y el modelo de umbral lineal. Los investigadores profundizan en el problema de encontrar un número límite de nodos influenciables que sean efectivos para el esparcimiento de la información en los modelos anteriores. (Saito, Kimura, Ohara, & Motoda, Efficient discovery of influential nodes for SIS models in social networks, 2012)

Un usuario es una persona, la cual se suscribe a Facebook, hace amistad con otros usuarios, se convierte en un miembro de una comunidad. (Safaei, Sahan, & Ilkan, 2009)

Los usuarios pueden crear álbumes para sus fotos, en estas fotos otros amigos pueden ser etiquetados. (Safaei, Sahan, & Ilkan, 2009)

Se ha estudiado la propagación del material multimedia en las redes de persona a persona, es decir por interacción real del usuario o interacción frente a frente, los resultados de algunas de estas investigaciones arrojan que aunque las fotos incluso sean populares, se esparcen lentamente y el intercambio de información entre amigos cuenta para aproximadamente el 50% de su opinión al clasificar y provoca un retraso considerable en los saltos de información. (Fazeen, Dantu, & Guturu, 2011)

Las investigaciones pasadas acerca del tema muestran que el uso de algoritmos greedy pueden dar una buena solución aproximada al problema de identificación de nodos líderes en las redes sociales, sin embargo los algoritmos greedy enfrentan un problema computacional, sobre él uso de recursos y el tiempo de cómputo requerido. (Kimura, Saito, Nakano, & Motoda, 2010)

En investigaciones anteriores autores han puesto su atención en obtener las diferencias entre las distintas redes sociales, por lo que en base a estos estudios podemos tomar la información para el desarrollo de estructuras, algoritmos y aplicaciones efectivas. (Fazeen, Dantu, & Guturu, 2011) Facebook cuenta con elementos que lo vuelven una plataforma de interacción social que muestra un entorno muy similar al del mundo real, que permite realizar una investigación sobre el comportamiento de las personas de una forma más similar a la que se realiza de persona a persona, ya que incluye objetos tales como: publicaciones que se pueden comparar con las críticas entre personas, *likes* que expresan gustos sobre los flujos de información que se generan incluyendo comentarios, fotos y publicaciones, ofrece la posibilidad de compartir elementos generados por otros usuarios que se pueden relacionar con los saltos de información directos entre personas, sentimientos agregados en los comentarios en conjunto con *smileys* o *emoticones* que ofrecen una interacción parecida a la de los gestos y el lenguaje corporal que afectan la forma en que la información puede ser interpretada y chats que simulan las conversaciones entre personas cara a cara o en un aspecto más general inclusive entre medios electrónicos o no electrónicos.

El uso de la inteligencia artificial en el apoyo de la investigación sobre las redes sociales ya se ha realizado con anterioridad, Cheong y Lee identificaron mensajes sobre las últimas tendencias del momento o eventos especiales en una red social usando técnicas de visualización y minería de datos basada en métodos inteligencia artificial. (Fazeen, Dantu, & Guturu, 2011)

El problema de identificación de actores o usuarios de redes sociales es actualmente un problema inherente asociado con los métodos de reconocimiento de patrones. El problema que podemos llegar a enfrentarnos en este ámbito es que mientras más grande sea la dimensión del vector de patrones, con lo que veríamos que las muestras de patrones para el entrenamiento estaría bastante disperso en un espacio dimensional mayor, por lo que se volvería insuficiente para el entrenamiento preciso de los clasificadores. Por lo que, al ser mayor la dimensionalidad del vector de patrones se requiere de una disponibilidad mucho mayor de muestras. (Fazeen, Dantu, & Guturu, 2011)

Para poder extraer información de gente anónima, necesitamos considerar la confiabilidad de los datos y desentendernos de un propósito. En adición, el trasfondo de las habilidades de los usuarios lectores, estados y condiciones mentales podrían ser de ayuda para obtener información. (Soo Cho, Yoel Yoon, Joon Kim, Yeon Lim, Kwan Kim, & Kim, 2011)

# 2. Planteamiento del problema

*Enriquecer la diversidad en algoritmos para analizar las redes sociales, en este caso sobre el análisis del comportamiento humano específicamente en el campo de la influencia de líderes sobre otras personas. Ya que un creciente número de personas se están uniendo a las redes sociales donde ellos cultivan sus amistades, comparten intereses comunes y discuten diferentes tópicos, un ejemplo del impacto de las redes sociales puede ser visto a través del número de usuarios que tienen estas, un ejemplo puede ser facebook (ver Imagen 1). A propósito de estas interacciones miembros de redes sociales se afectan entre sí en su comportamiento y opinión.*(Carolin, Johannes, & Freimut, 2013)

******

*Imagen 1: Usuarios de Facebook activos al més durante el segundo cuarto del 2013, fuente: (Facebook, Inc., Million of Monthly Active Users, 2013)*

Muchas veces se hace un descubrimiento o se otorga una herramienta al mundo, la cual es aceptada y utilizada por una gran cantidad de personas alrededor del planeta sin embargo, muy raras veces se es consciente de todos los posibles usos que se le pueden llegar a dar a tal herramienta o descubrimiento, ya que muchos surgen conforme la gente va dándoles cierto uso. Además aunque se sea consciente de la cantidad de propósitos o posibles usos de tales herramientas o descubrimientos por falta de recursos o tiempo, no se puede adentrar demasiado en el tema.

Actualmente el lenguaje de programación Python carece herramienta alguna que permita el análisis de la Red Social Facebook. Este lenguaje empieza a ser utilizado ampliamente en el desarrollo y análisis Web dadas sus características para generar scripts, su simpleza y rapidez al momento de codificar, lo que brinda una amplia ventaja sobre algunos lenguajes similares en el área, por lo que esta investigación brindará un Framework Llamado **PythonSocialNetworkAnalyzer** que permitirá realizar análisis social mediante una arquitectura de procesos distribuidos, utilizando la Minería Web, la Minería de Texto y la Minería de Datos, este Framework se encuentra disponible en la siguiente URL: *https://github.com/idcodeoverflow/SocialNetworkAnalyzer.*

Entender como los usuarios se influencian unos a otros, puede tener varios beneficios. Por ejemplo, el marketing viral, recomendaciones, dispersión de la información, entre otros. (Liu, Tang, Han, & Yang, 2012) Un claro ejemplo de lo anterior se puede observar en la (Imagen 2). Además del hecho de conocer características sobre los perfiles que pueden llegar a ser posibles líderes, brinda una gran cantidad de oportunidades a los negocios tanto para mejorar sus ventas, como para ampliar y mejorar sus estrategias de mercadeo y tomar decisiones más acertadas sobre los posibles productos que podrían entrar al mercado. Siendo esto para el apoyo a la toma de decisiones de máxima importancia y el análisis de riesgo.



*Imagen 2: Fuentes de Marketing más efectivas durante el 2014, fuente: (Content Marketing Institute, marketingcharts.com, December 2014).*

Con el poder de la influencia, una compañía puede mercadear un nuevo producto empezando por convencer un pequeño número de usuarios influenciables para adoptar el producto y preparar futuras adopciones a través del efecto del “el mundo de la boca” que se conoce como maximización de influencias. Por ejemplo, en las redes académicas, gracias a la influencia entre los investigadores colaboradores, las ideas novel o innovaciones, se esparcen rápidamente y llevan a nuevas ideas y nuevos direcciones académicas. En las redes sociales como Facebook y Twitter, los usuarios son frecuentes a seguir amigos influyentes en su círculo social para retwettear un metroblog o dar like una foto o comentario. (Liu, Tang, Han, & Yang, 2012)

Asimismo ayudará a resolver en base a la información obtenida sobre los perfiles y sus relaciones en cuestión de lazos, puede ayudar a tomar mejores decisiones sobre que personas podrían servir como actores para publicidad de ciertos productos, de acuerdo al público al que vayan dirigidos. (Zhang, Wang, & Xia, 2010)

Otra problemática que puede venir a apoyar para su resolución es la mejora de los algoritmos para sugerencia de amigos en las redes sociales, ya que en base a tu perfil y el de otro usuario de acuerdo a la estadística y análisis arrojado por los algoritmos se podría llegar a predecir la relación entre dos usuarios, y de esta forma sugerir que comienzan una amistad sobre la plataforma y así ampliar las redes de flujo de información. (Saito, Kimura, Ohara, & Motoda, Efficient discovery of influential nodes for SIS models in social networks, 2012)

Un conflicto común con él que se está lidiando al hacer este tipo de investigaciones es el hecho de que la complejidad del espacio de búsqueda y análisis al igual que la magnitud de este representan un reto para el uso de recursos computacionales (Al-Saggaf & Islam, 2014), por lo que un mejor rendimiento sobre los algoritmo utilizados puede representar un ahorro tanto en energía, tiempo y recursos a las compañías encargadas de realizar este tipo de análisis. (Livingston, Triquenaux, Fighiera, Beyler, & Jalby, 2014)

De una forma increíble se puede llegar a mejorar el ambiente, las políticas ambientales y recursos mercadológicos. En el primer caso de mejorar al ambiente me refiero a que se invertiría una menor cantidad de energía para realizar ciertos cálculos lo que se puede traducir en una menor cantidad de emisiones de contaminantes al ambiente por concepto de producción de energía, al igual de que al mejorar el rendimiento sobre los equipos ya existente se requeriría una menor cantidad de los mismos para realizar dichas investigaciones lo que implica una menor manufactura y uso de recursos naturales, al igual que menos energía empleada para poder producir los equipos de cómputo necesarios para realizar dichos cómputos. (Livingston, Triquenaux, Fighiera, Beyler, & Jalby, 2014)

Los efectos directos de la diversidad están basados en evidencia, “… que la existencia de la diversidad, por sí misma, puede afectar ciertos procesos organizacionales, creatividad, y resolución de problemas, que están directamente relacionados al rendimiento”. Cox (Cox, Cultural Diversity in Organizations, 1993), Copeland (Copeland, 1988), Mandell y Kohler Gray (Mandell & Kohler-Gray, 1990) y Lobel and McLeod (Cox, Lobel, & McLeod, Effects of ethnic group cultural difference on cooperative versus competitive behaviour in a group task, 1991) defienden que cuando la diversidad es manejada apropiadamente, diversas organizaciones y grupos presentan ventajas de desempeño, sobre elementos homogéneos. Esta línea de pensamiento es conocida como “Valor-en-diversidad” en la filosofía. (Diversity and Managing Diversity: Theory and State of the Art, 2008) Basados en este principio es que puedo decir que al enriquecer la colección de algoritmos para el análisis social en la plataforma de *Facebook* se hace una aportación al rendimiento, ya que al realizar una correcta administración sobre los mismos al momento de analizar y llevar a cabo estudios se pueden obtener mejores resultados que al contar únicamente con un elemento que pudiera no ofrecer las características requeridas o deseadas para efectuar el estudio, de esto es que parto a decir que: “la alimentación o inclusión de nuevos algoritmos para incrementar la diversidad de opciones al momento de requerir estudiar un tópico es una problemática constante que debe ser atendida por los investigadores”, que viene a ser atendida por esta investigación.

Al ser una rama de la ciencia con poco tiempo de empezar investigaciones sobre ella, se viene a atacar la problemática sobre la mejora, nuevos enfoques y datos obtenidos por algoritmos con fines de obtener nuevos resultados y motivos sobre las relaciones personales entre los usuarios de las redes sociales, y la relación de estas con el comportamiento humano en la interacción frente a frente, y como es que esta relación y plataformas pueden llegar a afectar las relaciones personales en la vida personal y trato persona a persona. (Raad, Chbeir, & Dipanda, 2013)

Los círculos sociales están en constante evolución de acuerdo a lo investigado y los seres humanos también enfrentan cambios a lo largo de su vida de acuerdo a la generación a la que pertenezcan (Bentley, Ormerod, & Batty, 2011), esto nos dice que las empresas y los negocios también deben de evolucionar, aquellas empresas que logran visualizar dichos cambios antes que los demás logran tener una ventaja competitiva sobre otras, por lo que generalmente logran tener mejores ingresos al tener mejores conocimientos sobre la manera de relacionarse de los usuarios y por ende también sobre ciertos aspectos de sus consumidores, esta es la forma en que se puede llegar a predecir el comportamiento de las personas e incluso las relaciones futuras con otros usuarios. Esto representa una herramienta muy poderosa para las entidades económicas. (Lusch & Laczniak, 1987) Yendo más allá representaría una ventaja importante también a partidos políticos, ya que estos necesitan información sobre la población para poder dirigir sus campañas y objetivos para el cargo de sus representantes, además de que representan una oportunidad importante para las universidades que requieren llegar a ofrecer mejores alternativas de estudio a sus estudiantes para mejorar las técnicas de enseñanza y el proceso de aprendizaje e interacción de sus estudiantes con la universidad, además de un adecuado flujo de la información es indispensable en toda organización ya sea lucrativa o sin fines de lucro. (Peters & Araya, 2007)

# Justificación

# 2.1. Planteamiento del problema

*Enriquecer la diversidad en algoritmos para analizar las redes sociales, en este caso sobre el análisis del comportamiento humano específicamente en el campo de la influencia de líderes sobre otras personas. Ya que un creciente número de personas se están uniendo a las redes sociales donde ellos cultivan sus amistades, comparten intereses comunes y discuten diferentes tópicos, un ejemplo del impacto de las redes sociales puede ser visto a través del número de usuarios que tienen estas, un ejemplo puede ser Facebook (ver Imagen 1). A propósito de estas interacciones miembros de redes sociales se afectan entre sí en su comportamiento y opinión.*(Carolin, Johannes, & Freimut, 2013)

******

*Imagen 1: Usuarios de Facebook activos al més durante el segundo cuarto del 2013, fuente: (Facebook, Inc., Million of Monthly Active Users, 2013)*

Muchas veces se hace un descubrimiento o se otorga una herramienta al mundo, la cual es aceptada y utilizada por una gran cantidad de personas alrededor del planeta sin embargo, muy raras veces se es consciente de todos los posibles usos que se le pueden llegar a dar a tal herramienta o descubrimiento, ya que muchos surgen conforme la gente va dándoles cierto uso. Además aunque se sea consciente de la cantidad de propósitos o posibles usos de tales herramientas o descubrimientos por falta de recursos o tiempo, no se puede adentrar demasiado en el tema.

Actualmente el lenguaje de programación Python carece herramienta alguna que permita el análisis de la Red Social Facebook. Este lenguaje empieza a ser utilizado ampliamente en el desarrollo y análisis Web dadas sus características para generar scripts, su simpleza y rapidez al momento de codificar, lo que brinda una amplia ventaja sobre algunos lenguajes similares en el área, por lo que esta investigación brindará un Framework Llamado **PythonSocialNetworkAnalyzer** que permitirá realizar análisis social mediante una arquitectura de procesos distribuidos, utilizando la Minería Web, la Minería de Texto y la Minería de Datos, este Framework se encuentra disponible en la siguiente URL: *https://github.com/idcodeoverflow/SocialNetworkAnalyzer.*

Entender como los usuarios se influencian unos a otros, puede tener varios beneficios. Por ejemplo, el marketing viral, recomendaciones, dispersión de la información, entre otros. (Liu, Tang, Han, & Yang, 2012) Un claro ejemplo de lo anterior se puede observar en la (Imagen 2). Además del hecho de conocer características sobre los perfiles que pueden llegar a ser posibles líderes, brinda una gran cantidad de oportunidades a los negocios tanto para mejorar sus ventas, como para ampliar y mejorar sus estrategias de mercadeo y tomar decisiones más acertadas sobre los posibles productos que podrían entrar al mercado. Siendo esto para el apoyo a la toma de decisiones de máxima importancia y el análisis de riesgo.



*Imagen 2: Fuentes de Marketing más efectivas durante el 2014, fuente: (Content Marketing Institute, marketingcharts.com, December 2014).*

Con el poder de la influencia, una compañía puede mercadear un nuevo producto empezando por convencer un pequeño número de usuarios influenciables para adoptar el producto y preparar futuras adopciones a través del efecto del “el mundo de la boca” que se conoce como maximización de influencias. Por ejemplo, en las redes académicas, gracias a la influencia entre los investigadores colaboradores, las ideas novel o innovaciones, se esparcen rápidamente y llevan a nuevas ideas y nuevos direcciones académicas. En las redes sociales como Facebook y Twitter, los usuarios son frecuentes a seguir amigos influyentes en su círculo social para retwettear un metroblog o dar like una foto o comentario. (Liu, Tang, Han, & Yang, 2012)

Asimismo ayudará a resolver en base a la información obtenida sobre los perfiles y sus relaciones en cuestión de lazos, puede ayudar a tomar mejores decisiones sobre que personas podrían servir como actores para publicidad de ciertos productos, de acuerdo al público al que vayan dirigidos. (Zhang, Wang, & Xia, 2010)

Otra problemática que puede venir a apoyar para su resolución es la mejora de los algoritmos para sugerencia de amigos en las redes sociales, ya que en base a tu perfil y el de otro usuario de acuerdo a la estadística y análisis arrojado por los algoritmos se podría llegar a predecir la relación entre dos usuarios, y de esta forma sugerir que comienzan una amistad sobre la plataforma y así ampliar las redes de flujo de información. (Saito, Kimura, Ohara, & Motoda, Efficient discovery of influential nodes for SIS models in social networks, 2012)

Un conflicto común con él que se está lidiando al hacer este tipo de investigaciones es el hecho de que la complejidad del espacio de búsqueda y análisis al igual que la magnitud de este representan un reto para el uso de recursos computacionales (Al-Saggaf & Islam, 2014), por lo que un mejor rendimiento sobre los algoritmo utilizados puede representar un ahorro tanto en energía, tiempo y recursos a las compañías encargadas de realizar este tipo de análisis. (Livingston, Triquenaux, Fighiera, Beyler, & Jalby, 2014)

De una forma increíble se puede llegar a mejorar el ambiente, las políticas ambientales y recursos mercadológicos. En el primer caso de mejorar al ambiente me refiero a que se invertiría una menor cantidad de energía para realizar ciertos cálculos lo que se puede traducir en una menor cantidad de emisiones de contaminantes al ambiente por concepto de producción de energía, al igual de que al mejorar el rendimiento sobre los equipos ya existente se requeriría una menor cantidad de los mismos para realizar dichas investigaciones lo que implica una menor manufactura y uso de recursos naturales, al igual que menos energía empleada para poder producir los equipos de cómputo necesarios para realizar dichos cómputos. (Livingston, Triquenaux, Fighiera, Beyler, & Jalby, 2014)

Los efectos directos de la diversidad están basados en evidencia, “… que la existencia de la diversidad, por sí misma, puede afectar ciertos procesos organizacionales, creatividad, y resolución de problemas, que están directamente relacionados al rendimiento”. Cox (Cox, Cultural Diversity in Organizations, 1993), Copeland (Copeland, 1988), Mandell y Kohler Gray (Mandell & Kohler-Gray, 1990) y Lobel and McLeod (Cox, Lobel, & McLeod, Effects of ethnic group cultural difference on cooperative versus competitive behaviour in a group task, 1991) defienden que cuando la diversidad es manejada apropiadamente, diversas organizaciones y grupos presentan ventajas de desempeño, sobre elementos homogéneos. Esta línea de pensamiento es conocida como “Valor-en-diversidad” en la filosofía. (Diversity and Managing Diversity: Theory and State of the Art, 2008) Basados en este principio es que puedo decir que al enriquecer la colección de algoritmos para el análisis social en la plataforma de *Facebook* se hace una aportación al rendimiento, ya que al realizar una correcta administración sobre los mismos al momento de analizar y llevar a cabo estudios se pueden obtener mejores resultados que al contar únicamente con un elemento que pudiera no ofrecer las características requeridas o deseadas para efectuar el estudio, de esto es que parto a decir que: “la alimentación o inclusión de nuevos algoritmos para incrementar la diversidad de opciones al momento de requerir estudiar un tópico es una problemática constante que debe ser atendida por los investigadores”, que viene a ser atendida por esta investigación.

