Pontificia Universidad Católica del Perú

Escuela de Posgrado

Modelo computacional de minería de microblogs para el análisis del comportamiento del consumidor de telefonía celular

Tesis para optar el grado de Magíster en Informática con mención en Ciencias de la Computación que presenta

Santiago Hernán Apaza Delgado

Dirigido por

Dr. César Armando Beltran Castañon

Jurado

Dr. Héctor Andrés Melgar Sasieta

Dr. César Armando Beltran Castañon

Dr. Hugo Alatrista Salas

San Miguel, 2016

Este trabajo es dedicado a la mujer que amo, a Rosmery, que me apoya siempre, que me da fuerzas, que esta siempre para mi a veces olvidándose de si misma y que su amor y su sonrisa es el aliciente que necesito, cuando 5 minutos más de trabajo hacen la diferencia. A mi madre que nos quiere a sus hijos por igual, que me enseñó que quemarse las pestañas siempre tiene una recompensa y que siempre estará ahí para nosotros. A mis hermanos que nunca dudaron de mi y me apoyaron siempre, a mis amigos de la universidad junto a los cuales aprendí y tuve la oportunidad de enseñar. A mi asesor el Doctor César Beltran que me guió para poder culminar este trabajo que no es el final, sino solo el inicio de muchas cosas más. Y sobre todo a Diosito que me brinda cada día la oportunidad de seguir aprendiendo, de seguir amando y de seguir siendo amado.

Abstract

Twitter messages are being increasingly used to determine the sentiment of consumers of services or products. To do this using various computational techniques are made and adapted from traditional text classification problems and recent models using machine learning. In both approaches must be developed a series of stages ranging from pre - processing to evaluation.

This document shows the result of apply several Sentiment Analysis techniques to assign a positive, negative or neutral polarity from tweets of cellphone consumers in Peru, with the purpose to identify which is the behavior that customers manifest to cellphone companies represented in the opinions expressed in comments on Twitter.

Was collected 26,917,539 publications from Twiter that were extracted in 2 periods, each one of 30 days. These publications are the tweets of the followers of three cellphone companies in Peru, including a relatively new one in the Peruvian market. The procedure included the following tasks: a) Collection of tweets of the followers of the cellphone companies; b) Pre–processing of the data obtained to identify important elements of each tweet; c) Filtering irrelevant elements or noise; and d) Classification of each publication based on the characteristics obtained in previous stages.

The results show that the introduction of a dictionary of lexicons increased the number of terms that can be considered for classification. Likewise, the use of this dictionary allowed to increase the classification rate on a 0,75 %. Finally, thanks to these techniques sentiment analysis, it is possible to exploit the content of social networks so that they can serve corporations for decision-making, especially service to its users.

Keywords: Sentiment Analysis, Opinion Mining, Twitter message classification, Dictionary of lexicons, hashtags

Resumen

Los mensajes de Twitter están siendo cada vez más usados para determinar el sentimiento de los consumidores de servicios o productos. Para ello se hacen uso de diversas técnicas computacionales, desde las tradicionales adaptadas de problemas de clasificación de textos y las recientes que usan modelos de aprendizaje de máquina. En ambos enfoques se debe desarrollar una serie de etapas que van desde el pre–procesamiento hasta la evaluación.

El presente documento muestra el resultado del proceso de aplicación de diversas técnicas de Análisis de Sentimiento para poder asignar una polaridad positiva, negativa o neutral a los tweets de los consumidores de telefonía celular en el Perú, con la finalidad de poder identificar cual es el comportamiento que presentan los clientes de las empresas de telefonía celular representado en opiniones vertidas en la red social Twitter.

Para ello se extrajeron 26,917,539 publicaciones de la red social Twiter durante 2 periodos, cada uno de 30 días. Estas publicaciones corresponden a los tweets de los seguidores de tres empresas de telefonía celular en el Perú, incluyendo una relativamente nueva en el mercado peruano. El procedimiento seguido comprendió las siguientes tareas: a) Recolección de tweets de los seguidores de las empresas de telefonía celular; b) Pre–procesamiento de la data obtenida para poder identificar elementos importantes de cada tweet; c) Filtrado de elementos poco relevantes, o ruido; y d) Clasificación de cada publicación basado en las características obtenidas en etapas previas.

Los resultados obtenidos nos muestran que la introducción de un diccionario de lexicones incrementó el número de términos que pueden ser considerados para la clasificación. Así mismo, el uso de este diccionario al cual se le aumento nuevos términos permitió incrementar la tasa de clasificación en un 0,75 %. Finalmente, gracias a estas técnicas de análisis de sentimiento, es posible explotar el contenido de redes sociales de manera que puedan servir a las corporaciones para la toma de decisiones, especialmente de servicio a sus usuarios.

Palabras Clave: Análisis de Sentimiento, minería de opinión, Clasificación de mensajes de Twitter, diccionario de lexicones, hashtags

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Generalidades

1.1. Introducción

Con la llegada y posterior expansión de las redes sociales, se ha observado cómo la cantidad de datos que se generan día a día en Internet se incrementa. Para tomar un ejemplo, la red social Twitter[1], en la cual un promedio de 500 millones de tweets son publicados por día[2], donde cada tweet en promedio posee 100 caracteres[3], por lo tanto diariamente se publicaría un promedio de 5.8GB de contenido escrito cargado de información, como por ejemplo imágenes, enlaces a páginas web, menciones a otros usuarios, etc.

Los tweets en su mayoría contiene información valiosa cargada de opiniones (Aisopos et al., 2012) y estos son definidos como expresiones subjetivas que describen sentimientos, actitudes, valoraciones, ideas, juicios y creencias expresadas acerca de hechos y sus características (Karamibekr y Ghorbani, 2013).

El cuerpo de conocimiento que busca el tratamiento computacional de la opinión, sentimiento y subjetividad en un texto es conocido como "minería de opinión", "análisis de sentimiento" y/o "análisis de subjetividad" (Aisopos et al., 2012; Bosch, 2013; Li et al., 2012; Pang y Lee, 2008), también denominado como "Clasificación de Sentimiento" (Bravo-Marquez et al., 2013). Karamibekr y Ghorbani (2013) lo define como una técnica para distinguir entre información pobre e información rica en contenido.

En nuestro medio el análisis de sentimiento (AS), minería de opinión o en general la minería de datos está comenzando a dejar el ambiente académico y está empezando a introducirse a un nivel empresarial, desde emprendimientos como Simi–Labs[4] a empresas

[1] https://twitter.com

[2] https://about.twitter.com/company

[3] Según en análisis realizado por TRACK Social en el año 2012 en promedio cada tweet posee 100 caracteres: http://tracksocial.com/blog/2012/10/optimizing-twitter-engagement-part-3-tweet-length/

[4] http://www.simi-labs.com

como PeruStat[5], los cuales tienen como objetivos el realizar consultorías, investigación y análisis de datos a solicitud de empresas que lo requieran, ya sea general o específico; por ejemplo un estudio específico cómo el realizado por el INEI[6] para la encuesta ENAHO (Encuenta Nacional de Hogares) donde se realiza un análisis específico poblacional; o el realizado por Parlakuy[7] para el análisis de las elecciones presidenciales en el Perú en el 2016.

Así mismo, en el estudio realizado por Arellano Marketing llamado "Comportamiento Digital del consumidor Peruano" se identifica que el 13 % y 31 % de la muestra analizada esta *Totalmente de acuerdo* y *De acuerdo* respectivamente en quejarse mediante las redes sociales por servicios brindados por el proveedor de algún servicio. Cabe destacar que, uno de los rubros de mayor crecimiento en los últimos años en el Perú es el de telecomunicaciones; y que en el año 2014, según Osiptel[8], se dio un incremento de aproximadamente 10 % en sus servicios de Internet y telefonía, esto debido a la incursión de Operadores Móviles Virtuales (OMV), mayor uso de redes 4G e inversión pública para la conectividad a nivel nacional.

Debido a la aparición de un mayor número de operadores móviles en el Perú y la gran cantidad de opiniones que existe en las redes sociales es que el porcentaje de la población que hace uso de estas para poder demostrar algún tipo de descontento se ha incrementado. A raíz de esto es que surge el interés por identificar cómo se comporta el usuario de telefonía celular en el Perú cuando un nuevo operador de telefonía móvil ingresa al mercado, si el descontento con su actual operador móvil se incrementa o no, y cuáles podrían ser las razones. En base a ello, brindarle al operador un conjunto de puntos a mejorar; por ejemplo en el trabajo de Garza et al. (2008) se propuso como mejorar la atención al cliente, la calidad de producto y el servicio, basado en la percepción que tienen los clientes hacia la empresa.

Existen varias herramientas que apoyan en el proceso de generar analíticas para las empresas, como son: Google Analytics[9] o Salesforce Analytics Cloud[10], en las cuales se puede analizar entre varias cosas: donde ingresan o como llegan los usuarios, cuánto tiempo se quedan viendo algo, que compran o que no, etc. Pero estas solo cubren un pequeño espectro del comportamiento que expresan en Internet los clientes, olvidando el texto escrito en las diferentes redes sociales, el cual puede ser explotado gracias al análisis de sentimiento de mejor forma. En un artículo publicado en SearchSalesForce[11] se indica que el 97 % de 260 compañías habían incurrido en adquisiciones de productos de analítica de descontento

[5] http://perustat.com
[6] http://inei.gob.pe/
[7] http://parlakuy.com/
[8] https://www.osiptel.gob.pe
[9] https://analytics.google.com
[10] http://www.salesforce.com/analytics-cloud/overview/
[11] http://searchsalesforce.techtarget.com/news/4500273644/Users-discontent-with-data-insights-from-analytics-apps, consultado por última vez el 09/07/2016

del consumidor para el 2015, pero solo el 19 % de ellas reportaban satisfacción con estas herramientas.

Así mismo, en el mercado existen plataformas como Watson Personality Insight[12] el cual se promociona como una herramienta para extraer atributos de personalidad de textos provenientes de diversas fuentes, la falencia que posee esta herramienta es que solo trabaja con texto escrito en idioma Ingles; para poder utilizarlo en otros idiomas es necesario realizar la traducción previa del texto, para ello IBM presenta servicios que pueden ser añadidos en Bluemix[13] para realizar la traducción entre otras tareas más; pero, esta no realiza traducción de jerga o palabras coloquiales, lo cual limita el uso que posee, así mismo el tiempo que necesita para el entrenamiento es extremadamente alto.

En ese sentido, el problema que se buscó resolver con el presente trabajo consistió en desarrollar una metodología de análisis de microblogs que permita determinar el sentimiento de los usuarios de telefonía móvil de tres empresas del medio y de esa manera apoyar a estas sobre los servicios en los cuales se tienen mayor descontento de manera que los resultados sirvan como base para tomar acciones correctivas y/o reforzar aquellos que poseen un sentimiento positivo.

El resto de la investigación se distribuye de la siguiente forma: En el Capítulo 2 se describe el marco teórico de la investigación como aspecto metodológico del análisis de sentimiento. En el Capítulo 3 se presentan los resultados de la experimentación realizada a la data extraída de la red social Twitter bajo las condiciones indicadas previamente. Finalmente, en el Capítulo 4 se muestran las conclusiones y trabajos futuros basados en la experimentación realizada; así como los problemas y soluciones hallados en el transcurso del desarrollo de la investigación.

1.1.1. Objetivo General

Desarrollar un modelo computacional para el análisis de sentimiento de la opinión de usuarios de telefonía celular a partir de microblogs, para la identificación de términos que causan descontento de los usuarios con respecto a su empresa proveedora, frente al ingreso de un nuevo competidor en el rubro de telefonía.

1.1.2. Objetivos Específicos

1. Aplicar una estrategia de limpieza de tweets para el proceso de análisis de sentimiento.

[12] http://www.ibm.com/watson/developercloud/personality-insights.html
[13] https://console.ng.bluemix.net/catalog/

2. Definir un modelo de análisis de sentimiento para reconocer el grado de polaridad de los microblogs.

3. Identificar los términos que podrían estar relacionados al descontento de los usuarios de telefonía celular con respecto a su empresa proveedora frente al ingreso de una nueva mepresa de telefonía.