Al ser una rama de la ciencia con poco tiempo de empezar investigaciones sobre ella, se viene a atacar la problemática sobre la mejora, nuevos enfoques y datos obtenidos por algoritmos con fines de obtener nuevos resultados y motivos sobre las relaciones personales entre los usuarios de las redes sociales, y la relación de estas con el comportamiento humano en la interacción frente a frente, y como es que esta relación y plataformas pueden llegar a afectar las relaciones personales en la vida personal y trato persona a persona. (Raad, Chbeir, & Dipanda, 2013)

Los círculos sociales están en constante evolución de acuerdo a lo investigado y los seres humanos también enfrentan cambios a lo largo de su vida de acuerdo a la generación a la que pertenezcan (Bentley, Ormerod, & Batty, 2011), esto nos dice que las empresas y los negocios también deben de evolucionar, aquellas empresas que logran visualizar dichos cambios antes que los demás logran tener una ventaja competitiva sobre otras, por lo que generalmente logran tener mejores ingresos al tener mejores conocimientos sobre la manera de relacionarse de los usuarios y por ende también sobre ciertos aspectos de sus consumidores, esta es la forma en que se puede llegar a predecir el comportamiento de las personas e incluso las relaciones futuras con otros usuarios. Esto representa una herramienta muy poderosa para las entidades económicas. (Lusch & Laczniak, 1987) Yendo más allá representaría una ventaja importante también a partidos políticos, ya que estos necesitan información sobre la población para poder dirigir sus campañas y objetivos para el cargo de sus representantes, además de que representan una oportunidad importante para las universidades que requieren llegar a ofrecer mejores alternativas de estudio a sus estudiantes para mejorar las técnicas de enseñanza y el proceso de aprendizaje e interacción de sus estudiantes con la universidad, además de un adecuado flujo de la información es indispensable en toda organización ya sea lucrativa o sin fines de lucro. (Peters & Araya, 2007)

# 2. Justificación

# 2.2. Preguntas de Investigación

* ¿Qué variables debemos tener en cuenta al momento de generar un algoritmo para analizar las publicaciones en *Facebook*?
* ¿Cómo podemos medir la influencia de un usuario en *Facebook*?
* ¿Cómo desarrollar un Framework en Python que nos permita realizar minería Web para medir la influencia de un usuario en *Facebook*?

# 2.3. Objetivos

# 2.3.1. Objetivo General

Distinguir los nodos líderes que causan influencia sobre el albedrío de su círculo social en *Facebook* mediante el análisis de sus publicaciones y los comentarios hechos en estas, desarrollando un Framework que permite el Análisis Social en *Facebook* para el lenguaje Python.

# 2.3.2. Objetivos Específicos

1. Desarrollar un Framework que permita el análisis de la red Social Facebook para el lenguaje de programación Python que hasta el día de hoy no hay una herramienta que permita esto.
   1. El Framework permitirá Minería Web.
   2. El Framework permitirá Minería de Texto.
   3. El Framework permitirá Minería de Datos.
   4. Ofrecerá la posibilidad de instalarse en una arquitectura distribuida, para mejorar el rendimiento.
2. Obtener las publicaciones de usuarios de *Facebook* para su análisis.
   1. Es necesario obtener las publicaciones de usuarios de *Facebook* a partir de una semilla dada, ya que a partir de esto se obtendrá la información para su posterior pre-procesamiento y análisis en busca de patrones.
3. Realizar un pre-procesamiento adecuado de la información que nos permita limpiar los datos.
   1. Para una adecuada obtención de conocimiento es necesario eliminar ruido e inconsistencias en los datos seleccionados obtenidos, es una preparación para su análisis posterior. Me permitirá mejorar la precisión y certeza del conocimiento obtenido.
4. Construir un algoritmo para distinguir los nodos líderes, mediante el análisis de sus publicaciones en *Facebook* usando minería de texto.
   1. Se requiere de un algoritmo que nos permita efectuar interpretación y evaluación de los datos en busca de patrones, para obtener conocimiento a partir de estos.
5. Obtener un diccionario de palabras negativas, positivas, al igual que uno en el que se encuentren las más comunes entre las publicaciones populares (Es necesario incluir el periodo de tiempo).
   1. Estos diccionarios serán parte del proceso requerido para encontrar patrones dentro de los datos transformados a partir de la información seleccionada con el fin de encontrar a los usuarios líderes en círculos sociales.

# 2.4. Hipótesis

A través del conteo de palabras contenidas en diccionarios de palabras previamente etiquetadas como positivas o negativas de un post o comentario que estén contenidas en un texto es posible determinar si este es positivo o negativo, a través de la cantidad de likes un comentario o post puede ser medido en su grado de influencia, al relacionar la pertenencia de un texto a un sentido positivo o negativo en conjunto con su grado de influencia se puede obtener la tendencia sobre que probabilidad hay que un texto positivo obtenga una mayor aceptación en relación a la aceptación que puede ser esperada por un texto clasificado como negativo. (Chou, Sinha, & Zhao, 2008)

Al llevar una estadística sobre la cantidad de cadenas de información generadas por un usuario de Facebook, y la cantidad de cadenas que han sido catalogadas como populares en base a los likes, se puede determinar que usuarios tienen una mayor aceptación o influencia sobre sus círculos sociales y el tipo de influencia que estos ejercen ya sea positiva o negativa en base a la cantidad dominante de publicaciones marcadas como populares y la pertenencia de cada una de estas publicaciones a un sentido positivo o negativo. Identificando así a líderes positivos y negativos dentro de sus círculos sociales. (Al-Saggaf & Islam, 2014) A través de un Framework implementado en lenguaje Python y una base de datos implementada en MySQL.

# 3. Marco de Referencia

3.1. Redes Sociales y sus implicaciones psicológicas

Los orígenes de un enfoque a la estructura social explícitamente usando ideas de “red social”, son difíciles de discernir. Estructuralmente pensando tiene raíces profundas en la tradición sociológica, pero fue hasta 1930’s que específicamente hablando se pensó como una red dando un enfoque distinto a la estructura social. Los investigadores alemanes influenciados por George Simmel pusieron su énfasis en las propiedades formales de la interacción social para construir una “sociología formal” en las que los sociologistas se unieron para investigar las configuraciones de las relaciones sociales producidas a través de la entrevista de encuentros sociales. Intentando romper el consenso de lo que se había asumido por las principales corrientes de la Sociología Americana y reconocer conflictos y divisiones dentro de la estructura de la comunidad, ellos vieron un análisis de redes que proveía los principios para llegar al final de estas corrientes. Fue hasta 1954 que se propuso la idea de tomar seriamente la del concepto de una red de relaciones. Alfred Vierkant y Leopoldo Von Wiese fueron los proponentes claves de esta idea y explícitamente adoptaron una terminología de puntos, líneas, y conexiones para describir las relaciones sociales. Sus ideas influenciaron a un número de trabajadores en la psicología social y la psicoterapia quienes se interesaron en las maneras en las que las pequeñas estructuras grupales influencian a los individuos en sus percepciones, acciones y elecciones. (Scott, 2011)

La Sociometría es un método cuantitativo para medir las relaciones sociales. Esto introdujo la idea que las estructuras sociales son representadas como diagramas de redes. La sociometría se convirtió en el campo de investigación en la educación y en la psicología social, donde tuvo su punto más alto en el llamado enfoque de las “dinámicas de grupo”. (Scott, 2011) Una de las ramas de la ciencia que serán de apoyo para la realización de esta investigación ya que proveerán el criterio para poder generar un algoritmo adecuado en la clasificación de los posts y de los comentarios, que posteriormente permitirá clasificar a un usuario como influyente o no de acuerdo a la cantidad de posts que haya generado y se hayan clasificado como influyentes. Será un apoyo desde el punto de vista que nos permitirá conocer los fundamentos mediante los cuales una persona puede ser aceptada dentro de su grupo y el grado de aceptación de la misma, que vienen a formar un bloque crucial dentro de la investigación. Ya que esto representa el ajuste de los parámetros y las variables a tomar por la parte técnica. Visto desde otro punto de vista viene a representar el cerebro dentro del cuerpo del algoritmo, ya que en base a este conocimiento, podremos tener las métricas que evalúan a cada uno de los usuarios dentro de los grupos anteriormente descritos. Esta investigación viene a reunir varias entidades que habían estado separadas dentro de un algoritmo que podrá ser capaz de clasificar a los usuarios en base al trabajo de equipo de varias ramas de la ciencia.

Entre los objetivos de la sociometría podemos encontrar:

1. Conocer el nivel de aceptación que una persona tiene en su grupo.
2. Evaluar el grado de cohesión entre personas de un grupo.
3. Localizar a los individuos más rechazados y más valorados (líderes potenciales).
4. Localizar a los sujetos aislados, que no despiertan ni admiración ni rechazo.
5. Comprobar las consecuencias de la incorporación de nuevas personas al grupo.
6. Verificar el grado de aceptación e incorporación de personas a un nuevo lugar de trabajo.

Las dinámicas de grupo buscan explicar los cambios internos que se producen como resultado de las fuerzas y condiciones que influyen en los grupos como un todo y de cómo reaccionan los integrantes. Dentro de los grupos se han identificado 4 prototipos clásicos, que, según su rol son:

* Portavoz: es el responsable dentro de una organización para dirigirse a los medios de comunicación. Puede ser el presidente/consejero delegado, el director de relaciones públicas u otro alto directivo designado por la organización para tal fin; el puesto puede recaer también en alguien diferente especialmente designado. Es habitual que los portavoces reciban formación específica para mejorar su desempeño frente a los periodistas.
* Chivo expiatorio: es la denominación que se le da a una persona o grupo de ellas a quienes se quiere hacer culpables de algo de lo que no son, sirviendo así de excusa a los fines del inculpador.
* Líder: es esa persona comprometida en asumir una posición de poder debido a un compromiso y convicción dentro de un ambiente de equipo. Lo que diferencia a un líder con los demás, es su carisma y espíritu de lucha incesante con el fin de lograr un bien común o meta en la vida.
* Saboteador: es aquel que realiza una modificación, destrucción, obstrucción o cualquier intervención en una operación ajena, con el propósito de obtener algún beneficio para uno mismo.

Estos roles y dinámicas de grupos las podemos ver reflejados en las redes sociales, en este caso en *Facebook*. (Scott, 2011) Dentro de los círculos sociales humanos, podemos reconocer una jerarquía y una estructura, individuos que desean entrar al círculo social, otros que quieren salir y los que desean permanecer en él. Los elementos del círculo social tienden a compartir actividades, características, objetivos, gustos, opiniones y tendencias. Regularmente su posición en la jerarquía dentro del círculo social se ve determinada por su capacidad de relacionarse con los otros, personalidad y apego a las características que definen su círculo social.

3.2. Minería de Datos

Mucho del trabajo de investigación y análisis de las redes sociales se ha enfocado en el estudio matemático utilizando la teoría de grafos por lo que el enfoque que se le da a esta investigación es un campo relativamente nuevo, y su potencial empieza a ser reconocido a través de las ciencias sociales y los avances en el Análisis y Minería de Datos en las Redes Sociales es una marca de su potencial en el análisis en el campo de las redes sociales. Las nuevas técnicas de análisis de redes son más apropiadas para grandes data-sets del tipo que generalmente no son factibles analizar utilizando software comercial o común público. La Minería de Datos permite que dicha información sea analizada utilizando técnicas que prometen proveer conocimiento sustancial del área. (Scott, 2011) Al ser el caso de que la mayoría del trabajo de investigación sobre las redes sociales haberse enfocado sobre el estudio de teoría de grafos, nos deja un campo fértil para la investigación con bases en otros campos que pueden ser de apoyo, ya que algunos conceptos, técnicas, características u otras variables que se tomen en cuenta en otros campos pueden ser adoptados dentro de este, lo que ofrece una posibilidad de realizar una investigación más ágil y con más fundamentos.

La minería de datos es el proceso de exploración y análisis, por medios automáticos o semi-automáticos, de grandes cantidades de datos con la intención de descubrir patrones útiles. En otras palabras, la minería de datos es el proceso completo de revelar patrones útiles y relaciones en los datos usando técnicas como la inteligencia artificial, machine learning y estadísticas a través de herramientas de análisis de datos avanzadas. (Bozkır, Güzin Mazman, & Akçapınar Sezer, 2010) A partir de esta premisa es que obtengo la idea de involucrar agentes inteligentes para el proceso del análisis de la información a través de la minería de textos. Se contará con dos tipos de agentes uno que se encargara de ser vigía y buscar a aquellos posibles usuarios que puedan representar un nodo líder, mientras que otros tomarán los resultados obtenidos por estos agentes que estarán almacenados en la base de datos que servirá como un búfer para control de entradas y salidas, y analizarán sus publicaciones utilizando minería de texto, para decidir si una publicación puede ser catalogada como influyente o no influyente a través de la cantidad de palabras que tenga de acuerdo a diccionarios, que tendrán pre-clasificadas palabras como positivas o negativas, todo esto posteriormente de realizar el pre-procesamiento de los datos para permitir que estos sean tratados de una forma más sencilla y que se presten a ofrecer resultados más certeros. Sin dicho pre-procesamiento de los datos se puede convertir en una tarea imposible la trata y la búsqueda en forma de encontrar información útil, ya que un espacio de búsqueda demasiado amplio puede ocasionar problemas en los algoritmos al momento de buscar en los diccionarios de datos, con las palabras pre-clasificadas. También se es necesario el llevar a algunas variantes de palabras a una forma básica para reducir la dimensionalidad del espacio que puede llegar a ser un problema como se verá más adelante, para este paso dentro del pre-procesamiento también será necesario el eliminar algunas de las palabras que no se consideran como relevantes para el estudio de la publicación y que pueden categorizarse como ruido, y de esta forma reducir tiempo de ejecución al igual que el uso de recursos e inclusive también mejorar la calidad de la precisión en la salida obtenida.

Los métodos de minería de datos son clasificados en dos categorías: predictiva y descriptiva. El propósito de los métodos de la minería de datos predictiva es hacer predicciones en casos no vistos utilizando casos vistos en modelos entrenados. El propósito de las técnicas descriptivas de la minería de datos es descubrir relaciones profundas, correlaciones y propiedades descriptivas de los datos. (Bozkır, Güzin Mazman, & Akçapınar Sezer, 2010) Para propósitos de esta investigación utilizare técnicas descriptivas de la minería de datos, ya que el objetivo es encontrar relaciones entre los usuarios, para encontrar a aquellos nodos que pueden calificarse como líderes, a través del análisis de las palabras que se encuentran en sus publicaciones y en los comentarios hechos sobre las mismas publicaciones. Se evitarán las técnicas de predicción de la minería de datos ya que no pretendo predecir el comportamiento de los usuarios en base a la información pasada, sino únicamente encontrar aquellos vínculos que permitan calificarlos como líderes en sus círculos sociales.