4. Analizar y describir el comportamiento del usuario de telefonía a partir de la evaluación de polaridad realizada.

1.1.3. Resultados Esperados

1. Para el Objetivo Específico 1 se plantea obtener una base de datos de tweets de 2 periodos de 1 mes cada uno, libre de términos poco relevantes, como de tweets no referidos al dominio de telecomunicaciones, para el Análisis de Sentimiento.

2. Para el Objetivo Específico 2 se plantea obtener un modelo que evalúa un microblog asignándole un grado de polaridad (positivo, negativo, neutro).

3. Para el Objetivo Específico 3 se plantea obtener una conjunto de términos negativos que permiten identificar la razón del descontento de los usuarios de telefonía celular para con su empresa.

4. Para el Objetivo Específico 4 se plantea obtener una reporte del análisis de sentimiento, respecto a la polaridad de las opiniones de los usuarios hacia su empresa de telefonía celular.

1.2. Delimitación del Proyecto

1.2.1. Alcance

El presente trabajo abarca las etapas de recolección, pre–procesamiento, filtrado, clasificación y un breve análisis de los resultados obtenidos del análisis de los microblogs de la red social Twitter de los seguidores de las empresas de telefonía celular del Perú más resaltantes. Las empresas a analizar serán Movistar Perú, Claro Perú y Entel Perú, teniendo como premisa que los seguidores a analizar serán solo aquellos que sigan solo y únicamente a una de las 3 empresas de telefonía celular comentadas previamente.

1.2.2. Riesgos

Debido a la inmensa cantidad de seguidores por empresa y la cantidad de tweets que cada uno publica, es que se debe tener en cuenta la forma de cómo se entrena el modelo seleccionado, así como los componentes del vector de características a seleccionar, los cuales deberán ser elegidas de forma que no se introduzca ruido al corpus.

1.2.3. Justificación

En Marzo del 2015 la consultora Arellano Marketing realiza un estudio acerca del comportamiento digital del consumidor peruano de la cual se logró identificar entre varias cosas que el consumidor convive con varias redes sociales para realizar diferentes tareas, entre ellas: *interactuar con otras personas, publicar o comentar información, informarse sobre noticias, informarse sobre productos y temas laborales*. Así también, se identificó que el 44 % de la muestra, si usaría una red social para realizar algún reclamo o interacción con alguna marca o empresa. Debido a las conclusiones obtenidas por parte de la investigación de Arellano Marketing y la aparición de una nueva empresa de telefonía celular en el mercado peruano es que se opta por hacer uso de técnicas de análisis de sentimiento sobre microblogs (de la red social Twitter) para poder identificar, no solo comportamiento sino también la afinidad de las usuarios de telefonía con sus respectivas empresas.

Capítulo 2

Metodología de análisis de sentimiento en microblogs

2.1. Proceso de análisis de sentimiento

Según Mejova (2012) el análisis de sentimiento tiene por objetivo extraer opiniones y emociones de textos, clasificando las emociones expresadas en estos textos a lo largo de un espectro de polaridad *positiva – neutral – negativa*, que es la forma en que otros autores clasifican los microblogs, publicaciones, comentarios o tweets[1] (Bosch, 2013; Gebreselassie y Date, 2011). También se puede realizar una clasificación más avanzada al considerar múltiples estados emocionales como *decepción, excitación, enojo, etc.* (Karamibekr y Ghorbani, 2013).

La *clasificación de polaridad de texto* es considerada una tarea típica para la cual se identifican 2 enfoques (Mejova, 2012):

- Haciendo uso de lexicones con polaridad de sentimiento conocida, pero que tendría por inconveniente que estos diccionarios son construidos para un contexto específico.
- Construyendo modelos de lenguaje usando data de entrenamiento y técnicas de aprendizaje de máquina, con el fin de construir un clasificador que también sea entrenado para un contexto específico pero que capture peculiaridades del lenguaje usado.

Esto se complementa con la arquitectura mostrada por R-Moreno et al. (2013) donde enumera 3 módulos: *un minador* que realizaría la tarea de recolección, *un clasificador* que realiza las tareas de entrenamiento, pre–procesamiento y filtrado y *un analizador* que realizaría el proceso de clasificación.

[1] Según Bermingham y Smeaton (2010) un microblog es un nuevo dominio textual representado en un texto de tamaño corto que brinda un sentimiento y que es representado mediante publicaciones en redes sociales.

Para el presente trabajo se siguió la metodología *Knowledge Discovery in Databases* (KDD) descrita por Fayyad et al. (1996) que es ampliamente aceptada por la comunidad científica del área. Esta propuesta en AS comprende cuatro etapas bien definidos y una etapa adicional: selección, pre–procesamiento, transformación y minería de datos acompañado por una etapa adicional de interpretación y evaluación. Algunos autores modifican la nomenclatura en el AS, como por ejemplo (Alhumoud et al., 2015): recolección, pre–procesamiento, filtrado, clasificación y una etapa adicional de interpretación respectivamente. La Figura 2.1 muestra un flujograma general basado en KDD para hacer una análisis de tweets, en el cual se puede apreciar que el trabajo se desarrolla en forma de cascada, cabe destacar que existe una etapa previa al proceso que en el gráfico es llamado *Keyword* en la cual lo que se realiza es una selección de palabras clave; las cuales serán palabras mediante las cuales se seleccionará los microblogs a extraer.