El método de árbol de decisión es probablemente el más utilizado para la clasificación, entre las distintas técnicas de clasificación de la minería de datos, debido a la fácil interpretación visual que puede ser representada, así mismo su uso e implementación son relativamente sencillas. Típicamente, una tarea de minería de datos para un árbol de decisión es la clasificación; por ejemplo, identificar el riesgo de un usuario de tarjetas de crédito. La idea principal de un árbol de decisión es dividir la información de manera recursiva en subconjuntos de tal forma que cada subconjunto cubra de la mejor forma posible los estados de una variable dependiente. En cada división en el árbol, todas las variables independientes son recalculadas para ver su impacto en la variable dependiente. Cuando este proceso recursivo es detenido, y el árbol de decisión se encuentra en un estado estable, se dice que el árbol de decisión se encuentra en un estado formado. En este caso nuevos casos pueden ser clasificados vía el árbol de decisión formado. Este punto también es llamado el árbol de deducción. (Bozkır, Güzin Mazman, & Akçapınar Sezer, 2010) Se utilizará esta técnica de clasificación una vez que se tengan clasificadas las publicaciones y relacionadas con sus respectivos dueños o publicadores, para poder realizar la clasificación de dichos usuarios, opto por elegir esta técnica ya que si tomamos en cuenta la cantidad de usuarios que se pueden llegar a analizar y esta ya representa una importante cantidad con la que se puede llegar a tener problemas al lidiar con el tiempo y la complejidad para su procesamiento, el hecho de pensar que cada uno de estos usuarios tendrá varias publicaciones o al menos una, el problema se vuelve aún más complejo, dada esta situación y teniendo en cuenta que este tipo de clasificador utiliza el enfoque de divide y vencerás, tenemos que es un elemento de alta importancia para lidiar con la complejidad del problema y el espacio de búsqueda tan grande que llega a representar. Se ha visto en las ciencias de la computación que este tipo de enfoque de divide y vencerás se utiliza en casos en los cuales el espacio de búsqueda es demasiado amplio y una técnica de tratamiento de la información por fuerza bruta representa tiempos de pre-procesamiento bastante largos, y siendo este el caso en el que buscamos optimizar el tratamiento de este problema de clasificación de usuarios, es que nos percatamos que este enfoque se adapta perfectamente a la intención de esta investigación y que en problemas que cuentan con características parecidas a este han ofrecido buenos resultados, inmediatamente se opta por escoger esta clase de esquema para el tratamiento de la información en busca de una clasificación adecuada y que ofrezca buenos tiempos de respuesta, tomando en cuenta que una vez armado el árbol clasificador se tiene una caché de procesamiento importante que reduce aun más el tiempo de procesamiento.



Imagen 3: Proceso de obtención de conocimiento a través de la minería de datos.

3.3. Minería Web

La minería de datos web puede ser dividida en 3 categorías principales: minería de la estructura web, minería de contenido web y minería de uso web. La minería de estructura web es la técnica de extracción de links y estructura de sitios web con el fin de descubrir relaciones previas desconocidas entre las páginas web. La minería de contenido de web, algunas veces referenciada como minería de texto, es el proceso de escaneo y extracción de contenido de una página web, por ejemplo, texto, imágenes, gráficas, audio, video e hipervínculos, para determinar la relevancia del contenido de un query de búsqueda. La minería de contenido web es particularmente útil en el análisis de redes sociales debido al ambiente rico en contenido de los sitios de redes sociales. La minería de uso web es una técnica usada para rastrear y analizar el comportamiento de navegación de los usuarios. (Al-Saggaf & Islam, 2014) Para propósitos de esta investigación se utilizara la minería de texto, ya que como se describe anteriormente sus características y propósitos se relacionan ampliamente con el fin de este estudio, ya que no buscamos categorizar a *Facebook* dentro de un conjunto de páginas web de acuerdo a su estructura, ni tampoco deseamos categorizar la página dentro un conjunto de aplicaciones o sitios web de acuerdo al uso que los usuarios le dan, sino que se tiene la intención de catalogar a los usuarios de acuerdo a una clasificación dado su comportamiento sobre este sitio web de *Facebook* que representa una red social ampliamente utilizada alrededor del mundo y que además simula de una forma muy parecida a la realidad la interacción entre las personas, la forma en la que se puede realizar esta clasificación de los usuarios es sobre su comportamiento, el cual viene descrito en las publicaciones y comentarios que se realizan sobre esta red social, lo que nos pone a pensar, que la minería de texto es la categoría que debe ser usada para llevar a cabo el proceso de categorización de los usuarios, dado que se puede utilizar para investigar el contenido expuesto en cada uno de los comentarios y publicaciones hechos por los usuarios, y llevar a cabo el pre-procesamiento de los datos, el análisis, la generación de patrones y la obtención del conocimiento a partir de estos.

3.4. Minería de Datos Sociales

El análisis de redes sociales podemos dividirlo en dos áreas: el análisis cualitativo y el análisis cuantitativo. Los últimos trabajos de Berkowitz y Scott sobre análisis cualitativo de redes sociales se enfocan en patrones de relaciones entre las personas, estados, organizaciones, etc. En este campo también podemos encontrar el concepto de homofilia o la esclavitud entre tipos de personas similares, tales como matrimonio, amistad, sociedades, trabajo, intercambio de información, etc. y limita el mundo social de las personas con respecto al intercambio de información, interacciones humanas, y formación de actitudes. Es importante en este campo tomar en cuenta varias dimensiones socio-demográficas de la homofilia como la raza, etnia, sexo, religión, educación, etc. y algunos otros factores como la educación, prestigio ocupacional, y clases sociales están fuertemente en el mismo nivel homofílico que la religión y el sexo. (Fazeen, Dantu, & Guturu, 2011) Este trabajo se basará en el análisis de datos cualitativo ya que mas de referirnos a números o a obtener estadísticas de uso como resultado, se tiene la finalidad de categorizar perfiles en base a sus publicaciones, que estas a su vez serán clasificadas en base a su contenido, en pocas palabras la investigación consta de pasar un proceso subjetivo tal como la categorización en base a contenido, a un computador para que pueda realizar esta clasificación acorde a un conjunto de algoritmos, y técnicas de procesamiento, pre-procesamiento y trata de la información en busca de conocimiento con un buen grado de aceptabilidad y precisión que permita generar premisas verídicas sobre la clasificación de un perfil de usuario. Esta clasificación puede llegar a obtenerse únicamente a través del análisis adecuado de la información, que ya ha sido pasada por una etapa de limpieza de la misma, y que antes de esta etapa para mejorar el tiempo de respuesta un agente inteligente ha seleccionado los perfiles sobre los cuales se va a trabajar y empezar a monitorear su comportamiento sobre la red social de *Facebook* y de esta forma poder ofrecer un juicio adecuado sobre si es un líder, un seguidor o un spammer.

El modelado de bloques para análisis cuantitativo de las redes sociales que inductivamente descubre la subyacente estructura de roles tales como la similitud de posiciones en una organización, sociedad de una comunidad o asociación, etc. (Fazeen, Dantu, & Guturu, 2011) El análisis cuantitativo únicamente se verá reflejado en las estadísticas de interacción y comportamiento entre usuarios, de tal forma que se pueda llevar a cabo un conteo de *likes* y calcular un rango de tiempo sobre las publicaciones que han de ser analizadas ya que un factor importante que se debe delimitar es el tiempo máximo en que una publicación ha de ser comentado, ya que el hecho de analizar posts de los usuarios de esta red social, no debe hacerse de una forma ilimitada en cuanto tiempo ha de hablarse ya que se describe que el tiempo es un factor clave, pero si se piensa en que esta será una razón un elemento de división o representará un factor de multiplicación puede disparar los resultados o hacer que estos tiendan a cero afectando la precisión de los resultados. Sin embargo no se debe de pensar en dejar como un elemento a excluir el tiempo ya que en base a este se puede medir, la velocidad en que un usuario esparce sus opiniones a través de la web.

Imagen 3.1: Gráfica de tendencia de cambio hacia una constante que califica el impacto de interacción de los usuarios con un post. Se observa que la interacción de un post tiende a llegar a un punto en el que sus cuantificadores cualitativos como los likes permanecen en un estado constante o de cambio mínimo en su defecto. El eje de las ‘Y’ representa la cantidad de likes, mientras el eje de las ‘X’ representa el tiempo transcurrido en minutos.

Imagen 3.2: Gráfica de relación interacción/tiempo obtenida de dividir la cantidad de likes sobre el tiempo transcurrido en minutos. Se observa que la relación de cambio comienza a tender a 0, esto debido al aumento de tiempo y la tendencia a una constante descrita en la imagen 3.1, y el cambio incesante en el tiempo, el punto máximo describe la influencia del post. El eje de las ‘Y’ representa la cantidad de likes, mientras el eje de las ‘X’ representa el tiempo transcurrido en minutos.

El comportamiento visto en las gráficas anteriores se da debido a que el tiempo es un factor en constante aumento mientras que como se ve en la gráfica de la imagen 3.1, el aumento en la cantidad de likes tiende a quedarse fijo dado un punto en el que la influencia de ese post llega a un estado inerte, siendo que vemos la relación entre el tiempo y la cantidad de likes que en este caso describen la influencia a la cual el post se hace acreedor, al tener como factor divisor el tiempo vemos que el comportamiento describe una curva gaussiana dado por un inicio en 0 likes y al aumentar el tiempo el acercamiento al 0 en el eje de las X.

En algunos estudios el tiempo es visto como un factor, y se asume que los usuarios solo son influenciables de dos diferentes formas: influenciados dentro de un cierto periodo de tiempo y des influenciados fuera del periodo de tiempo. (Tsai, Tzeng, Lin, & Chen, 2014) El tiempo en este estudio es visto como un factor de acuerdo a varios motivos, primero en investigaciones anteriores como la cita anterior lo refiere mientras más rápido interactúan los usuarios con un flujo de información proveniente de un nodo de su red, mayor puede considerarse la influencia y la amplitud de esta, teniendo en cuenta la cantidad de amigos que este tenga en su lista, ya que no es lo mismo 20 interacciones sobre la publicación de un usuario que tiene 1200 amigos, a 20 interacciones sobre la publicación de otro usuario que tiene 100 amigos, en base a esto es que se debe hacer este análisis del tiempo en base a porcentaje, y el tiempo al afectar la cantidad de interacciones de manera que mientras más corto sea el lapso de tiempo en el que esta alcanza un porcentaje considerable de interacción y mientras más tarda en pasar esto se ve afectada la calificación del post del usuario de forma negativa, es que encontramos que el tiempo es un factor de relación en base a la interactividad, por lo que se describiría de la siguiente manera:

Fórmula 1. Relación de cambio de la cantidad de likes de un texto en *Facebook en relación al tiempo.*

Siempre y cuando el divisor se encuentre dentro del rango aceptado de tiempo de vida donde la diferencia debe ser igual al valor del parámetro “*P*” que representa el delta de tiempo aceptado para el análisis de dicha publicación.

Muchos textos de los usuarios contienen importantes y variadas informaciones. Podremos obtener esta información si la semántica puede ser extraída de los textos en una red social. (Soo Cho, Yoel Yoon, Joon Kim, Yeon Lim, Kwan Kim, & Kim, 2011) He aquí que la idea se ve respalda por el hecho de que la información puede ser extraída siempre y cuando la semántica lo permita y como se menciona anteriormente, investigaciones preliminares arrojan resultados de que el análisis de texto se puede realizar a través de palabras simples de una forma efectiva, ya que las palabras expresan en sí el contenido y el orden viene dado en su origen por la sintaxis, y para cuestiones de análisis de texto basta la comparación de estas palabras contenidas, todo esto posteriormente hecho un pre-procesamiento de la información, para así tener reducida la complejidad, descartar elementos que no sean necesarios y asimismo ofrecer un mejor resultado, utilizando las técnicas de para este paso descritas en párrafos anteriores, este representa un paso muy importante ya que puede depender de este que se puede realizar con éxito la clasificación de un comentario y a su vez de una publicación en base a los mismos y a las palabras contenidas en el cuerpo de la publicación. Esta técnica ha sido utilizada anteriormente para catalogar si el contenido de las páginas web se podía clasificar como abuso de Internet, es decir que en base a las palabras que se contenían dentro del cuerpo de la página web, se podía clasificar el contenido en base a temas pre-clasificados utilizando diccionarios o bolsas de datos analizando cada palabra o token, de forma separada y en base a la cantidad y relación cantidad inversa, se añadían a clases o grupos predefinidos de temas. (Chou, Sinha, & Zhao, 2008)

Se debe lidiar también con el hecho de la complejidad del algoritmo y el descartar cadenas de influencia irrelevantes. (Tsai, Tzeng, Lin, & Chen, 2014) La complejidad del algoritmo siempre debe ser tomada en cuenta ya que el hecho de tener una idea sobre cómo resolver un problema no garantiza que el algoritmo llegue a converger al menos en un tiempo razonable o inclusive factible, ya que si no se analiza la estructura y funciones del mismo en relación a una notación tal como la “función O” no se puede conocer cuánto tiempo consumirá en encontrar una solución que da a reflejar sus desventajas sobre todo en entradas grandes o con aquellas que en base a la configuración de los parámetros del algoritmo pueden ocasionar un mayor uso de recursos computacionales y permita que el algoritmo se cicle, generando un tiempo de ejecución infinito si no se detiene la corrida del mismo. Por ejemplo, si tenemos un algoritmo de complejidad O(n3) y tenemos una entrada que genera un valor de “n=10,000” tendremos que serán necesarios al menos 1,000,000,000,000 ciclos del algoritmo para poder converger, suponiendo que nuestro ordenador puede realizar un millones de operaciones de nuestro algoritmo por segundo, tendremos que tardaría 11 días 13 horas 46 minutos y 40 segundos, esto siendo todavía una entrada pequeña, por esto es que hay soluciones infactibles para un problema dada su complejidad computacional.

Para poder analizar el efecto de la propagación de las cadenas de información debemos elegir un método de propagación, encontramos 2 posibles SIR (Susceptible, Infectado, Recuperado) donde un elemento susceptible puede ser infectado pero un recuperado no, esto nos dice que un elemento no puede ser infectado más de una vez, en contraste SIS (Susceptible, Infectado, Susceptible) que permite que un elemento pueda ser infectado más de una vez. En esta ocasión se utilizará SIS. Este método ha sido utilizado en otras investigaciones tal como en (Saito, Kimura, Ohara, & Motoda, Efficient discovery of influential nodes for SIS models in social networks, 2012). Utilizaré Susceptible Infectado Susceptible, ya que este método de propagación define mejor la interacción entre los usuarios de las redes sociales en la mayoría de los casos, en el que un usuario puede verse influenciado por la publicación de otro más de una vez, teniendo en mente esto, podemos pensar en la premisa que la influencia de *Facebook* es proporcional a la cantidad de usuarios que tenga, ya que mientras más usuarios utilicen este servicio una mayor cantidad de cadenas de la información se moverán a través de su red, a su vez también llegando a una mayor cantidad de usuarios tomando en cuenta factores como la cantidad de amigos del publicador del post, y la privacidad del mismo, ya que puede expandirse solo a la lista de amigos de este en caso de ser privado, llegar a un cierto porcentaje de sus amigos en caso de tener privacidad personalizada o ser pública y llegar a cuanta persona que se vea interesada en recibir dicha cadena de información desee ver pueda acceder a esta inclusive si no posee una cuenta en la red social. Por otra parte este método de infección ha sido utilizado en investigaciones sobre dicho tema anteriormente. (Saito, Kimura, Ohara, & Motoda, Efficient discovery of influential nodes for SIS models in social networks, 2012) En resumen este tipo de infección nos permite analizar a un usuario que ha interactuado con otro más de una vez y no vernos en la necesidad de recurrir a una Lista Tabú, para restringir el análisis entre usuarios, únicamente este recurso será utilizado para poder excluir de análisis a ciertos usuarios que han sido catalogados como no líderes durante un cierto periodo de tiempo, o hasta que haya un aumento en su lista de amigos considerable, sea el primero en ocurrir de estos casos.

3.5 Entendiendo una Red Social

Es importante tomar en cuenta el cambio de la red social y el círculo social con respecto al tiempo ya que la edad es un factor clave dentro de la categorización y conducta humana. Algunos autores presentan modelos de redes sociales con estructuras sociales convincentes, y explican cómo una búsqueda rápida de entidades objetivo en las redes sociales puede ser alcanzada porque la partición jerárquica de las redes sociales debido a las relaciones de las redes y las identidades de los individuos forman la red social. Las identidades por definición, son conjuntos de características atribuidas a los individuos, por si mismos y por otros en virtud de su asociación en las redes sociales. (Fazeen, Dantu, & Guturu, 2011) Los círculos sociales están en constante evolución de acuerdo a lo investigado y los seres humanos también enfrentan cambios a lo largo de su vida y también de acuerdo a la generación a la que pertenezcan, esto nos dice que las empresas y los negocios también deben de evolucionar, y aquellas empresas que logran visualizar dichos cambios antes que los demás logran tener una ventaja competitiva sobre otras empresas, por lo que generalmente logran tener mejores ingresos al tener mejores conocimientos sobre la manera de relacionarse de los usuarios y por ende de ciertos aspectos de las personas, es que se pueden llegar a predecir el comportamiento de las personas e incluso las relaciones futuras con otros usuarios. Esto representa una herramienta muy poderosa para las entidades económicas. Yendo más allá representaría una ventaja importante también a partidos políticos, ya que estos necesitan información sobre la población para poder dirigir sus campañas y objetivos para el cargo de sus representantes, además de que representan una oportunidad importante para las universidades que requieren llegar a ofrecer mejores alternativas de estudio a sus estudiantes para mejorar las técnicas de enseñanza y el proceso de aprendizaje e interacción de sus estudiantes con la universidad, además de un adecuado flujo de la información es indispensable en toda organización ya sea lucrativa o sin fines de lucro.