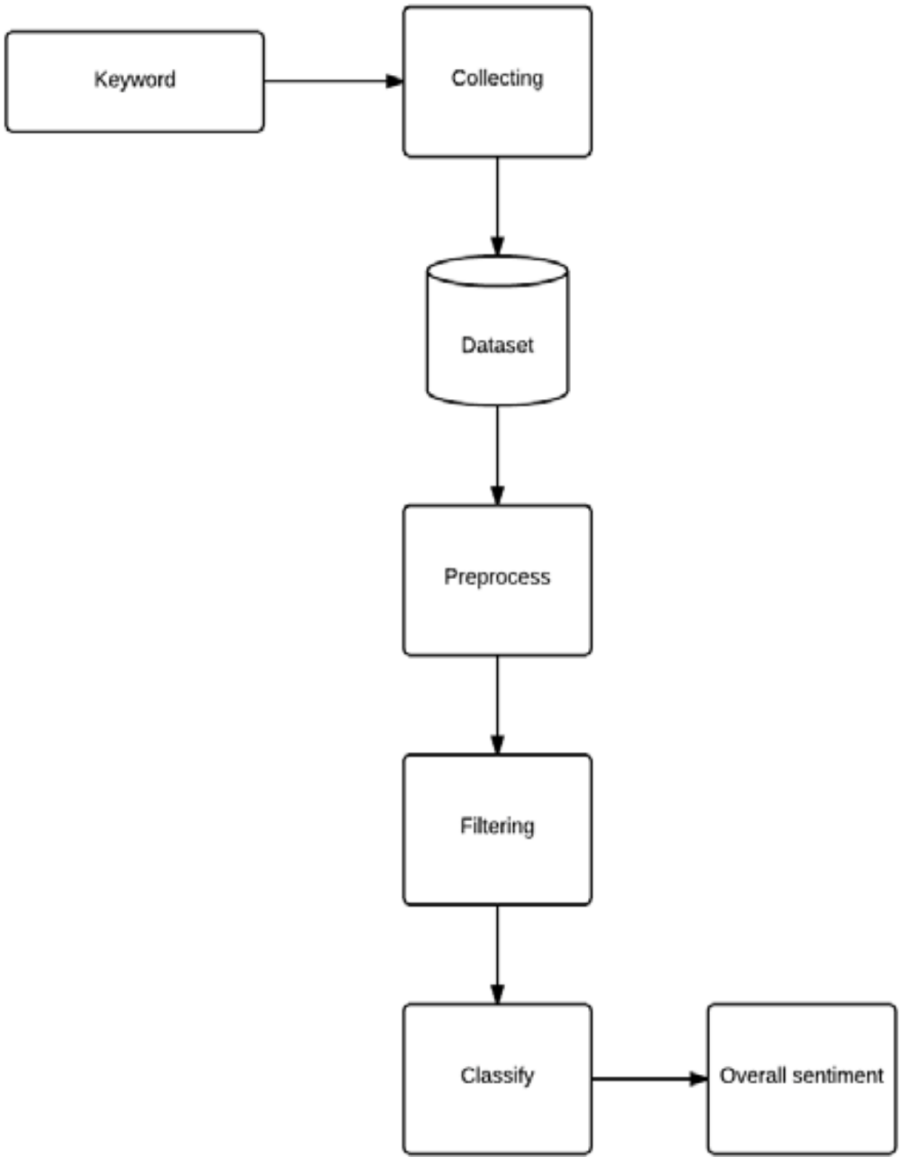

Figura 2.1 Flujograma del proceso general de análisis de sentimiento (Alhumoud et al., 2015).

A continuación se describen las técnicas más usadas para cada una de las etapas del proceso de análisis de sentimiento:

2.2. Etapas

2.2.1. Recolección

Para esta etapa se realiza la obtención de la base de datos o corpus sobre el cual actuarán las siguientes etapas. En específico para la red social Twitter se presentan 2 formas para poder recolectar tweets y están presentes en su API, el *streaming* y *REST*[2] cada uno se utiliza dependiendo de qué tipo de tweets se desee obtener, en tiempo real o de tiempo pasado almacenado en las bases de datos de Twitter respectivamente, para ello existen conjuntos de bibliotecas propias para la mayoría de lenguaje de programación para realizar esta tarea[3].

2.2.2. Pre–procesamiento

En esta etapa se presentan un número variado de técnicas, debido a que es en esta etapa donde el data extraída comenzará a transformarse en información y como Bosch (2013) lo indica: *"AS es el arte de la selección de características"*, es por ello que esta es la etapa más importante de AS. A continuación se comentan algunas de las más utilizadas:

Bag–of–words

Es la técnica que hace uso de un vector en el cual cada palabra distinta representa una posición en el vector y el contenido de cada posición puede ser representado por un valor binario para representar ausencia o presencia del término; también puede ser usado para representar el número de veces que aparece el término en un texto dado.

n–gramas

Son representaciones de *n* términos contiguos los cuales en conjunto encierran una idea. Mejova (2012) indica que el rendimiento de los *n–gramas* no depende de si se usa un *n* igual a 1, 2 o 3, pero explica que esto si presenta variaciones en el rendimiento si se acompaña con la utilización de presencia o frecuencia de términos. Esta técnica es parte de la identificación del segundo tipo de subjetividad[4] comentado por Wiebe et al. (2004) llamado *colocación* y de forma interesante estos n–gramas poseen un conjunto de palabras llamados *stopwords*.

[2]https://dev.twitter.com/streaming/overview
[3]Por ejemplo twitteR para R: http://cran.r-project.org/web/packages/twitteR/index.html
[4]Los 3 tipos de subjetividad son: *hapax legomena*, *colocación* y *adjetivos y verbos*.

Part–of–speech

Tiene por objetivo reconocer la categoría gramatical a la cual una palabra pertenece (Boldrini y Balahur, 2010). Así mismo gracias a esta técnica se puede identificar la subjetividad de un documento (Bosch, 2013). Autores como Wiebe et al. (2004) definen al producto del uso de POS como el tercer tipo de subjetividad, ya que los *verbos*, *adverbios* y *adjetivos* son buenos indicadores de subjetividad. Adicionalmente mediante esta técnica se puede reconocer intensificadores, abreviaciones y emoticones, los cuales según experimentos llegan a introducir ruido (Bravo-Marquez et al., 2013).