Dentro de las redes sociales podemos distinguir dos tipos de entidades aparte de las propuestas, organizaciones e individuos, esto en la temática del tipo de persona. Sin embargo también se es posible clasificarlos de acuerdo a su conducta dentro de la red, o por el fin principal al que ingresan a dicha red, en esta clasificación podemos encontrar los siguientes grupos líderes, mirones (conocidos como stalkers), spammers y allegados. (Fazeen, Dantu, & Guturu, 2011) Considerando que las redes sociales representan una nueva forma de relacionarse entre las personas, también podremos encontrar que los individuos de la lista de amigos del dueño de un perfil de *Facebook* y twitter representan parte de su círculo o círculos sociales, son miembros de un círculo social en el que este desea entrar, son conocidos o elementos de un entorno familiar, personas que se desconocen pero atraen la atención del propietario del perfil o individuos solitarios que se agregan a su entorno virtual para simular el entorno real en que se desarrolla. En cualquiera de estos dos últimos casos podemos ver motivada esta acción por el deseo de mostrar que la lista de amigos del propietario tiene mayor cantidad de individuos, y de esta forma caer en el mismo proceso de buscar ser elegible a una mayor cantidad de grupos de interés, o incluso demostrar supremacía dentro de su grupo social. Un elemento importante que determina el valor de la lista de amigos de un perfil no solo es la cantidad sino la calificación social de los miembros de esta, y así de esta forma para cada perfil de *Facebook*. (Al-Saggaf & Islam, 2014) En la jerarquía del círculo social encontramos elementos muy influyentes en los demás cuya información se puede llegar a esparcir muy rápidamente, y en contraste elementos cuya influencia es poca y que la información que comparten tiende a pasar casi desapercibida. También podemos encontrar elementos que desean entrar a un círculo social y que buscarán demostrar que se identifican con las características comunes del mismo, y tratarán de crear vínculos importantes con los líderes o elementos influyentes del mismo, haciendo notar su aceptación hacia la información que los mismos comparten. Dentro de este grupo podemos encontrar individuos que serán aceptados y otros que no lo serán en cualquier caso durante el periodo en el que entrar al círculo social sea un objetivo de estos, buscarán lograr la aceptación de sus miembros y captar su atención mediante acciones. Asimismo podemos distinguir otros individuos solitarios que admiran ciertos grupos sociales y que se mantendrán en un constante proceso de demostrar su admiración pero sin atreverse a intentar formar parte del grupo social. En contraste podemos encontrar individuos solitarios que por resentimiento social buscarán ignorar todo flujo de información de ciertos grupos sociales, en estos encontraremos a los que realmente los desaprueban y los que buscan refugiarse fingiendo el desagrado a estos. En este último caso al sentir atención por parte de algún elemento con influencia conocida en el grupo en cuestión, provocara el individuo solitario el inicio de un proceso como el descrito anteriormente para llamar la atención al pensar formar parte del grupo. En ciertos individuos solitarios encontraremos en su flujo de información que no es aceptado e incluso sus tendencias llegan a ser rechazadas y tachadas por gran cantidad de grupos sociales, e incluso rechazo entre individuos solitarios.

La identificación de usuarios como spammers (este término es acuñado por la práctica frecuente publicación de mensajes o contenido considerado como spam), se realiza básicamente con la detección de la repetición temporal de contenidos y links persistentes en base a un patrón. Esta información representa ruido o basura para la investigación de contenidos. (Fazeen, Dantu, & Guturu, 2011) Los usuarios identificados como spammers deben excluirse del análisis ya que el propósito de esta investigación no es medir su influencia en su círculo social, y para nuestros fines su contenido se puede clasificar como no admisible, por lo que en la etapa de selección de la información para nuestro dataset a analizar se habrán de excluir y así aportar una ganancia de tiempo para el análisis y detección de posibles líderes. Un usuario de la red social puede clasificarse como spammer, si el contenido que publica es persistente y similar. Por ejemplo un usuario que venda autos u ofrezca sus servicios de consultoría en Tecnologías de la Información tenderá a publicar el mismo post repetidas veces, esta es una de las posibles técnicas que pueden utilizarse para la detección de spammers dentro de las redes sociales.

Para este estudio se clasificará a los usuarios (individuales) de *Facebook* en 2 grupos:

* Líderes: aquellos por los que empiezan las grandes cadenas o flujos de información en base a la cantidad de personas en su círculo social, y la cantidad de círculos sociales en las que dan a notar un liderazgo en relación a la cantidad de círculos sociales en los que se ven inmersos.
* Pasivos: permanecen inactivos y pocas ocasiones interactúan con los demás.

Dentro de este último grupo reclasificaremos a los que son seguidores e influenciables de ciertos líderes en relación a su perfil.

Una pregunta interesante es: cómo los amigos en una red social se influencian unos a otros y como la influencia se esparce sobre la red social? La respuesta a esto es no trivial. En vez de esto, es retadora en los siguientes aspectos: Primero, Cuáles son los mecanismos fundamentales micro-nivel de influencia social en la redes sociales? En particular, cuando las redes sociales son heterogéneas (consistentes de objetos heterogéneos, como usuarios, grupos y blogs), ¿cómo la influencia es afectada por diferentes tipos de objetos en diferentes tópicos? Recientemente, los usuarios de la web disfrutan compartir o esparcir contenido de usuario generado que sea interesante. Por lo que además de la estructura de la red, el esparcimiento del contenido en el tope de la red se convierte en un factor clave para la minería de la influencia social en redes heterogéneas. Mientras, que los hobbies pueden ser principalmente influenciados por los miembros de sus familia o amigos cercanos en su vida diaria. Por lo que la fuerza de la influencia puede variar también con los tópicos. El problema de juntar el conocimiento distribuido de tópicos asociados con otros usuarios y la influencia a nivel tópico entre los usuarios no ha sido direccionada antes. (Liu, Tang, Han, & Yang, 2012) Un elemento importante a considerar dentro de la investigación y por el motor del algoritmo de clasificación de las publicaciones de los usuarios, es la temática o el tópico que se trata, ya que como se menciona anteriormente no es lo mismo hablara sobre una graduación o la obtención de un título universitario en un publicación, a otra en la que se publique que se está comiendo en puesto de comida rápida sin un motivo de celebración aparente. En base a esto es que propongo la creación de un diccionario de datos en el que no solo se incluyan las palabras principales de las publicaciones marcadas como influyentes o importantes, sino que también se agregue la fecha, para descartar top-trends o modas, de una característica repetible general dentro de las mismas publicaciones influyentes.

En segundo lugar, ¿pueden los amigos de tus amigos tener algún tipo de influencia en tu comportamiento?, Interesantemente, la respuesta es sí. Por ejemplo, se ha estudiado con anterioridad la influencia de la felicidad, y se encontró que la felicidad se esparce a tres grados de separación. (Liu, Tang, Han, & Yang, 2012) Esto no solo nos dice que los amigos de nuestros amigos pueden tener influencia sobre nosotros en nuestras publicaciones, asimismo nos dice que la influencia de una persona o usuario de una red social, en este caso *Facebook*¸ puede transmitir su influencia, y afectar el comportamiento, sentimientos y/o pensamientos de aquellos que se ven vulnerables hacia este, y que lo reciben en su flujo de información. Esto respalda y habilita el supuesto de una publicación puede transmitir influencia de un usuario hacia otros, y que no necesariamente debe encontrarse en su lista de amigos, un posible tema de investigación posterior puede ser el análisis de grados de profundidad en una red social.

3.6. Análisis de aspectos psicológicos en base a texto

Hay dos métodos para analizar la psicología de las personas a través de un texto, el primero es el análisis del contenido temático, y el segundo es la cuenta de palabras estratégicas. (Soo Cho, Yoel Yoon, Joon Kim, Yeon Lim, Kwan Kim, & Kim, 2011) Para propósitos de esta investigación utilizaremos la cuenta de palabras estratégicas para el análisis de la influencia de un usuario, a través de la coincidencia de palabras dentro de diccionarios que las clasificarán. Técnicas similares se han utilizado en estudios anteriores, ya que desde la década de 1950 se ha tenido la intención de investigar la psicología humana a través de textos. En investigaciones pasadas se había encontrado que personas que utilizan pronombres en primera persona frecuentemente, bastantes palabras que representan emociones negativas, y menos verbos que una persona normal, tienen una alta probabilidad de padecer depresión. Además es más probable que una persona cometa suicidio si expresa frecuentemente palabras relacionadas con la muerte. También que una persona que utiliza palabras de causa y visión de una forma habitual es más saludable que otras de manera apreciable. Además, usan palabras en tercera y primera persona menos comúnmente para expresar exclusiones, emociones negativas y palabras de acción cuando dicen una mentira. Por otra parte, al analizar los textos se puede determinar si el escritor es hombre o mujer, si es joven o viejo. (Soo Cho, Yoel Yoon, Joon Kim, Yeon Lim, Kwan Kim, & Kim, 2011) Estas investigaciones habilitan el supuesto de que una publicación puede ser analizada de acuerdo a sus palabras para clasificarse como popular o no popular, y a su vez calificar como influyente o no influyente a un usuario, ya que a través de las palabras se puede obtener información sobre su escritor, y estas mismas reflejan tendencias de las personalidades, y si se toma en cuenta el hecho de que el ser humano es un ser social y tiende a adoptar prácticas o tendencias que ve a su alrededor, se ve completado el círculo de que las palabras escritas transmiten emociones, tendencias, pensamientos, y por lo tanto pueden transmitir influencia, y así influenciar a otros miembros de la red social.

El objetivo de la indexación es identificar un conjunto de atributos que son informativos sobre la categoría de un comentario, específicamente hablando categorizarlo como comentario popular, comentario no popular o spam, la indexación nos permitirá tener un mejor desempeño en la clasificación. Investigaciones sobre el flujo de palabras sugieren que funcionan bien como unidades de representación y el orden de estas unidades en un documento no tiene mayor importancia. Incluso investigaciones pasadas utilizando minería de texto descubrieron que el uso de pares de palabras da resultados más pobres que el uso de palabras simples. (Chou, Sinha, & Zhao, 2008) Básicamente esto describe el proceso por el cual tomaré las palabras clave, después de realizar la limpieza de elementos poco descriptivos del texto, y la eliminación de sufijos para dejar únicamente las palabras en su expresión “pura”, y buscarlas en los diccionarios de datos. Es en este momento que se describe que es factible y funcional, brindando buenos resultados el analizar un texto en base a sus palabras individuales, y que de hecho trae mejores resultados que analizar pares de palabras y demás, ya que las palabras mismas son unidades que representan un sentido, y cada una de estas unidades en conjunto dan un sentido al texto, es decir mientras mayor cantidad de palabras que expresen cierto sentido en una publicación se encuentra, mas estará orientada esta publicación hacia dicho sentido.

Esta técnica de pre-categorización del texto es conocida como categorización de texto o clasificación de textos. Es la actividad de etiquetar texto del lenguaje natural en categorías temáticas de un conjunto predefinido. Esto implica un proceso de conocimiento supervisado, donde un modelo de predicción llamado clasificador, es automáticamente inducido por un conjunto de ejemplos. Después de la preparación de un conjunto de ejemplos, un paso crítico está determinado por definir una buena forma de representar la clasificación de texto en un archivo. (Chou, Sinha, & Zhao, 2008) Este paso de pre-procesamiento requiere de dos procesos mayores: indexación y reducción de la dimensionalidad. En este caso la representación de la información dentro de un archivo estará dada por un diccionario de datos.

A través de esquemas en los que se les da un peso a los términos es la forma en la que se utilizará para trabajar. La frecuencia de términos y la frecuencia inversa de términos se ha encontrado como efectiva en la categorización de textos y es de las más ampliamente usadas. La frecuencia de términos pesará una palabra de un comentario de *Facebook* por el número de veces que aparecen palabras de su misma categoría dentro del comentario o publicación que se está analizando. Mientras más términos de una misma categoría aparezcan más representativa es dicha categoría para la publicación o comentario. La frecuencia inversa de términos ajusta la frecuencia en que aparece un término de una misma clase en un comentario o publicación por la cantidad de veces que dicha clase es utilizada durante todo el conjunto de comentarios en la publicación que se está analizando. (Chou, Sinha, & Zhao, 2008) Lo mencionado en el párrafo anterior se sustenta en esta investigación en la que se dice que efectivamente la categorización de textos se puede realizar a través de la frecuencia en la que los términos del texto aparecen, y que la frecuencia de estos conlleva a la publicación misma hacia el sentido que estos términos más frecuentes representan. Teniendo así que una clasificación a través de tokens previamente tratados y representativos del texto, permiten la clasificación del texto mismo, y de la misma forma, cierta cantidad de publicaciones clasificadas de tal o cual forma, permiten clasificar a un usuario de una red social, en este caso de *Facebook* como influyente o no influyente, dada la tendencia en que sus flujos de información son asimilados, rechazados o admitidos por otros usuarios de la red social, y el tiempo en que estos eventos se pueden llegar a producir.

3.7. Reduciendo el espacio de búsqueda

La reducción de la dimensionalidad llega a ser casi siempre necesaria dado el gran número de términos que se pueden llegar a encontrar en el conjunto de publicaciones y comentarios a analizar. Ya que conjuntos demasiado grandes de palabras únicas pueden llevar a un clasificador a una sobrecarga, esto en el sentido de perder precisión en los resultados arrojados. La reducción de la dimensionalidad puede ser llevada a cabo de dos formas. La primera es que la dimensionalidad se reduzca deshaciéndonos de las palabras de paro y sufijos. (Chou, Sinha, & Zhao, 2008) Las palabras de paro son unidades frecuentes, que no proveen gran información como: un, una, y, el, al igual que algunas preposiciones como dentro, en y sobre, conjunciones como: pero, aunque y porque. También es necesario unificar términos como conectado, conectarse, conectando, conectó, todos estos pertenecen al mismo término “conectar” solo que se les agregan sufijos, es necesario remover estos sufijos para unificar estos términos. El algoritmo “Porter Stemming” es uno de los más utilizados para remover sufijos. La segunda forma de reducir la dimensionalidad es seleccionar un conjunto de atributos significativos de acuerdo a la característica de seleccionabilidad, hay mucha información sobre funciones de seleccionabilidad. (Chou, Sinha, & Zhao, 2008) Para uso de esta investigación se optará por el primer método, ya que se adapta perfectamente a la reducción de términos para su localización dentro de diccionarios de datos previamente preclasificados, y así poder realizar una clasificación de las publicaciones en base al sentido de sus palabras. Este paso mejorará el tiempo de respuesta de los algoritmos y mejorará la precisión de las salidas, ya que será más probable que una palabra se encuentre dentro de un diccionario de datos previa la eliminación de elementos que diversifican su estructura pero no su significado original, o hacia el cual van orientados o dirigidos.

Utilizaré cuatro métodos de clasificación que ya se han utilizado en estudios anteriores para la categorización de textos. (Chou, Sinha, & Zhao, 2008)

3.8. Métodos de clasificación

Naive Bayes: es una técnica de modelos probabilísticos para la clasificación que se basa en la regla de Bayes de la probabilidad condicional que asume que los atributos son condicionalmente independientes, dada la clase, para simplificar la estimación de probabilidades condicionales. El naive Bayes multinomial es un es una forma modificada del naive Bayes que es más apropiada para la clasificación de textos, ya que no solo acomoda las palabras por su aparición sino que también toma en cuenta la frecuencia de aparición de las palabras. (Chou, Sinha, & Zhao, 2008) Esta técnica se adapta a las características del problema y del algoritmo, ya que los tokens representan variables independientes, mientras que el texto o la publicación representan una variable dependiente, y el algoritmo asume que los atributos son independientes y estima como es que se relacionan con las variables dependientes. Por lo que puede presentar un buen resultado en la clasificación de los individuos, que pueden considerarse como los elementos dependientes, mientras que los textos pertenecientes a este individuo son variables independientes que tienen una clasificación por sí mismos.

K-nearest neighbors es una técnica basada en una técnica de aprendizaje. Compara una instancia nueva con un conjunto de instancias almacenadas con salidas que ya son conocidas, y asigna la nueva instancia a la clase que represente una mayoría de acuerdo a elementos cercanos. (Chou, Sinha, & Zhao, 2008) Está técnica de clasificación también se adapta al problema y a la forma que se plantea de tratarlo, en la que se puede ubicar a un perfil de usuario de la red social, de acuerdo a la cantidad de posts que tenga clasificados de cierta forma, e ir iterando hasta que converja en una de las clases que se tienen especificadas, y así clasificar al individuo.

Árbol de decisión: es una técnica que toma un enfoque de divide y vencerás, genera un árbol de modelos secuenciales. Diferentes algoritmos pueden diferir principalmente en la forma en que los atributos se seleccionan en nodos intermedios del árbol. Por ejemplo, el árbol C4.5 utilizado en esta investigación usa la información obtenida y genera una proporción de ganancia como criterio para selección de atributos. (Chou, Sinha, & Zhao, 2008) Esta técnica puede ofrecer un gran desempeño y ventaja en cuanto a tiempo, debido al enfoque que da para tratar el problema de clasificación de los individuos, utilizando el principio de divide y vencerás que tienen como base las soluciones de programación dinámica, que se utilizan para tratar problemas que pueden ser divididos en sub-problemas, y ofrecer una respuesta en una complejidad O(nLog(n)) o O(Log(n)). Lo que puede ser un elemento clave si tratamos de mejorar nuestros tiempos de respuesta. Además del hecho de que posteriormente de que el árbol se ha estabilizado y los cambios en sus variables son mínimos se obtiene lo que se conoce como árbol clasificador, que representa una caché que puede clasificar a los elementos entrantes de una forma más rápida, mejorando aun así mas nuestros resultados en cuanto a tiempo, ofreciendo una precisión aceptable.

3.9. Recolección de datos

Para la minería de datos se emplean varias técnicas para extraer información de los sitios de redes sociales incluyendo el uso de Web Crawlers, que es la técnica más común, ataques de Phishing, aplicaciones de terceros, creando perfiles falsos en la red social a ser analizada. El uso de Web Crawlers implica crear scripts para Crawling comúnmente conocidos como arañas, para recolectar a través de los perfiles públicos de *Facebook.* (Al-Saggaf & Islam, 2014)Para la recolección de datos de esta investigación se emplearán Web Crawlers que se desplazarán a través de los perfiles, seleccionando a posibles usuarios que puedan ser líderes e ingresándolos a una base de datos que servirá como búfer para el control de entradas y salidas, mientras que otros agentes se encargarán de monitorear la actividad de estos perfiles seleccionados con los que formaremos nuestro dataset, y que posteriormente en base a su análisis realizado clasificarán las publicaciones de estos usuarios como influyentes o no influyentes en base a los parámetros, variables y criterios anteriormente descritos. Que posteriormente en base a esta información se realizará una clasificación en base a los 4 algoritmos clasificadores tratados en el párrafo anterior para determinar si un usuario pertenece al grupo de líderes, spammers o no pertenecen a ninguno de estos dos. De ser este último el resultado se marcará al usuario y se agregará a una lista tabú para evitar su futura reevaluación, hasta que se cumpla cierto periodo de tiempo o se generé un cambio considerable en relación a su lista de amigos, específicamente hablando un aumento, lo que podría denotar una posible clasificación como elemento influyente, como se trato en párrafos anteriores.