Palabras enriquecidas con negación

Por la cual la significancia de una palabra cambia cuando algún término de negación se encuentra alrededor suyo (Amiri y Chua, 2012; Mejova, 2012). Aunque para algunos autores comoMejova (2012), señalan que el uso de esta representación de texto no trae buenos resultados ya que introduce ruido. Para prevenir la introducción de ruido por el uso de negación de términos se puede generar reglas para casos específicos, como por ejemplo Amiri y Chua (2012) usó 36 palabras negadas en un conjunto de datos de preguntas–respuestas de Yahoo![5]. Por su parte Councill et al. (2010) se enfocó en poder identificar de mejor manera la negación en documentos mediante la generación de un lexicón *ad–hoc* y la identificación del ámbito de la negación con resultados que lograron mejorar sustancialmente la identificación de polaridad positiva y negativa.

Utilización de diccionarios de lexicones

Los cuales son un conjunto de palabras las cuales ya poseen una polaridad de sentimiento y que poseen la cualidad de que con ellas se puede clasificar textos que las contengan, por ejemplo pudiéndose hacer uso de un diccionario aumentado con sinónimos, antónimos, hipónimos además de lexicones no estándar como jergas o palabras mal deletreadas (Amiri y Chua, 2012). Mejova (2012) analizó diversas técnicas y enfoques en AS entre las cuales le presto mayor importancia a la representación matemática de la interacción entre personas denominada **Affect Control Theory (ACT)**, específicamente a la generación de un diccionario de lexicones mediante el enfoque **Evaluation, Potency & Activity (EPA)**, con lo cual se adquiere 3 dimensiones de polaridad para representar las palabras existentes en el diccionario. Así mismo, compara este diccionario con el diccionario SentiWordNet[6], el cual

[5] Yahoo! Webscope Dataset: http://webscope.sandbox.yahoo.com/
[6] http://sentiwordnet.isti.cnr.it

es una extensión de WordNet[7] al cual se le adicionó el factor de sentimiento y la polaridad a las palabras, encontrando que para el caso de estudio político brindaba mejores resultados el diccionario basado en el enfoque EPA. Acerca de la utilización de diccionarios de lexicones basados en el enfoque EPA también se puede usar el algoritmo de *Label Propagation* para expandir los puntajes de las palabras que se encuentran en un diccionario y llevarlos hacia palabras etiquetadas como *actor, verbo* u *objeto* (Hoey, 2015). Para este tipo de diccionarios se generó un diccionario estándar para el enfoque EPA base el cual sirvió para el proyecto INTERACT[8] el cual contiene palabras en varios idiomas y que poseen las 3 dimensiones EPA (Hoey y Schr, 2015).

Presencia de términos

También llamado **pesos binarios** o **frecuencia de términos** en la etapa de generación del vector de características. Cuando se hace uso de presencia de términos se elimina las palabras que aparecen raramente debido a que podrían ser palabras con errores de escritura (Mejova, 2012). Así también, dependiendo del investigador se puede o no eliminar palabras que aparecen una sola vez ya que estas tienen gran precisión de subjetividad y son definidas por Wiebe et al. (2004) como *hapax legomena* [9].

2.2.3. Filtrado

En la etapa de filtrado se realiza la limpieza de data, se puede enumerar algunas actividades para esta etapa como son: *Eliminación de letras repetidas*, *Eliminación de stop–words*[10] o palabras que no brindan significancia al texto y *Normalización* o intercambio de elementos no utilizables por otros que presten mayor ayuda a las siguientes etapas (Alhumoud et al., 2015). Los filtros más comúnmente utilizados cuando se realiza esta etapa es la eliminación de *menciones* – @, *hashtags* – #, *URLs* y *signos de puntuación* (R-Moreno et al., 2013); en general se utilizan expresiones regulares para encontrar patrones que contengan los filtros anteriores.

[7] https://wordnet.princeton.edu
[8] http://www.indiana.edu/s̃ocpsy/ACT/interact.htm
[9] Es una palabra que ha aparecido registrada solamente una vez en un idioma dado.
[10] Una lista puede ser encontrada en: http://www.ranks.nl/stopwords

2.2.4. Clasificación

Técnicas de aprendizaje de máquina

En esta etapa se hace uso de técnicas de aprendizaje de máquina, entre las más usuales se encuentran:

Pointwise Mutual Information (PMI) Es un método de análisis de orientación semántica que no requiere de un corpus extenso para el proceso de entrenamiento y se basa en poder indicar cuan cercana es una frase a un conjunto de palabras *paradigma*, en el cual fue definido el sentimiento que se desea capturar (Bosch, 2013). También representa el grado de dependencia estadística entre 2 términos cuyo indicador es la Orientación del sentimiento de una frase p ($SO(p)$) (Liu, 2012).

Los puntajes que se obtienen posterior al uso de PMI, son generalmente convertidos a características binarias para posteriormente ser parámetros de entrada para un clasificador Naïve Bayes. Este método es comúnmente utilizado haciendo uso de un motor de búsqueda en donde se hace una consulta de la frase p y se determina el número de co–ocurrencias contra las palabras paradigmas (*hits*) (Mejova, 2009).

Naïve bayes El clasificador Naïve Bayes considera que cada una de las características de un elemento contribuye independientemente a la probabilidad de asegurar que un elemento sea de la clase a la cual se le asigna (Bosch, 2013).

Support Vector Machine (SVM) Se define en la generación de hiper–planos capaces de asignar correctamente los documentos a una clase c, recordando que se usa el término *hiper–plano* y no *plano* por el número de dimensiones de las características que se desea considerar. Para ello se calcula la derivada de la fórmula del Lagrangiano Dual para hallar los valores α que son necesarios para calcular los hiper–planos (Bosch, 2013). Bravo-Marquez et al. (2013) señala que se obtienen mejores resultados en el proceso de clasificación haciendo uso de SVM en comparación con el uso de Naïve Bayes.

Pre-entrenamiento de clasificadores

Cuando se tiene información con emoticones, se desarrolla una labor previa denominada de pre–entrenamiento, la cual consiste en entrenar al modelo haciendo uso de los emoticones, asignando la polaridad de cada tweet basado en la existencia de emoticones positivos y negativos. En algunos casos se realiza un entrenamiento manual, como por ejemplo Bravo-Marquez et al. (2013); Hoey (2015); R-Moreno et al. (2013) desarrollaron una interfaz donde

se identifica y asigna características además de clasificar un determinado conjunto de tweets o documentos de forma manual.