La información generada por las técnicas de minería de datos elude la normativa de protección de la privacidad de la información personal. Esto es porque los datos recolectados tienden a ser, primeramente públicamente accesible (no confidencial) en naturaleza, y después, la nueva información generada consiste de análisis de patrones y relaciones que son meramente inferidos o implicados desde las vastas cantidades de instancias individuales de información personal dentro de grandes datasets. Como resultado, el método de data mining no parece violar directamente la información personal. (Al-Saggaf & Islam, 2014)

La protección de la información personal de los usuarios de sitios de redes sociales no está explícitamente declarada dentro de la mayoría de las redes de privacidad. La mayoría de las leyes de privacidad de las agencias de gobierno y otras instituciones discuten la recolección y uso de los datos personales en general sin mencionar la minería de datos en los sitios de redes sociales. (Al-Saggaf & Islam, 2014)

Dado que las redes sociales son usados por cientos de millones de personas, es seguro que estas influencian profundamente la vida de las personas, la mayoría del tiempo sin su conocimiento de las implicaciones. (Agrawal, Budak, El Abbadi, Georgiou, & Yan, 2014)

# 4. Método de Investigación utilizado

1. Montar arquitectura del clúster.
2. Creación de la base de datos sobre MySQL.
3. Desarrollo de los algoritmos para análisis de la información y corrección de errores.
4. Integración de los algoritmos desarrollados con el Framework desarrollado.
5. Montar los algoritmos y la base de datos en el clúster.
6. Recopilación de la información desde Facebook a la base de datos.
7. Realizar un pre-procesamiento de la información en la BD.
8. Ejecución de los algoritmos desarrollados en conjunto con la base de datos.
9. Si la calidad de los algoritmos es el deseado, y el funcionamiento del Framework es continuo sin interrupciones por bugs, se continuará al paso 9 sino se regresará al paso 3.
10. Análisis de los resultados obtenidos.
11. Presentar conclusiones.

En el esquema de la base de datos se almacenará la información obtenida de las publicaciones de los usuarios de *Facebook*, donde para cada usuario identificado como posible nodo líder además de ser almacenado el vínculo a dicho perfil se guardará la información sobre la actividad que tienen los demás interactuando en la información publicada en el mismo (comentarios sobre las publicaciones).

Los algoritmos se encargarán de alimentar la base de datos con la información recaudada en *Facebook* sobre sus publicaciones y su actividad con los nodos que estos algoritmos consideran líderes.

El análisis de los resultados nos permitirá respaldar o rechazar la hipótesis que se plantea, mientras el desarrollo del experimento de la forma expuesta nos permitirá llegar al objetivo del experimento.

# 4.1. Experimentación

Antes de partir al proceso de experimentación es necesario comprender que es un Framework.

4.1.1. ¿Qué es un Framework Web?

El concepto Framework se emplea en muchos ámbitos del desarrollo de sistemas software, no solo en el ámbito de aplicaciones Web. Podemos encontrar Frameworks para el desarrollo de aplicaciones médicas, de visión por computador, para el desarrollo de juegos, y para cualquier ámbito que pueda ocurrírsenos.

En general, con el término Framework, nos estamos refiriendo a una estructura software compuesta de componentes personalizables e intercambiables para el desarrollo de una aplicación. En otras palabras, un Framework se puede considerar como una aplicación genérica incompleta y configurable a la que podemos añadirle las últimas piezas para construir una aplicación concreta. (Gutiérrez., 2015)



Imagen 4.0: Representación visual abstracta de un Framework, dado que es una pieza de software genérica que es completada, por el desarrollador quien lo implementa en su software. (Gutiérrez., 2015)

Para analizar la influencia de una publicación de acuerdo a Text Mining de acuerdo a sus comentarios se efectuará el siguiente proceso:

1. Generar el nuestro dataset guardando la información sobre usuarios, posts y comentarios.
2. Generar un diccionario de datos con palabras negativas y positivas, este proceso representa la preclasificación de las palabras en clases que posteriormente se evaluarán para clasificar un texto o comentario de acuerdo al sentido mayormente expuesto y al que estas palabras mayormente coinciden, es decir a mayor cantidad de palabras positivas expuestas se puede considerar como una publicación o comentario positivo, en contraste a una mayor cantidad de palabras negativas expuestas en una publicación o comentario se puede considerar como negativo. (Minqing & Bing, 2004)
3. Se tokenizará el texto de la publicación y los comentarios. Este proceso nos permitirá iniciar un proceso de previo al análisis, que representa la separación de un todo en sus partes, a partir de esta acción tendremos los diferentes elementos que conforman el texto de la publicación disponibles de forma individual para tratarlos como variables independientes. (Studies in Computational Intelligence, 2006)
4. Realizar un pre-procesamiento del texto. El pre-procesamiento constará de una indexación de las palabras, y de una reducción de la dimensionalidad para así efectuar una evaluación de la información de una manera más sencilla. La reducción de la dimensionalidad se efectuará eliminando prefijos mediante el algoritmo de Porter Stemming (Chou, Sinha, & Zhao, 2008) y así tener una mayor probabilidad de encontrar las palabras relevantes del texto dentro de los diccionarios pre-clasificados. (Studies in Computational Intelligence, 2006)
5. En relación a la cantidad de likes de los posts que tenga en base a vayas denotadas por percentiles y un promedio sobre los likes se clasificará como popular o no. (Kamel & Campilho, 2009)
6. Se contarán la cantidad de palabras coincidentes de cada diccionario en los comentarios. En base a estos comentarios y su clasificación como negativos o positivos se afectará la clasificación del post al que pertenecen. (D. I. Kramer, E. Guillory, & T. Hancock, 2014)
   1. De acuerdo a lo anterior, el conteo sobre palabras contenidas del post en cada diccionario, se tomará el que sea más significativo para calificar una post como positivo o negativo.
   2. En el caso de que un post carezca de palabras contenidas en el diccionario de palabras positivas o en el diccionario de palabras negativas, se calificará como neutro, sin embargo también se calificará su influencia de acuerdo a los likes a los que se ha hecho acreedor, como se describe anteriormente.
7. Se guardarán los valores obtenidos por este cómputo y el de las variables en la base de datos. Esto con la finalidad de generar una base de conocimientos y en caso de ser necesaria una reevaluación de los comentarios, no sea necesario realizar cada cómputo de nuevo sino solo aquellos en los que se ha realizado alguna modificación en sus algoritmos.
8. En base a un promedio y un parámetro K para el valor mínimo a aceptar, se calificará como influyente o no influyente. Es decir se asignará un valor para definir si al momento de ser evaluado un post, debe ser tomado como positivo o negativo, es decir popular o no. (Kamel & Campilho, 2009)
9. Posteriormente se evaluarán los perfiles de usuario en base a los resultados obtenidos por el proceso de clasificación de sus comentarios anteriormente descrito, por las técnicas de clasificación: Naive Bayes, Decision Tree, K-Nearest Neighbors. (Chou, Sinha, & Zhao, 2008)

4.1.2. ¿Por qué Facebook?

Se optó por utilizar *Facebook* de acuerdo a la alta cantidad de usuario activos al mes, esto representa una importante ventaja para la diversidad del análisis ya que también se debe tomar en cuenta que el acceso a la información estará limitado por aquella que está marcada como pública en cada uno de los perfiles.

La diversidad de información nos permite comparar los resultados sobre los distintos temas que se tratan, asimismo permite comprobar la afinación de los parámetros del algoritmo al ejecutarse sobre un entorno con un amplio espacio de búsqueda, que está dado por entradas grandes que permiten testear realmente la optimización y tiempos de respuesta.

Se encuentran una gran cantidad de herramientas e investigaciones sobre el tema que permiten obtener acceso a una amplia base de datos, para facilitar la investigación, y resolución de dudas que llegan a surgir mediante la práctica y ejecución de los algoritmos.

Es una plataforma gratuita que ofrece la posibilidad de acceso a la información pública lo que abre una ventana a la investigación, que es necesaria para evitar conflictos legales o con las políticas de privacidad del sitio, y que se respeta la decisión del usuario sobre él trato que se le debe dar a su información.

El ambiente de interacción entre los usuarios ofrece elementos que permiten simular de una forma muy similar a la realidad el trato y manejo de la información entre los usuarios, esto permite ofrecer resultados más cercanos y precisos a los que se obtendrían mediante la investigación en un entorno real y no virtual, que posteriormente podrían traer ventajas adicionales, a las conclusiones y elementos obtenidos por la investigación.

A diferencia de Twitter cuenta con “likes” que ofrecen un elemento que facilita el análisis de popularidad. Ya que como se mencionaba anteriormente las palabras expresan y generan sentimientos, que en este caso al compararse con un favorito de Twitter genera un esquema mental completamente diferente, ya que un favorito es más restrictivo o se tienen en cuenta más elementos antes de indicar que algo entra en esta categoría a diferencia de un simple “like” que expresa un gusto por algo.

Ofrece la posibilidad de agregar un sentimiento o estado de ánimo a las publicaciones, como bien se había dicho anteriormente los estados de ánimo pueden descartar o aumentar la participación de una publicación para su análisis ya que puede poner en duda la veracidad de la información que se está transmitiendo, siendo esta una característica esencial para ofrecer un buen resultado preciso sobre el análisis.

Al tener acceso a esta plataforma desde una amplia variedad de dispositivos, permite a los usuarios estar en contacto con la red social durante el transcurso de su día, y así interactuar con esta y ver cómo reaccionan los demás usuarios ante el comportamiento y uso durante estas distintas facetas que experimenta una persona a lo largo de su día. Es decir permite que los distintos contrastes y emociones de una persona se vean reflejados.

Ofrece el uso de smileys que dan un elemento que se ve en las conversaciones cotidianas en persona, que son los gestos y el lenguaje corporal lo que enriquece la información que se puede llegar a publicar en esta red social.

4.1.3. Proceso de desarrollo del Framework PythonSocialNetworkAnalyzer

Se decidió utilizar el lenguaje de programación Python ya que por la facilidad y velocidad al momento de desarrollar, es perfecto para una investigación amplia en la que el tiempo es corto, además de ser un lenguaje de una generación relativamente nueva, cuenta con herramientas de última generación que permiten trabajar con la última tecnología de desarrollo. (SANNER, 2015)

MySQL fue el motor para la base de datos elegido ya que es una herramienta potente y OpenSource para desarrolladores, al ser un amplio espacio de búsqueda y una gran cantidad de información la que sería necesaria para su almacenamiento, se requiere un motor de almacenamiento que soporte operaciones concurrentes al ser un Framework que soporta operaciones distribuidas y un acceso a la información en tiempo real, este motor de base de datos específicamente hablando InnoDB, provee las características necesarias para este Framework e investigación. (Xiaolong, Weiming, & Yonghao, 2012)

Fue necesario indagar en la documentación del lenguaje para obtener ayuda sobre las librerías de soporte para peticiones y respuestas desde la Web, que se encuentra disponible en la siguiente URL: <https://docs.python.org/3/>, al igual que el manejo de los componentes para la conexión a la base de datos en MySQL, que se encuentra disponible en la siguiente URL: <http://dev.mysql.com/doc/connector-python/en/connector-python-example-connecting.html>.

**Capa de Acceso a la Base de Datos**

**Capa de Abstracción (Entidades)**

**Capa de Pre-procesamiento**

**Procesamiento HTML / CSS / Javascript / JSON**

**Procesamiento de Lenguaje Natural**

**API**

**Base de Datos**

4.1.4. Arquitectura del Framework

Imagen 4.0.1: Arquitectura del Framework PythonSocialNetworkAnalyzer.

En la imagen 4.0.1 vemos la arquitectura que el Framework PythonSocialNetworkAnalyzer utiliza para su funcionamiento, la programación está dividida en varias capas y cada una cumple con una función específica para lograr el almacenamiento, pre-procesamiento y análisis adecuado de la información extraía de la red social *Facebook*. Cada capa es descrita a continuación:

* Base de datos: Es la encargada de almacenar la información recabada por los agentes de recolección de perfiles de usuarios, comentarios y publicaciones, también sirve como columna vertebral para el funcionamiento distribuido ya que se encuentra en el servidor y controla las operaciones redundantes, para facilitar y optimizar los procesos de búsqueda y tratamiento de la información.
* Capa de acceso a la base de datos: Es la encargada de realizar la conexión a la base de datos para generar nuevos registros, enviar consultas y modificaciones a los datos almacenados en registros de la base de datos, esta capa se encuentra almacenada en cada instancia del Framework lo que le permite solicitar información al servidor.
* Capa de abstracción o de entidades: En esta capa se representa de forma unitaria cada uno de los elementos sobre los cuales se habla en esta investigación que contiene la red social, se encuentra el código que representa a los usuarios, perfiles de usuario, comentarios, y publicaciones, además de algunas clases para el control interno de la interacción entre las entidades mismas.
* Capa de pre-procesamiento: Es la encargada de tratar la información, de forma previa a su análisis en busca de encontrar nuevo conocimiento, en esta es donde se realiza la limpieza de código HTML, Javascript, interpretación de JSON, CSS, limpieza de caracteres no permitidos, eliminación de sufijos, tokenizacion, identificación de palabras positivas y negativas, re-interpretación de caracteres en formato hexadecimal.
* API: en esta capa es donde ocurre el proceso y la interacción con el desarrollador o investigador final, se encuentran los métodos para búsqueda de nuevo conocimiento a partir de los datos, y para generar nuevos registros sobre información actual de la red social en la base de datos auxiliándose de las capas anteriormente descritas, podría ser vista como una interfaz de usuario carente de formas gráficas, o una interfaz de métodos que permite a investigadores y desarrolladores manipular ciertos aspectos para obtener datos sobre la red social de Facebook.

A continuación se describe cada elemento de las capas del Framework.

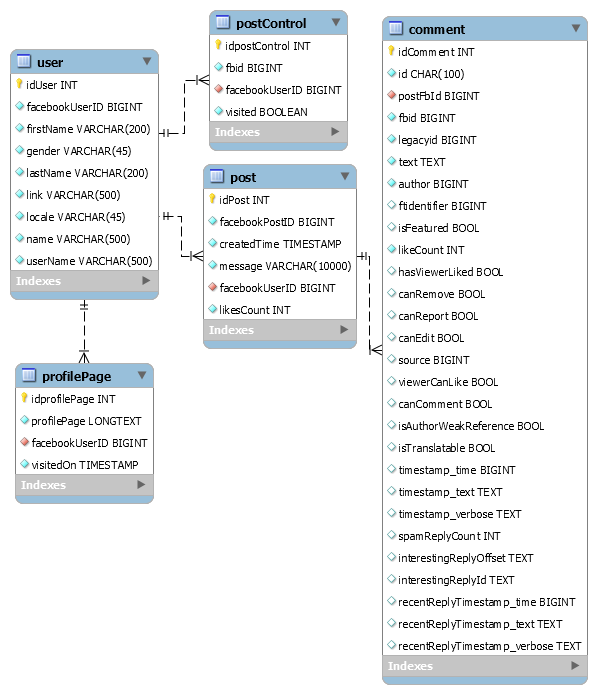


Imagen 4.0.2: Diagrama Entidad-Relación de la base de datos.

La tabla “user” es aquella en la que se almacena la información de los usuarios de Facebook, cabe mencionar que los campos firstName y lastName son solo para completar la estructura de la tabla sin embargo sin un consentimiento especial y expresado en el código no son utilizados al momento de almacenar información en dicha tabla. A continuación se describe la función de cada uno de los campos de esta tabla:

* idUser: es el ID de control interno para cada uno de los registros de la base de datos, es un campo auto-incremental, para evitar problemas al referenciar y hacer las consultas más rápidas al ser un tipo de dato más pequeño que el utilizado por *Facebook*.
* facebookUserID: es el ID del dueño de la cuenta, este ID es único para cada perfil agregado y es el que se describe en la base de datos de *Facebook*.
* firstName: almacena el nombre de pila del propietario del perfil.
* gender: describe el género de la persona registrada.
* lastName: almacena el apellido del dueño de la cuenta.
* link: contiene la URL que vincula a la cuenta del perfil de *Facebook* en la web.
* locale: cadena que describe información sobre el lenguaje y variante del mismo utilizado por el usuario de *Facebook.*
* name: nombre completo de la persona dueña del perfil.
* userName: nombre de usuario elegido o creado por *Facebook* para hacer referencia al usuario del perfil.

La tabla “profilePage” es aquella en la que se almacena el código fuente de la página de perfil de los usuarios que la tengan disponible para ser visualizada de forma pública en *Facebook.* A continuación se describe la función de cada uno de los campos de esta tabla:

* idProfilePage: es un index auto-incrementable de la tabla para llevar un control interno de cada uno de los registros que se almacenan en ella de forma independiente a los otros elementos que en ella se almacenan.
* profilePage: en este campo se almacena el código fuente obtenido de explorar la URL de un perfil de usuario recibido en el response.
* facebookUserID: es el ID del dueño de la cuenta, este ID es único para cada perfil agregado y es el que se describe en la base de datos de *Facebook*.
* visitedOn: en este campo se almacena la fecha en la que fue visitado y almacenado el código fuente del perfil de usuario.