En otras ocasiones el proceso de pre–entrenamiento se hace de forma automática mediante la utilización de conjuntos de datos pre–etiquetados. El inconveniente de este enfoque yace en que la mayoría de estos conjuntos de datos son específicos de un contexto dado, en otros casos lo que se suele hacer es generar un conjunto de datos específico para lo que se desea analizar mediante la utilización de otras herramientas. Por ejemplo, Kiritchenko et al. (2014) y Lim (2014) usaron la base de datos *Sentiment140 Base Lexicon* la cual contiene un conjunto de tweets etiquetados mediante emoticones y fue generada a partir del *Sentiment140 Corpus* (Go et al., 2009).

Por otro lado, Bravo-Marquez et al. (2013) descarta el uso de emoticones, debido a que introduce ruido y tampoco recomienda el uso de Sentiment140 porque este corpus no está enfocado a la clasificación de subjetividad, la cual como Pang y Lee (2008) indica: la clasificación de subjetividad va más allá de solo clasificar un texto como positivo o negativo.

Dificultad de clasificación en análisis de sentimiento

El análisis de sentimiento es una tarea más complicada que la clasificación de texto tradicional debido a varios factores, como son (Aisopos et al., 2012; Gebreselassie y Date, 2011):

- *Escasez*, debido a que los textos comúnmente usados son cortos.
- *Vocabulario no estándar*, los textos al ser creados en redes sociales presentan alto porcentaje de jergas y contracciones gramaticales propias del individuo.
- *Ruido* debido a la presencia de errores gramáticas, incomprensibilidad de contenido, etc.
- *Multi–idioma*, textos en el idioma materno de quien escribe con ciertas palabras o frases en un idioma foráneo.
- *Presencia de sarcasmo*, entre otros.

Así mismo, los textos por si solos carecen de subjetividad en su totalidad, pero hoy en día los textos poseen un contenido adicional como son URL, vídeos e imágenes, etc. los cuales son comúnmente son filtrados en la etapa previa (Bosch, 2013). Este contenido adicional guarda subjetividad en especial los emoticones y *hashtags* que como se comentó previamente suele ser ignorado (Bukhari et al., 2014).

Por su parte Ghenname et al. (2014) trabajó en la obtención de subjetividad de los *hashtags* con su enfoque *folksionary* donde se genera una matriz de similaridad que contiene como componentes a palabras y *hashtags* y como entradas las distancias entre los pares y para la generación del diccionario utilizó Markov Clustering Algorithm (MCA) el cual interpretaba las entradas desde la matriz previamente comentada como similaridades y simulaba una caminata aleatoria (random walk) para cambiar las probabilidades de transición en la matriz de adyacencia.

2.3. Trabajos Relacionados

A lo largo de la presente investigación se observó diversos trabajos los cuales se asemejaban al desarrollado en este, la mayoría de ellos en otros ámbitos diferentes a las telecomunicaciones pero que el AS se ve inmerso y ayuda a solucionar problemáticas similares a la investigada.

Así como se comentó inicialmente AS podría ser utilizado en muchos campos de la ciencia, entre ellos para predecir el futuro como lo indican Bosch (2013); R-Moreno et al. (2013). En este aspecto Bouillot et al. (2012) utilizó un AS orientado a la opinión política y los resultados obtenidos no diferían mucho de la realidad ya que se indicaba con un pequeño margen de error al ganador de las elecciones presidenciales de Francia del 2012; así también se indicó en otro trabajo similar que existe una dificultad en la clasificación de sentimiento político, obligando a hacer una re–definición del "sentimiento" y concluyéndose que debe existir interacción entre el *framework* desarrollado y el ser humano (Mejova, 2012).

Boldrini y Balahur (2010) realizó un estudio usando técnicas de pre–procesamiento, y un clasificador SVM del WEKA[11] sobre un conjunto de contenido obtenido de diferentes blogs y diferentes contenidos; como son: *Calentamiento Global* y *Política* en Zimbabwe y USA, etiquetando bi, tri y tetra–gramas con anotaciones de opinión, intensidad y emoción basados en el enfoque propuesto llamado *Emotiblog* al cual se le asignaba diferentes pesos para posteriores evaluaciones.

Así también, Li et al. (2012) propuso un marco de trabajo denominado *Topic-Level Opinion Influence Model* en el cual los históricos del usuario y sus registros de interacción social son el punto de apoyo de este modelo para poder construir un histórico e influencia de opinión de los vecinos a través de procesos de aprendizaje estadístico que son utilizados para predecir futuras opiniones del usuario sobre un tópico específico, algo similar como lo propuesto por Hoey y Schr (2015) donde mediante el uso de ACT usando procesos de decisión de Markov y mediante un enfoque Bayesiano se puede intentar descubrir la identidad

[11] http://www.cs.waikato.ac.nz/ml/weka/

afectiva de las personas y cómo interactúan así como predecir futuros sentimientos mediante la *deflexión*.

En referencia a los enfoques de clasificación dependiente del objetivo, se identificó que el enfoque tradicional posee falencias ya que los tweets por naturaleza son ambiguos, por ello es que el enfoque por objetivos trabaja con la incorporación de características dependientes del objetivo y también la capacidad de considerar tweets relacionados. Para lo mencionado fue diseñado un proceso en 3 etapas que se basaba en clasificación de subjetividad, clasificación de polaridad acerca del objetivo (dependiente del paso anterior) y optimización basada en grafos para mejorar el rendimiento tomando en cuenta tweets relacionados al que se está analizando (Jiang et al., 2011).

El AS no solo se realiza para textos en idioma ingles sino también es usado en otros idiomas como el Árabe, para textos de diferentes conjunto de caracteres se presta importancia a la sub–etapa de Normalización en la etapa de Filtrado. De igual forma se revisa la posibilidad de utilizar emojis (emotion faces), diccionarios y clasificadores no supervisados (Alhumoud et al., 2015).