La tabla postControl es la encargada de llevar una sincronización para la actividad distribuida de los agentes, en pocas palabras en ella se lleva el control de los fbid de cada uno de los posts que se visitan y almacenan para evitar que más de uno de los agentes visite dicho post y mejorar el desempeño de los mismos. A continuación se describen cada uno de los campos de esta tabla:

* idPostControl: es un index de control interno de la base de datos que le permite identificar de forma única cada uno de los registros que en ella se almacenan es un campo auto-incrementable.
* fbid: este campo almacena la información del index asignado a la publicación en la base de datos de Facebook con el que se hace referencia la información para poder acceder a su contenido desde el sitio web de dicha red social.
* facebookUserID: es el ID del dueño de la cuenta, este ID es único para cada perfil agregado y es el que se describe en la base de datos de *Facebook* y que en cuyo muro se encuentra la publicación o post.
* visited: es un campo en el cual se almacena si fue exitosa o hubo un error al momento de visitar el perfil del usuario en busca de la publicación.

La tabla post es la encargada de almacenar la información referente a cada uno de los posts o publicaciones que han sido obtenidos por los agentes de extracción, a través de la consulta del fbid perteneciente al post o publicación y al facebookUserID de su propietario. A continuación se describen cada uno de los campos de esta tabla:

* idPost: es un index de control interno y campo auto-incrementable para el control e identificación única de cada uno de los registros que están almacenados en ella.
* facebookPostID: este campo almacena la información del index asignado a la publicación en la base de datos de Facebook con el que se hace referencia la información para poder acceder a su contenido desde el sitio web de dicha red social.
* createdTime: en este campo se almacena la información sobre la fecha en la cual fue guardado el registro con la información del post en la base de datos.
* message: contiene el texto que fue ingresado en la publicación.
* facebookUserID: es el ID del dueño de la cuenta, este ID es único para cada perfil agregado y es el que se describe en la base de datos de *Facebook* y que en cuyo muro se encuentra la publicación o post.
* likesCount: almacena el número de likes que fueron dados al post por otros usuarios de la red social.

La tabla comment contiene la información sobre los comentarios hechos a una publicación de algún registro de la tabla post. A continuación se describe cada uno de los campos que contiene esta tabla:

* idComment: es un index de control interno de la base de datos y campo auto-incrementable, en base a este se identifica cada registro como único.
* id: el id del comentario dentro de la base de datos de Facebook es un campo de tipo entero que lo diferencia de los demás en la base de datos de Facebook.
* postFBId: este campo almacena la información del index asignado a la publicación en la base de datos de Facebook con el que se hace referencia la información para poder acceder a su contenido desde el sitio web de dicha red social, es el id del post al cual está ligado el comentario.
* fbid: identificador interno del comentario en las bases de datos de Facebook.
* legacyid: (campo que Facebook no define ya que ha habido varios issues de seguridad por este)
* text: contenido en texto expresado por el usuario que genera el comentario sobre un post.
* author: id de la base de datos de Facebook de la persona que generó el comentario sobre el post.
* ftidentifier: identificador de contenido extra del comentario.
* isFeatured: campo de tipo bandera que marca cuando un comentario tiene contenido extra.
* likeCount: cantidad de likes que tiene el comentario de la publicación.
* hasViewerLiked: campo bandera que marca cuando el usuario que está visualizando el comentario ha dado like sobre este.
* canRemove: campo bandera que marca cuando el usuario que está visualizando el comentario puede eliminarlo.
* canReport: campo bandera que marca cuando un usuario que está visualizando el comentario puede reportarlo a Facebook por los motivos que este marca.
* canEdit: campo bandera que marca cuando un usuario que está visualizando el comentario puede modificar el contenido del mismo.
* source: (campo no descrito por motivos de seguridad de Facebook).
* viewerCanLike: campo bandera que marca cuando un usuario que está visualizando el comentario puede dar like al mismo.
* canComment: campo bandera que marca cuando un usuario que está visualizando el comentario puede comentar o responder al mismo.
* isAuthorWeakReference: campo bandera que marca cuando el objeto del autor está ligado por una referencia fuerte o una referencia débil, es decir la diferencia de estas dos es al momento de llamar al “recolector de basura” una referencia fuerte impide la recuperación de la memoria a diferencia de una referencia débil.
* isTranslatable: campo bandera que marca cuando el texto del comentario puede ser traducido.
* timestamp\_time: hora en formato de entero largo en la que se generó el comentario.
* timestamp\_text: fecha y hora en formato texto en la que se generó el comentario.
* timestamp\_verbose: fecha y hora en formato de texto largo en la que se generó el comentario.
* spamReplyCount: cuenta de los usuarios que respondieron al comentario como spam.
* interestingReplyOffset: campo no definido en la documentación de Facebook.
* interestingReplyId: campo no definido en la documentación de Facebook.
* recentReplyTimestamp\_time: hora en formato de entero largo en la que se respondió el comentario.
* recenteReplyTimestamp\_text: fecha y hora en formato texto en la que se respondió el comentario.
* recentReplyTimestamp\_verbose: fecha y hora en formato de texto largo en la que se generó el comentario.

A continuación se describe la arquitectura de la capa de acceso a la base de datos:

Imagen 4.0.3: Arquitectura de la capa de acceso a la Base de Datos.

**Capa de Acceso a la Base de Datos**

**Conector a la Base de Datos**

**Transacciones en Comentarios de Facebook**

**Transacciones en Posts de Facebook**

**Transacciones en Comentarios de Facebook**

**Transacciones en Usuarios de Facebook**

**Transacciones en Control de Posts de Facebook**

**Transacciones en Páginas de Perfil de Facebook**

* Conector a la base de datos: en él se configura el protocolo de conexión TCP/IP a la base de datos en el servidor además de contener las credenciales de inicio de sesión en la base de datos.
* Transacciones en Comentarios de Facebook: este elemento de la capa de transacciones de la base de datos contiene los métodos necesarios para realizar las altas bajas y consultas para las entidades de comentarios en la base de datos.
* Transacciones en Posts de Facebook: esta unidad de la capa de acceso a la base datos contiene los métodos necesarios para realizar las altas bajas y consultas para las entidades de comentarios en la base de datos.
* Transacciones en Usuarios de Facebook: este elemento de la capa de acceso a la base de datos contiene los métodos necesarios para realizar las altas bajas y consultas para las entidades de usuarios en la base de datos.
* Transacciones en Control de Posts de Facebook: este elemento de la capa de acceso a la base de datos contiene los métodos necesarios para realizar las altas bajas y consultas para el control de almacenamiento de las entidades de posts de Facebook.
* Transacciones en las Páginas de Perfil de Facebook: este elemento de la capa de acceso a la base de datos contiene los métodos necesarios para realizar las altas bajas y consultas para el control de almacenamiento de las entidades de páginas de perfil de base de datos.

A continuación se describe la arquitectura de la capa de abstracción o entidades:

**Capa de Abstracción (Entidades)**

**Comentarios de Facebook**

**Posts de Facebook**

**Usuarios de Facebook**

**Control de Posts de Facebook**

**Páginas de Perfil de Facebook**

Imagen 4.0.4: Arquitectura de la capa de abstracción de la base de datos.

* Comentarios de Facebook: contiene los atributos y métodos que son necesarios para representar la información que puede ser extraída de un comentario de Facebook por el Framework.
* Posts de Facebook: contiene los atributos y métodos que son necesarios para representar la información que puede ser extraída de un post de Facebook por el Framework.
* Control de Posts de Facebook: contiene los métodos y atributos utilizados para enviar la información que es utilizada en la base de datos para sincronizar el almacenamiento de los agentes a la base de datos.
* Páginas de Perfil de Facebook: contiene los atributos y métodos que son necesarios para representar la información que puede ser extraída de un perfil de Facebook por el Framework.

**Capa de Pre-procesamiento**

**Pre-procesamiento de Lenguaje Natural**

**Mapeo de Símbolos Restringidos**

**Diccionario de Palabras Positivas (Inglés)**

**Diccionario de Palabras Negativas (Inglés)**

**Procesador de Lenguaje**

**Pre-procesamiento HTML / CSS / Javascript / JSON**

**Mapeo de Caracteres Hexadecimales**

**Procesador Lenguajes Web**

* Usuarios de Facebook: contiene los atributos y métodos que son necesarios para representar la información que puede ser extraída de un perfil de Facebook por el Framework.

Imagen 4.0.5: Arquitectura de la capa de pre-procesamiento.

* Pre-procesamiento de Lenguaje Natural: En esta capa se encuentran los elementos necesarios para limpiar y tratar los datos contenidos en el texto en materia de fonética y léxico.
  + Procesador de Lenguaje: En este elemento encontramos los métodos necesarios para hacer el tratamiento y procesamiento adecuado para tratar la parte del lenguaje humano en el texto dejando de lado el código.
  + Diccionario de Palabras Positivas: Estructura de datos de tipo hash table en la que se encuentran mapeadas las palabras que fueron denominadas como positivas en el idioma inglés en la investigación (Chou, Sinha, & Zhao, 2008).
  + Diccionario de Palabras Negativas: Estructura de datos de tipo hash table en la que se encuentran mapeadas las palabras que fueron denominadas como negativas en el idioma inglés en la investigación (Chou, Sinha, & Zhao, 2008).
  + Mapeo de Símbolos restringidos: Estructura de datos de tipo hash table en la que se encuentran mapeados los símbolos que han de ser eliminados del texto para su adecuada interpretación.
* Pre-procesamiento HTML / CSS / Javascript / JSON: En esta capa se encuentran los elementos necesarios para limpiar y tratar los datos contenidos en el texto en materia de código HTML, Javascript e interpretar los formatos JSON para obtener sus datos.
  + Mapeo de Caracteres Hexadecimales: Estructura de datos de tipo hash table en la que se encuentran mapeados los símbolos que han de ser reemplazados por sus caracteres UTF-8 en el texto para su adecuada interpretación.
  + Procesador de lenguajes Web: En este elemento encontramos los métodos necesarios para hacer el tratamiento y procesamiento adecuado para tratar la parte del texto en forma de código.

A continuación en la imagen 4.0.6 se muestra el diagrama de clases detalladas del Framework, para ayudar a futuras investigaciones a implementar dicha herramienta diseñada en Python, y que facilite su exportación a otros lenguajes y herramientas de ser necesario.

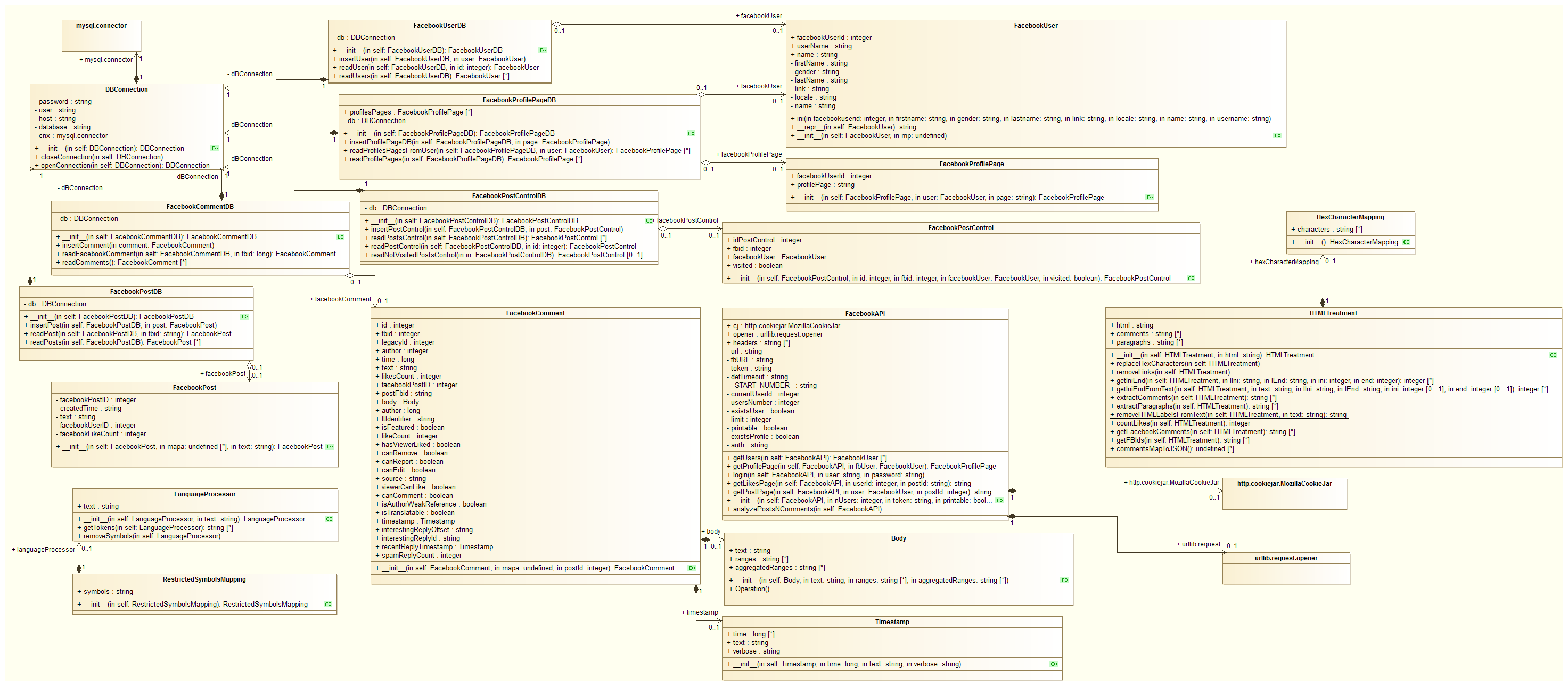


Imagen 4.0.6: Diagrama de Clases Detalladas del Framework FacebookSocialNetworkAnalyzer, lenguaje UML 2.0.

4.1.5. Problemática en el desarrollo del Framework y en la experimentación

Se encontró con varias dificultades en el proceso de desarrollo, y es poca la documentación sobre el manejo de esta red social en el lenguaje Python ya que no hay un Framework parecido que permita el manejo de Minería Web en Facebook para Python.

Antes que nada fue necesario el iniciar sesión en *Facebook*, ya que esta red social no muestra el contenido de los perfiles de usuario a personas que no están registradas, a excepción de los perfiles de personas públicas u organizaciones que en este caso estarían representados por las Fan’s Page que tienen contenido público accesible sin iniciar sesión. Esta característica de seguridad es nueva, ya que tiempo atrás se era posible el acceder a la información pública de los perfiles personales sin necesidad de tener una sesión abierta de un usuario registrado, esto representó un problema ya que el punto de interés en esta investigación son los perfiles personales. Para evitar este problema de seguridad fue necesario abrir una cuenta de Facebook con propósitos de esta investigación que no tuviera a ningún “amigo” en su lista de amigos, para que la información de los perfiles a la que se tuviera acceso fuera únicamente la estrictamente marcada como pública, de igual forma la API Graph de Facebook restringe el acceso a personas que no han iniciado sesión en la plataforma. Se mitigó el problema iniciando sesión en la plataforma con dicha cuenta mandando los siguientes encabezados, simulando ser un navegador Mozilla Firefox 3.5.7 en un sistema operativo Windows NT 5.1.

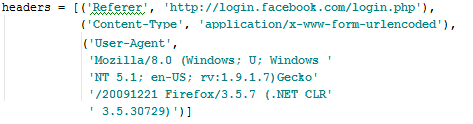


Imagen 4.1: Encabezados necesarios para iniciar sesión en Facebook, desde el Framework desarrollado para Python.

Otra problemática que se enfrentó durante el proceso de recabar información desde el sitio Web de Facebook, es que dado cierto número de peticiones al servidor de Facebook o de la API Graph que fue utilizada para obtener información de los perfiles de usuario, es que se bloquean las respuestas del servidor a la aplicación por lo que es necesario cambiar los encabezados de la aplicación cuando se obtiene continuamente una excepción por Timeout, es que se ha llegado al límite de peticiones impuestos por el servidor de Facebook, en un periodo de tiempo dado, en caso de que no se quiera cambiar los encabezados de la sesión será necesario esperar entre 2 y 24 horas acorde a la sanción que imponga el servidor de Facebook por exceder el límite de peticiones en un periodo de tiempo dado, si es que se quiere continuar recabando información de esta red social, por lo que este proceso puede convertirse en un procedimiento lento y arduo, una arquitectura distribuida con varios agentes que recolectan información utilizando este Framework es parte de la mitigación a este problema de ralentización de la recaudación de la información, ya que se disminuye en proporción a la cantidad de agentes que se tengan recabando información.

Al momento de recabar información de los usuarios mediante la API Graph de Facebook, se obtiene únicamente el ID de usuario, nombre y apellidos, username, url de su perfil (sólo en algunos casos), su sexo y el Locale, este último describe información sobre el idioma y los símbolos de formato para ese usuario en específico. La información obtenida por esta API no siempre muestra disponible la página del perfil de usuario sin embargo sus datos si, por lo que es necesario visitar la página de perfil que es arrojada para analizar si existe o no, lo que aumenta el tiempo necesario para obtener el número de perfiles deseados, sin embargo es más rápido que checar la existencia de cada ID de usuario posible, ya que hay varios huecos en estos ID’s tal vez causados por bajas de usuarios, perfiles cancelados, bloqueados por Facebook, o errores al momento de dar una alta lo que produce un error en los índices auto-incrementables de la tabla, sin restaurarse, esto en conjunto produce grandes huecos entre ID’s de usuario, ejemplo de solicitud de perfil de usuario de acuerdo a su índice: <https://www.facebook.com/155>, esta sentencia hace una petición sobre el perfil de usuario con el ID 155 en las bases de datos de Facebook. A continuación se muestra un ejemplo sobre los espacios en el consecutivo de los índices del auto-incrementable en la tabla de usuarios de Facebook.

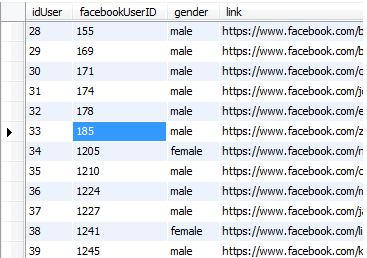


Imagen 4.2: Resultado del query sobre la tabla user en el modelo de la base de datos descrito posteriormente en su diagrama de Entidad Relación, se puede apreciar que el “facebookUserID” se salta índices lo que hace más lenta una búsqueda por fuerza bruta sin utilizar primeramente la API Graph de Facebook.