Cuando se habla de metodologías también se presta importancia al contexto, ya que se indica que cuando un análisis es realizado dependiente del contexto, entonces si se le aplica a un estudio, campo o contexto diferente entonces no brinda los resultados esperados. En el caso de una metodología no dependiente del contexto para minería de opinión Bosch (2013) realizó un caso de estudio donde se extrajo tweets de productos comerciales para identificar la confiabilidad de las mismas y así generar una clasificación de opinión social. Se pre–entrenó el modelo con emoticones, concluyendo que estos no brindaban un espectro completo de opinión en tweets y no era adecuado, como también se confirmó en el trabajo de Bravo-Marquez et al. (2013).

Para evitar el problema de la cobertura del espectro de opinión que solamente con los emoticones no se logra cubrir lo más común es realizar un pre–entrenamiento manual (R-Moreno et al., 2013) y posteriormente usar un diccionario de lexicones

Capítulo 3

Determinación del sentimiento en microblogs

En el presente capítulo se muestra la aplicación de diversas técnicas de AS, tomando como caso de estudio el análisis del comportamiento del consumidor de telefonía celular peruano, para ello se utilizó algunas técnicas de minería de textos como TF–IDF, entre otras como los presentados por Bouillot et al. (2012) cuando realizo un análisis de los tweets de los seguidores de los candidatos presidenciales para las elecciones de Francia 2012, y metodología de Li et al. (2012) ya que realiza un análisis de microblogs de microblogging Tencent Weibo.

3.1. Fuente de información sobre opiniones de usuarios de telefonía celular

Para el análisis del comportamiento se decidió realizar un análisis temporal de las opiniones de los usuarios de 3 de las empresas de telefonía móvil en el Perú más representativas, estas son: *Claro Perú*[1], *Movistar Perú*[2] y *Entel Perú*[3]. Para este análisis se decidió tomar los tweets del *25 de Junio de 2014 al 25 de Julio de 2014* y adicionalmente del *13 octubre de 2014 al 13 de noviembre 2014*, la razón de la selección de estos rangos de tiempo se debe a que: el primer lapso de tiempo representa el primer mes de actividad en la red social Twitter de la empresa *Entel Perú*, aún en ese entonces llamada *Nextel Perú*, mientras que para el segundo lapso de tiempo, fue donde esta empresa pasó por un proceso de cambio de razón

[1] http://www.claro.com.pe
[2] http://www.movistar.com.pe
[3] http://www.entel.pe

social, el cual hizo que el nombre de la empresa cambiara y así como también, su forma de trabajo, cambiando así también la percepción de los usuarios, por ello el segundo lapso de tiempo representa el primer mes de actividad en el mundo real de la empresa *Entel Perú*.

El proceso de análisis de opiniones de los usuarios de telefonía móvil estuvo basado en las etapas de KDD que, de ahora en adelante se denominará *seguidores* a los clientes de las empresas, esto debido a que para la red social Twitter los clientes o usuarios se les denomina *seguidores* y para no introducir ruido de los usuarios que sean seguidores de más de una empresa de telefonía en Twitter se decidió limitar el análisis solamente a aquellos que sigan solo a una única empresa.

A continuación se comenta las tareas y resultados obtenidos en las etapas del proceso de AS:

3.2. Recolección de información de microblogs

Para la etapa de recolección se utilizó el *Application Programming Interface* (API) REST de Twitter para extraer los *identificadores* de las 3 empresas de telefonía celular en esta red social, al tener estos identificadores mediante el API se extrajo los identificadores de todos sus seguidores para tenerlos mapeados como sus clientes. En la Tabla 3.1 se puede observar el número de seguidores que poseían cada una de las empresas para Mayo del año 2015. En la fila *Total* se observa el número total de seguidores sin importar si seguían a una o más empresas al mismo tiempo y en la fila *Limitados por empresa* se observa el porcentaje del Total de usuarios que solo son seguidores de su respectiva empresa (se le denominará seguidores únicos), esto se realiza para disminuir el ruido de la base de datos.

Tabla 3.1 Seguidores en Twitter de las Empresas de Telefonía Celular

Seguidores	**Claro**	**Movistar**	**Entel**
Total	383083	388362	34325
Limitados por empresa	49,08 %	55,22 %	29,56 %

Producto de extraer los identificadores de los seguidores únicos se procedió a recolectar los tweets de estos seguidores que hubieran sido escritos en la red social desde el 25 de Junio del 2014 hasta la fecha en que se realizaba el proceso de minado con la finalidad de poder tener una amplia base de datos que pudiera servir para futuras comparaciones, producto de ello 26,917,539 tweets fueron recuperados. En la Tabla 3.2 se muestra el número de tweets que fueron extraídos de los seguidores de las 3 empresas operadoras para los lapsos de tiempo correspondientes a:

- **2014-06-25 00:00:00** a **2014-07-25 23:59:59** → 1 mes posterior de la creación del usuario @*Entelperu* en Twitter.
- **2014-10-13 00:00:00** a **2014-11-13 23:59:59** → 1 mes posterior al re–branding de Nextel Perú a Entel Perú.

Tabla 3.2 Tweets extraídos para el análisis de sentimiento

	Tweets
Lapso de tiempo 1	1,746,948
Lapso de tiempo 2	2,753,664

3.3. Generación del espacio de características

3.3.1. Pre–procesamiento

Para la generación del espacio de características se decidió omitir ciertos filtros que comúnmente se hacen, por ejemplo en el caso de los emoticones se trabajó de forma similar al trabajo realizado por Lim (2014) en el cual se utilizaba grupos de emoticones como indicadores de emoción, en nuestro caso se realizó un etiquetado y reemplazo de los emoticones por su significado textual. En la Tabla 3.3 se muestra un ejemplo de lo comentado, en donde se etiquetó un total 191 emoticones, mostrándose que de todos los tweets un 41.68 % de ellos contenían emoticones los cuales hubieran sido eliminados, pero al realizar esta tarea se logra reforzar el tweet con el significado textual de los emoticones.