Otra problemática es que no se puede tener acceso a toda la información de un usuario utilizando la API Graph de Facebook de manera libre ya que se requiere un “token” de autenticación otorgado por Facebook en el que se conciben ciertos permisos de acuerdo a la petición hecha, y consta de una larga cadena de caracteres, sin esta no se puede tener acceso a la información necesaria para leer las publicaciones de los usuarios, por lo que en su defecto fue necesario implementar este motor de búsqueda y consultas dentro del Framework para facilitar a los investigadores, su labor y reduciendo el tiempo que estas toman, un factor importante que se tuvo que tomar en cuenta fue el uso de las condiciones de uso y privacidad de Facebook por lo que no es posible almacenar nombres, ni acceder a ciertas publicaciones de los usuarios, de ahí la necesidad de que el usuario que se encargaría de recopilar información de otros usuarios de Facebook, no contara con amigos en su lista de amigos, ya que si esto fuera e indagara en el perfil de alguno de sus “amigos” obtendría acceso a publicaciones que no han sido marcadas como públicas.

Otro factor que se debe tomar en cuenta es que dada la amplitud de la red social, y la diversidad de idiomas fonéticos y caracteres de escritura se tuvo que limitar a ofrecer una utilidad sobre los lenguajes y caracteres descritos en el “Locale”: “en\_US” y “en\_BR”, que describen al idioma inglés en sus variantes: Americano y Británico, esta utilidad puede ser ampliada agregando nuevos diccionarios de palabras positivas y negativas sobre otros lenguajes, cambiando la cabecera “**en-US;”** descrita al momento de iniciar sesión en el código de la API PythonSocialNetworkAnalyzer que fue desarrollada en esta investigación, y realizando las modificaciones que sean pertinentes en el mapeo de símbolos prohibidos, pre-procesamiento, o en su defecto reemplazar los diccionarios existentes, con nuevos que describan los elementos que cada uno contiene de acuerdo al rol que cada diccionario juega en el Framework.

Otro incidente que se tuvo al momento de manejar de manejar los ID’s de los usuarios, las publicaciones y los comentarios sobre las publicaciones, fue la dimensión que tienen los enteros (su tamaño en bits) esto en la plataforma de la base de datos (MySQL) ya que no se habían tomado en cuenta tal dimensión hasta el momento que se tuvo corriendo el Framework, de tal modo que se comenzó a lanzar Excepciones por parte del motor de la base de datos, ya que dichos enteros podían ser manejados por el intérprete de Python pero no por el motor InnoDB en el esquema actual, por lo que fue necesario cambiar el tipo de dato de dichos ID’s a “BIG INT”, respaldando la información recabada pero no la estructura de la tabla para evitar que al momento de restaurar dicha información se cambiara el tipo de dato de la estructura de la tabla, posteriormente se realizó un “DROP TABLE”, sobre las tablas postControl, post y comment, dicha instrucción elimina del esquema de la base de datos las tablas y su información, después de esto se realizó el cambio en la estructura de la base de datos, posteriormente se sincronizó el modelo con el nuevo tipo de dato para estos ID’s y se realizó la restauración de los datos almacenados posterior al “DROP TABLE” sobre dichas tablas, este proceso solucionó el problema de almacenamiento con enteros grandes, ya que no se podía hacer el cambio de tipo de dato, por las llaves foráneas entre tablas que no permitían hacer un “ALTER TABLE” sobre dichas tablas ya que tendrían que cambiarse los tipos de datos de dos tablas de manera simultánea, la sentencia “ALTER TABLE” modifica la estructura de una tabla o alguno de sus meta-datos que tiene disponibles para su modificación tal como el contador auto-incrementable.

Al momento de analizar la información de los comentarios sobre los posts, fue su estructura irregular, ya que algunos campos aparecían y desaparecían, algunos inclusive no estaban descritos en la información proveída sobre la estructura de los comentarios en la documentación de la API Graph de Facebook, por lo que fue necesario dejar como no restrictivo en valores nulos la mayoría de los campos de la tabla “comment”, en la base de datos y realizar varias corridas para crear un diccionario con los posibles campos que podían ser obtenidos por los 1,039 usuarios obtenidos de forma aleatoria y que cumplían con las características deseadas, para posteriormente generar el modelo de la base de datos, al igual que fue necesario realizar un constructor para los objetos de la clase FacebookComment que recibiera como parámetro un diccionario, y dentro del constructor realizar las validaciones correspondientes para no obtener un KeyError no cachado en tiempo de ejecución que interrumpiera la corrida del programa o que descartara información por un error arrojado y no cachado adecuadamente.

Es necesario eliminar sufijos para llevar las palabras a una forma simplificada y así tener una mayor posibilidad de encontrar una coincidencia en alguno de los diccionarios de palabras ya sea positiva o negativa, de esta forma ofrecer un resultado más exacto, con este propósito se implementó el algoritmo “Porter Stemming” que es de los más utilizados para esta tarea, cuya implementación oficial para varios lenguajes se encuentra alojada en la página oficial de su desarrollador Martin Porter: <http://tartarus.org/martin/PorterStemmer/python.txt>. (Chou, Sinha, & Zhao, 2008)

También es necesario realizar un pre-procesamiento de los signos de puntuación, para esto en base a un diccionario, teniendo de llave un el signo de puntuación a ser removido, y la función “\_\_replace()\_\_” de la clase String de Python reemplazando los caracteres en este mapeo con espacios en blanco.

La forma de localizar posibles posts en el código fuente de un perfil de Facebook es buscando la expresión “fbid=”, posterior a esta expresión aparece un numero entero, se puede comprobar que dicho numero represente un post en el perfil del usuario visitando la siguiente URL: <https://www.facebook.com/> + username + “/posts” + fbid, si no se recibe un error 401, o una página con la expresión “Sorry” este “fbid”, representa un post en la base de datos de Facebook y se obtiene el código fuente con comentarios y el número de likes del post visitando la URL anteriormente descrita, sin embargo no todos los posts contienen texto ya que algunos son únicamente imágenes o fotografías publicadas por el usuario, por el momento el Framework PythonSocialNetworkAnalyzer no permite el análisis digital de imágenes por lo que estos posts deben descartarse, se puede comprobar si contienen texto, al buscar en la propiedad “body.text” si el valor de dicha propiedad es igual a “” (cadena vacía) se debe descartar dicho post para su almacenamiento en la base de datos.

Otra problemática que hasta el momento se tiene y debido al patrón de diseño que tiene Facebook al momento de cargar los comentarios hechos por los usuarios sobre alguna publicación, no es posible obtener más de 30 comentarios por publicación aproximadamente, para esta área se requiere más trabajo sobre la simulación de eventos hechos por el usuario, es el mismo caso para los likes hechos sobre alguna publicación, pero en este caso no se pueden obtener más 70 likes aproximadamente accediendo a la URL <https://www.facebook.com/browse/likes?id=> + fbid del post + '&actorid=' id de usuario, esta URL carga la página de likes de la publicación o post, sin embargo es necesario mejorar la simulación de eventos hechos por el usuario en el Framework para obtener una mayor cantidad de datos de usuario sobre los likes hechos a la publicación o post.

Un elemento que causó conflicto y se tuvo que resolver fue el hecho de que se usan caracteres en su representación hexadecimal en el texto, por lo que estos deben ser encontrados y reemplazados, para esta tarea primero se tuvo que extraer el texto, y posteriormente, utilizando un mapeo de dichos caracteres en su representación hexadecimal como llave de un diccionario, y como valor de dicha llave su representación en código UTF-8. La estructura de datos de dicho mapeo es la siguiente:

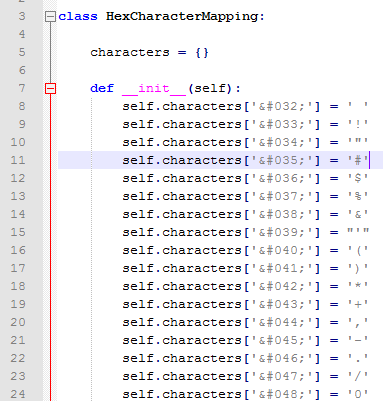


Imagen 4.3: Diccionario para el mapeo de los caracteres en su representación hexadecimal del lado de la llave y en el lado del valor asociado a dicha llave la representación en código UTF-8.

4.1.6. Arquitectura del Clúster



Imagen 5: Diagrama de red, que describe la arquitectura del clúster (Fuente: Cisco Packet Tracer).

4.1.7. Descripción del Clúster:

* Server MySQL:
  + HP ProLiant ML310e Gen8 V2
  + Disco duro estado rígido, 250 Gb para sistema
  + 2 Discos duros 1 Tb en RAID1
  + Tarjeta controladora smart array b120i
  + 8 Gb en RAM
  + 2 Tarjetas de red Gigabit redundantes
  + Sistema Operativo Windows Server 2008 R2
  + MySQL Server (v. 5.6.21)
* Leader Detection Nodes:
  + Sistema Operativo Windows 7 amd64
  + 4 Gb RAM
  + Procesador Intel Celeron 2 núcleos
  + Tarjeta de red Fast Ethernet
  + Disco duro 500 Gb
* Leader Analysis Nodes:
  + Sistema Operativo Windows 7 amd64
  + 4 Gb RAM
  + Procesador Intel Core 2 duo, 2 núcleos
  + Tarjeta de red Fast Ethernet
  + Disco duro 500 Gb
* Descripción de la red:
  + Cableado UTP Cat 5e
  + Switch TP-Link TL-SF1024
  + Conexión a Internet Simétrica 100 mbps
  + Protocolo TCP/IP v4

4.1.8. Funcionamiento del Clúster:

* Servidor MySQL:
  + Cumplirá con el rol de controlar la base de datos, la albergará y su motor está hecho sobre MySQL Server, cumplirá con el rol de búfer para el control de entradas y salidas de los algoritmos, hablando sobre la información que será analizada y los resultados de la información analizada.
  + Es el elemento central de la red, aunque no tiene ningún contacto directo con Internet, sino que su base de datos es alimentada por los algoritmos que se encuentran funcionando en la red del clúster.
  + Ejecutará los queries para ingresar, consultar, eliminar y modificar información en la base de datos, de acuerdo a lo que los algoritmos ubicados en los clientes.
  + Se decidió separar la base de datos de los algoritmos y ponerla como un nodo aparte para permitir que la información en tanto a ingresos, modificaciones y eliminaciones se efectúen mucho más rápido, y se tenga centralizada para que varios clientes tengan acceso a ella de forma concurrente.
* Leader Detection Nodes:
  + Su función es ejecutar los algoritmos para detectar aquellos nodos calificados como posibles líderes, esto lo realizará a manera de Web Crawler, donde navegará a través de los perfiles y en base a la cantidad de likes que se encuentren en sus publicaciones en comparación a la cantidad de personas en su lista de amigos, en comparación a un parámetro K, los agregará a lista de perfiles a analizar o no.
  + Al detectar un posible líder se agrega a la base de datos que alimentará un búfer para el control de entradas y salidas en este caso representa las entradas.
  + Volverá a considerar a un nodo como posible líder después de N cantidad de tiempo, o un incremento o decremento X en su lista de amigos, que indique un posible aumento en su popularidad y su influencia, de lo contrario se evitará su análisis en cuestión de volver a entrar en su perfil por medio de un Web Crawler.
* Leader Analysis Nodes:
  + Su función es ejecutar los algoritmos para calificar aquellos nodos calificados como posibles líderes, y en base a un porcentaje de liderazgo aceptable M rechazarlos o aceptarlos. Este porcentaje es calculado en base a la fórmula descrita en párrafos anteriores, y de acuerdo al resultado que esta arroje en comparación con el parámetro de aceptación.
  + En este nodo es donde se realiza el análisis a través de minería de texto, para detectar la influencia de las publicaciones.
  + Tomará los nodos a calificar de la base de datos albergada en el servidor, que tendrá un papel de búfer.
  + Al calificar a un nodo como líder agregará dicho nodo a la base de datos a la tabla “líder”, de lo contrario sólo marcará el nodo como calificado para descartarlo.

# 5. Resultados

Imagen 6.1: Comparativa de cantidad likes obtenidos, excluyendo a los usuarios 41 y 4 para ofrecer una mejor apreciación.

En la imagen 6 se describe el comportamiento de la cantidad de likes hecha sobre los posts de los usuarios en la red Social de Facebook tomando en cuenta una muestra de 1,034 usuarios, de los cuales se encontraron 1,227 posts hábiles para ser analizados, descartando aquellos que tuvieran características tales como sólo signos de puntuación, texto en blanco y solo imágenes, sólo emoticones o *smileys*. Se logra apreciar en la gráfica que el punto máximo representado por una cantidad de 2655 likes fue alcanzado por el usuario con ID = 7108, y el punto mínimo fue alcanzado por un conjunto de usuarios con una cantidad de 0 likes en todas sus publicaciones, cabe tomar en cuenta que en esta gráfica se descartaron los resultados obtenidos por los usuarios con ID = 4 y 41, ya que correspondían a niveles muy por arriba de la media de 549 likes por publicación que se obtuvo de dichos datos, con una cantidad de likes de 554,260 y 77,982 respectivamente, con el fin de ofrecer una mejor apreciación de los datos, a este par de usuarios hemos de referirnos como elementos altamente influyentes ya que si tomamos en cuenta la media de 549 likes por post y estos alcanzaron un total de más de 77,000 likes se puede decir que al menos dicha cantidad de personas concuerdan con lo que se publica en sus muros.

Imagen 6.2: Cantidad de likes representando la frecuencia esperada por usuario de Facebook.

En la imagen 6.2 sorprendentemente se puede observar que el número de likes más frecuentemente visto en los posts de los usuarios de Facebook es el 0 con un total de 203 posts de un total de 1,227 de este estudio, seguido por 1 like con un total de 88 posts, 2 likes con un total de 61 post. Teniendo en cuenta los 10 resultados con una mayor cantidad de posts vemos que el top 10 de cantidad de likes corresponde a los números: 0, 1, 2, 3, 7, 6, 9, 5, 4, 12 respectivamente siendo este último con acreedor a un total de 24 publicaciones que comparten esta cantidad de likes. El número máximo de likes alcanzado por más de un post fue 211, mientras que el número máximo de likes alcanzado por al menos un post fue de 156,366 likes. La media sobre la cantidad de likes sobre al menos 2 post es de 59 likes, mientras que en contraste la media de cantidad likes sobre un solo post es de 6,853 likes. Con fines de este estudio la probabilidad de ver un post con una cantidad de 45 likes es la misma que la de ver un post de 156,366 likes descrita por 1/1,227.

Imagen 6.3: Gráfica con la representación de la cantidad de posts clasificados como positivos, negativos o neutros.

En la imagen 6.3 se observa la cantidad de posts clasificados como positivos, negativos o neutros, se observa que la cantidad predominante de posts clasificados se encuentran etiquedos como neutros con una cantidad de 774 de los 1,227 posts analizados representando el 63.1%, seguido por los posts clasificados como positivos con una cantidad de 345 representado el 28.1% y en último lugar vemos a aquellos clasificados como negativos con un total de 108 representando el 8.8% del total de posts analizados.

Imagen 6.4: Esta gráfica representa la cantidad de likes recibidos acorde a cada clasificación.

De acuerdo a la imagen 6.4 se observa que es más factible obtener una mayor cantidad de likes si se hace una publicación positiva con un total de 608,688 likes recibidos, en contraste con un total de 3,897 likes recibidos por la categoría de publicaciones negativas, mientras que los posts marcados como neutros recibieron un total de 61,167 likes. Tomando en cuenta los datos anteriores podemos decir que se espera el 90% de los likes que se generan en Facebook sean recibidos por usuarios que generan publicaciones positivas, el 9.07% lo recibirán aquellos usuarios con publicaciones neutras, mientras que tan solo el 0.93% lo recibirán aquellas publicaciones marcadas como negativas.

Imagen 6.5: Cantidad de comentarios positivos, negativos y recibidos acorde a la clasificación de los posts.

En la imagen 6.5 se observa que la cantidad de comentarios positivos recibidos sobre una publicación positiva es mucho mayor que la recibida por las otras clasificaciones con un total 7,233 comentarios positivos, mientras que la clasificación neutral recibió 1,158 comentarios positivos y la negativa solo 224 comentarios positivos. En contraste la clasificación positiva recibió 6,933 comentarios negativos, mientras que la neutral 223 y la negativa 79. Acorde a esto podemos decir que las publicaciones negativas son las que reciben una menor cantidad de interacción sobre opiniones de las mismas, mientras que las positivas son las que tienden a generar una mayor cantidad de cadenas de información. Analizando los porcentajes de estos datos vemos que las publicaciones negativas los comentarios positivos representan un total de 66.3% mientras que sus comentarios negativos representan solo un 23.4% y el porcentaje restante un 10.1% representan aquellos comentarios clasificados como neutros. Las publicaciones neutras tienen un total de 54.2% de sus comentarios clasificados como positivos, un 10.4% de sus comentarios clasificados como negativos y el 35.4% restante clasificados como neutros. Las publicaciones clasificadas como positivas tienen un total de 24.8% de sus comentarios clasificados como positivos, un 23.7% de sus comentarios clasificados como negativos y el 51.5% de sus comentarios marcados como neutros. A partir de estos datos se puede decir que las publicaciones negativas tienden a recibir un mayor porcentaje de comentarios positivos, mientras que las publicaciones positivas reciben una mayor cantidad de comentarios neutros y las publicaciones neutras reciben una mayor cantidad de comentarios negativos y de las 3 clasificaciones son las que reciben una menor cantidad de comentarios negativos.

|  |  |
| --- | --- |
| **Id usuario** | **Suma de post’s\_likes** |
| 4 | 554260 |
| 41 | 77982 |
| 5 | 3489 |
| 7108 | 2655 |
| 7542 | 1647 |
| 9109 | 1558 |
| 1681 | 1187 |
| 3818 | 1080 |
| 6454 | 1055 |
| 5143 | 981 |
| **Total general** | **645894** |

Imagen 6.6 Tabla de los 10 usuarios más influyentes del estudio de acuerdo a la cantidad de likes de sus publicaciones.

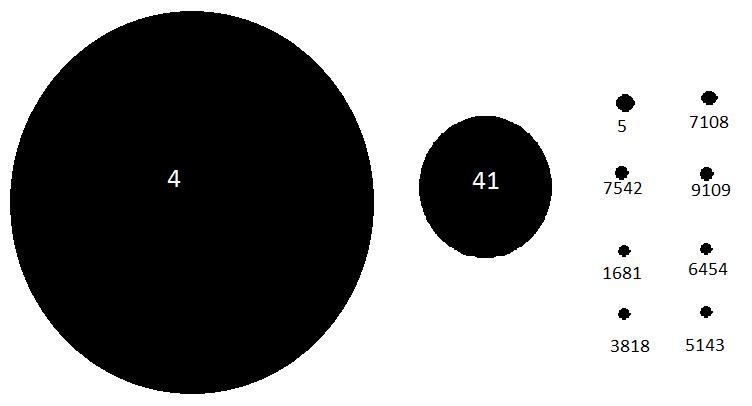


Imagen 6.7: comparativa gráfica de la cantidad de usuarios mesurable a la cual llegaron las publicaciones de sus autores, los números representan el id de usuario.

La imagen 6.6 y 6.7 describe el top 10 de los usuarios más influyentes del estudio en esta se muestra el ID del usuario de Facebook en la columna izquierda y el total de likes de las publicaciones analizadas pertenecientes a dicho ID, mientras que en la imagen 6.7 mientras mayor es el tamaño del circulo mayor fue la influencia de dicho usuario. Considerando la media obtenida de 549 likes por usuario, vemos que el último elemento de este conjunto rebasa por 442 likes a dicha cifra, mientras que el primer elemento de este conjunto rebasa por 553,711 likes a esta media. A través de estos usuarios se logró tener acceso a la sección de noticias en conjunto con las visitas al perfil de dichos usuarios de al menos 645,894 usuarios del total de 673,752 usuarios identificados que se movieron en la red de los perfiles puestos en estudio que representa un 95.9% del total de usuarios, es decir a través de estos 10 perfiles se tiene acceso a este porcentaje de usuarios de los círculos sociales que se vieron en este estudio.

# 6. Conclusiones

Es posible determinas aquellos miembros de un círculo social que juegan un papel de líderes mediante el análisis del código fuente y texto de sus perfiles en la red social de Facebook, tomando en cuenta elementos como la cantidad likes y comentarios para ver el impacto mesurable en su círculo social, ya que a través de estos número es que podemos ver al menos a cuantas personas llego dicha información representada en una publicación. La cantidad de likes implica una aceptación de los usuarios con lo información que les es presentada, un apoyo, o estar de acuerdo con tal información que les es presentada, lo que nos permitió determinar el grado de popularidad del post. Gracias a los diccionarios de palabras obtenidos de investigaciones realizadas anteriormente de los cuales se habla en específico en la sección 3. de este documento nos permite clasificar los post y comentarios en 3 grupos (positivos, negativos y neutros) acorde a la cantidad de palabras que contiene de cada uno de estos diccionarios, y en base a esta clasificación analizar el posible impacto esperado de la información presentada al pertenecer a cierto grupo y de esta forma en base a la cantidad de posts pertenecientes a cierto grupo de cada usuario poder determinar un posible elemento líder en su círculo social en conjunto con la cantidad promedio de likes recibidos en cada uno de sus posts, inclusive determinar el grado de impacto de una publicación al momento de ser lanzada a una red social acorde a la cantidad promedio de likes que contienen los posts de su autor y al grupo al cual estos pertenecen. Este último tópico que es tratado representa un campo fértil para la investigación y continuación de esta misma, ya que el hecho de predecir el impacto que tendrá una publicación antes de ser lanzada a una red social tomando en cuenta factores como su autor, la clasificación de dicha publicación y la cantidad promedio de likes por clasificación y estadísticos del autor, determinar un probable número de personas a los cuales esta información llegará, acorde a las estadísticas del usuario cuantos han de estar de acuerdo con él, cuantos en desacuerdo y cuantos tengan interacción con el post sin demostrar concordancia con esta o desacuerdo a la misma.

La mayor aportación de esta investigación a futuras investigaciones es el Framework en Python para análisis de datos sociales aplicando la minería de texto, los diagramas de ingeniería de software que permitirán migrar la plataforma a otros lenguajes de ser necesario para otras investigaciones y la base de datos generada para esta investigación todos estos datos están disponibles en la siguiente [URL: https://github.com/idcodeoverflow/SocialNetworkAnalyzer/](https://github.com/idcodeoverflow/SocialNetworkAnalyzer/tree/), y la comprobación de que se puede clasificar a los usuarios de una red social a partir de la información que en ella comparten utilizando minería de datos, minería de textos y minería web.

# Bibliografía

Abraham, A. (2012). Extraction and Analysis of Facebook Friendship Relations. In S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, & A. Provetti, *Computational Social Networks* (pp. 291-324). London: Springer London.

Adnan, M., Nagi, M., Kianmehr, K., Tahboub, R., Ridley, M., & Rokne, J. (2011). Promoting where, when and what? An analysis of web logs by integrating data mining and social network techniques to guide ecommerce business promotions. *Social Network Analysis Mining* , 173-185.

Agrawal, D., Budak, C., El Abbadi, A., Georgiou, T., & Yan, X. (2014). Big Data in Online Social Networks: User Interaction Analysis to Model User Behavior in Social Networks. In M. Aastha, K. Shinji, & B. Subhash, *Databases in Networked Information Systems* (pp. 1-16). Springer International Publishing.

Alhajj, R., & Memon, N. (2011). Introduction to the second issue of Social Network Analysis and Mining journal: scientific computing for social network analysis and dynamicity. *Social Network Analysis and Mining* , 73-74.

Ali Abbas, S. M. (2013). An agent-based model of the development of friendship links within Facebook. *Computational Mathematical Organization Therory* , 232-252.

Al-Saggaf, Y., & Islam, M. Z. (2014). Data Mining and Privacy of Social Network Sites Users (Implications of the Data Mining Problem). *Science Engineering Ethics* .

Bentley, R., Ormerod, P., & Batty, M. (2011). Evolving social influence in large populations. *Behavioral Ecology and Sociobiology* , 537-546.

Blenn, N., Doerr, C., Van Kester, B., & Van Mieghem, P. (2012). Crawling and Detecting Community Structure in Online Social Networks Using Local Information. In R. Bestak, L. Kencl, L. Erran Li, J. Widmer, & H. Yin, *NETWORKING 2012* (pp. 56-67). Berlin: Springer Berlin Heidelberg.

(2009). Investigation on the Relationships among Media Characteristics, Presence, Flow, and Learning Effects in Augmented Reality Based Learning. In K. Bokyung, *Multimedia and E-Content Trends* (pp. 21-37). Bruck, PeterA.

Bozkır, A., Güzin Mazman, S., & Akçapınar Sezer, E. (2010). Identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case. In S. Kurbanoğlu, U. Al, P. Lepon Erdoğan, Y. Tonta, & N. Uçak, *Technological Convergence and Social Networks in Information Management* (pp. 145-153). Springer Berlin Heidelberg.

Cao, L., Weiss, G., & Yu, P. (2012). A brief introduction to agent mining. *Autonomous Agents and Multi-Agent Systems* , 419-424.

Carolin, K., Johannes, K., & Freimut, B. (2013). Simulating the spread of opinions in online social networks when targeting opinion leaders. *Information Systems and e-Business Management* , 597–621.

Chou, C.-H., Sinha, A. P., & Zhao, H. (2008). A text mining approach to Internet abuse detection. *Information Systems E-Bussines Management* , 419-439.

Copeland, L. (1988). Valuing workplace diversity. 52-60.

Cox, T. J. (1993). Cultural Diversity in Organizations. *Research and Practice* , 17.

Cox, T. J., Lobel, S., & McLeod, P. (1991). Effects of ethnic group cultural difference on cooperative versus competitive behaviour in a group task. 827-847.

(2008). Diversity and Managing Diversity: Theory and State of the Art. In *Managing Diversity in Intergovernmental Organisations* (pp. 63-156). VS Verlag für Sozialwissenschaften.

Fazeen, M., Dantu, R., & Guturu, P. (2011). Identification of leaders, lurkers, associates and spammers. *Springer-Verlag* , 241-254.

Felton, B., & Berry, C. (1992). Groups as social network members: Overlooked sources of social support. *American Journal of Community Psychology* , 253-261.

Fleuriet, C., Cole, M., & Guerrero, L. K. (2014). Exploring Facebook: Attachment Style and Nonverbal Message Characteristics as Predictors of Anticipated Emotional Reactions to Facebook Postings. *Journal of Nonverbal Behavior* , 429-450.

Fogg, B., & Iizawa, D. (2008). Online Persuasion in Facebook and Mixi: A Cross-Cultural Comparison. In H. Oinas-Kukkonen, P. Hasle, M. Harjumaa, K. Segerståhl, & P. Øhrstrøm, *Persuasive Technology* (pp. 35-46). Berlin: Springer Berlin Heidelberg.

K. Tanbeer, S., & Kai-Sang Leung, C. (2012). Mining Social Networks for Significant Friend Groups. *Database Systems for Advanced Applications* , 180–192.

Kimura, M., Saito, K., Nakano, R., & Motoda, H. (2010). Extracting influential nodes on a social network for information diffusion. *Data Mining and Knowledge Discovery* , 70-97.

Kumara, S., & Cavdur, F. (2014). Network mining: Applications to business data. *Information Systems Frontiers* , 473–490.

Leung, C.-S., Tanbeer, S., & Cameron, J. (2014). Interactive discovery of influential friends from social networks. *Social Network Analysis and Mining* , 154.

Liu, B. (12, 05 2015). *Univerity of Illinois at Chicago*. Retrieved 05 15, 2015, from Univerity of Illinois at Chicago: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

Liu, L., Tang, J., Han, J., & Yang, S. (2012). Learning influence from heterogeneous social networks. *Data Mining and Knowledge Discovery* , 511-544.

Livingston, K., Triquenaux, N., Fighiera, T., Beyler, J., & Jalby, W. (2014). Computer using too much power? Give it a REST (Runtime Energy Saving Technology). *Computer Science - Research and Development* , 123-130.

Lusch, R., & Laczniak, G. (1987). The evolving marketing concept, competitive intensity and organizational performance. *Journal of the Academy of Marketing Science* , 1-11.

Mandell, B., & Kohler-Gray, S. (1990). Management development that values diversity. 41-47.

Memon, N., & Alhajj, R. (2011). Introduction to the first issue of Social Network Analysis and Mining journal. *Social Network Analysis and Mining* , 1-2.

Minazzi, R. (2015). Social Media Metrics and Analysis. In R. Minazzi, *Social Media Marketing in Tourism and Hospitality* (pp. 137-163). Springer International Publishing.

Moradian Zadeh, P., & Sadighi Moshkenani, M. (2008). Mining Social Network for Semantic Advertisement. *Third 2008 International Conference on Convergence and Hybrid Information Technology* , 611-618.

Morzy, M. (2008). New Algorithms for Mining the Reputation of Participants of Online Auctions. *Algorithmica* , 95–112.

Obradovic, D., Baumann, S., & Dengel, A. (2013). A social network analysis and mining methodology for the monitoring of specific domains in the blogosphere. *Social Network Analisis Mining* , 221–232.

Ozar, D. T. (1984). Social rules and the actions of the groups: control of physical objects. *J. Value Inquiry* , 18:23.

Peters, M., & Araya, D. (2007). Networks, Information Politics and the New Paradigm of Social Production. In P. Smeyers, & M. Depaepe, *Educational Research: Networks and Technologies* (pp. 33-42). Netherlands: Springer Netherlands.

Raad, E., Chbeir, R., & Dipanda, A. (2013). Discovering relationship types between users using profiles and shared photos in a social network. *Multimed Tools Appl* , 141–170.

Rabade, R., Mishra, N., & Sharma, S. (2014). Survey of Influential User Identification Techniques in Online Social Networks. In S. M. Thampi, A. Abraham, S. K. Pal, & J. M. Rodriguez Corchado, *Recent Advances in Intelligent Informatics* (pp. 359-370). Springer International Publishing.

Romsaiyud, W., & Premchaiswadi, W. (2011). Applying Mining Fuzzy Sequential Patterns. *IEEE* , 134 - 137.

Safaei, M., Sahan, M., & Ilkan, M. (2009). Social Graph Generation & Forecasting using Social Network Mining. *33rd Annual IEEE International Computer Software and Applications Conference* .

Saito, K., Kimura, M., Ohara, K., & Motoda, H. (2012). Efficient discovery of influential nodes for SIS models in social networks. *Knowledge and Information Systems* , 613-635.

Saito, K., Kimura, M., Ohara, K., & Motoda, H. (2012). Efficient discovery of influential nodes for SIS models in social networks. *Knowledge and Information Systems* , 613-635.

Salazar, J. M. (1979). *Psicología Social.* México: Editorial Trillas.

Scott, J. (2011). Social network analysis: developments, advances, and prospects. *Social Network Analysis and Mining* , 21-26.

Shahriari, M., & Jalili, M. (2014). Ranking Nodes in Signed Social Networks. *Social Network Analysis and Mining* , 4:172.

Shapira, B., Rokach, L., & Freilikhman, S. (2013). Facebook single and cross domain data for recommendation systems. *User Modeling and User-Adapted Interaction* , 211-247.

Shin, S.-j., Jeong, Y.-j., Kim, C.-M., Han, Y.-H., & Park, C. (2014). Study on Relation between Social Circles and Communities in Facebook Ego Networks. In Y.-S. Jeong, Y.-H. Park, C.-H. (. Hsu, & J. J. Park, *Ubiquitous Information Technologies and Applications* (pp. 567-572). Berlin: Springer Berlin Heidelberg.

Sidanius, J., & Pratto, F. (2001). *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression.* New York: Cambridge University Press.

Simpson, T., Barron, K., Frecker, M., Barton, R., & Ligetti, C. (2007). Impact of response delay and training on user performance with text-based and graphical user interfaces for engineering design. *Research in Engineering Design* , 49-65.

Soo Cho, K., Yoel Yoon, J., Joon Kim, I., Yeon Lim, J., Kwan Kim, S., & Kim, U.-M. (2011). Mining Information of Anonymous User on a Social Network Service. *International Conference on Advances in Social Networks Analysis and Mining* .

Steiner, T., Verborgh, R., Gabarró Vallés, J., & Van de Walle, R. (2011). Adding Meaning to Facebook Microposts via a Mash-up API and Tracking Its Data Provenance. *IEEE* , 4.

Stokes, J. (1983). Predicting satisfaction with social support from social network structure. *American Journal of Community Psychology* , 141-152.

Tang, C., Ross, K., Saxena, N., & Chen, R. (2011). What’s in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook. In J. Xu, G. Yu, S. Zhou, & R. Unland, *Database Systems for Advanced Applications* (pp. 344-356). Berlin: Springer Berlin Heidelberg.

Tang, L., Wang, X., & Liu, H. (2012). Community detection via heterogeneous interaction analysis. *Data Mining Knowledge Discovery* , 1–33.

Tchuente, D., Canut, M.-F., Jessel, N., Peninou, A., & Sédes, F. (2013). A community-based algorithm for deriving users’ profiles from egocentrics networks: experiment on Facebook and DBLP. *Social Network Analisys Mining* , 667–683.

Tsai, M.-F., Tzeng, C.-W., Lin, Z.-L., & Chen, A. (2014). Discovering leaders from social network by action cascade. *Social Network Analysis and Mining* , 165.

Tsung-Yu, L., Tan-Hsu, T., & Yu-Ling, C. (2009). QR Code and Augmented Reality-Supported Mobile English Learning System. In *Mobile Multimedia Processing* (pp. 21-37). Bokyung, Kye.

Uhlmann, E. L., Brescoll, V. L., & Machery, E. (2010). The Motives Underlying Stereotype-Based Discrimination Against Members of Stigmatized Groups. *Social Justice Research* , 1–16.

Wang, S., Wang, F., Chen, Y., Liu, C., Li, Z., & Zhang, X. (2014). Exploiting social circle broadness for influential spreaders identification in social networks. *World Wide Web* , 1-25.

Wang, Y., Huang, W., Zong, L., Wang, T., & Yang, D. (2013). Influence maximization with limit cost in social network. *Science China Information Sciences* , 1-14.

Yang, B., Huang, J., Liu, D., & Liu, J. (2009). A Multi-Agent Based Decentralized Algorithm for Social Network Community Mining. *Advances in Social Network Analysis and Mining* , 78-82.

Yao, H.-R., & Ting, I.-H. (2014). Topic Participation Algorithm for Social Search Engine Based on Facebook Dataset. In L.-L. Wang, J. June, C.-H. Lee, K. Okuhara, & H.-C. Yang, *Multidisciplinary Social Networks Research* (pp. 158-170). Berlin: Springer Berlin Heidelberg.

Zhang, E., Wang, G., Gao, K., & Yu, G. (2013). Finding critical blocks of information diffusion in social networks. *World Wide Web* , 1-17.

Zhang, Y., Wang, Z., & Xia, C. (2010). Identifying Key Users for Targeted Marketing by Mining Online Social Network. *IEEE 24th International Conference on Advanced Information Networking and Applications Workshops* .

Zimmer, M. (2010). ‘‘But the data is already public’’: on the ethics of research in Facebook. *Springer Science+Business Media* , 313-325.