Tabla 3.3 Emoticones

Emoticones							**Traducción**
:)	:D	:o)	:]	:c)	=]	8)	feliz
:(	:-c	:c	:<	:-[	:[	:{	triste
:-O	:O	:-o	:o	8-0	O_O		sorprendido

Realizada esta primera tarea se procedió a utilizar esta nueva base de datos como input en la generación de un diccionario de lexicones. Para el mismo se decidió realizar una comparativa entre *SentiWordNet*, con un puntaje tanto para polaridad positiva como negativa, y el diccionario de lexicones basado en el enfoque *EPA*, con sus 3 dimensiones de polaridad,

para ello se utilizó como diccionarios iniciales el provisto por el proyecto SentiWordNet[4] y para el segundo diccionario el brindado por el proyecto INTERACT[5].

Teniendo estos diccionarios iniciales se hizo uso del algoritmo de *Label Propagation* para poder ampliar estos diccionarios (Hoey, 2015). Inicialmente se realizó la adición de los *hashtags*, para lo cual se trabajó de una forma similar a lo propuesto por Ghenname et al. (2014) hasta el punto donde se genera una matriz de adyacencia de términos y *hashtags* en la cual las entradas de la matriz son los pesos, para nuestro enfoque es el número de documentos en los cuales el par ordenado de cada entrada (par de palabras) aparecen juntos y posteriormente mediante *Label Propagation* expandir las etiquetas de las palabras conocidas a las desconocidas, entre ellas los *hashtags*.

En la Figura 3.1 se muestra una nube de *tags* que contiene los *hashtags* más utilizados por los seguidores de las empresas de telefonía celular, cuyo tamaño en la nube indica la frecuencia de uso. Por otro lado, la Figura 3.2 se muestra un grafo cuyas aristas representan la existencia de una relación entre palabras esto se realiza para poder visualizar la conectividad entre *hashtags* y palabras que si poseen puntajes en los diccionarios utilizados.

Esta nube y grafo, fueron generados con la frecuencia de palabras resultantes del proceso de eliminación de términos, que cumplían con poseer un índice de dispersión de 0,99. El índice de dispersión en matrices de términos indica el porcentaje de ausencia de un término en un corpus de documentos, por ejemplo si se tienen 1000 documentos con un índice de dispersión de 0,99 indica que si el término no está en al menos 10 documentos, el término será eliminado. Para la visualización se utiliza un índice de 0,99 a modo de mostrar más términos pero en la investigación se utiliza un índice de 0,999 y así tener una mayor precisión en la asignación de los puntajes de polaridad a las palabras desconocidas.

[4]http://sentiwordnet.isti.cnr.it
[5]http://www.indiana.edu/ socpsy/ACT/data.html

Figura 3.1 Nube de *hashtags* más utilizados en tweets de seguidores de empresas de telefonía

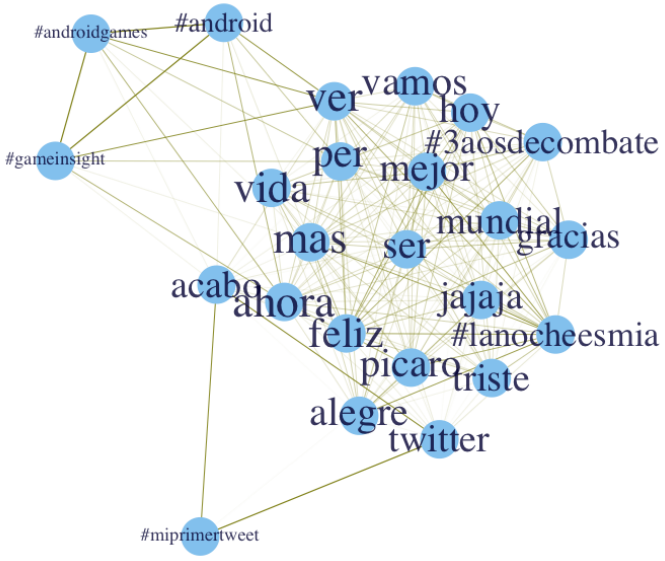

Figura 3.2 Grafo de relaciones entre palabras y *hashtags* de seguidores de empresas de telefonía

Posterior a la ejecución del algoritmo *Label Propagation* sobre el corpus de solo los tweets que poseían hasthtags se llegó a la generación de los diccionarios aumentados mediante este algoritmo. En la Tabla 3.4 se observa una muestra de los puntajes asignados a los *hashtags* tomando como diccionario base al diccionario del enfoque EPA, así como también, en la Tabla 3.5 se observa la muestra de los puntajes obtenidos tomando como base el diccionario SentiWordNet.

Tabla 3.4 Pesos asignados mediante *Label Propagation* sobre el diccionario EPA

Término	Evaluation	Potency	Activity
#100happydays	-3.75	-3.75	-3.8
#3anosdecombate	-2.61	-2.62	-2.77
#3aosdecombate	4.3	4.3	4.3
#3arelegendapp	-3.8	-3.8	-3.84
#4glteclaro	-3.77	-3.77	-3.82
#4yearsof1d	-2.49	-2.5	-2.66
#4yearsofonedirection	-1.27	-1.29	-1.56
#90depasin	-3.79	-3.79	-3.84
#90fl	-3.7	-3.7	-3.75

Tabla 3.5 Pesos asignados mediante *Label Propagation* sobre el diccionario SentiWordNet

Término	Positivo	Negativo
#100happydays	0.06	0.01
#3anosdecombate	0.17	0.03
#3aosdecombate	1.62	0.33
#3arelegendapp	0.05	0.01
#4glteclaro	0.05	0.01
#4yearsof1d	0.18	0.04
#4yearsofonedirection	0.29	0.06
#90depasin	0.05	0.01
#90fl	0.06	0.01

Acerca del diccionario SentiWordNet aumentado se puede observar que los puntajes positivos obtenidos con SentiWordNet predominan sobre los negativos en la muestra, al contrario de lo observado en el diccionario EPA aumentado donde se puede notar que hay más variedad en la asignación de los puntajes. Para verificar cual poseía un mejor acercamiento se procedió a comparar de forma manual una muestra de *hashtags* en los tweets, como por ejemplo los tweets que tengan en su texto el *hashtag #100happydays* el cual se puede entender como un término con puntaje de polaridad positivo siendo corroborado por los siguientes tweets obtenidos de la extracción de los seguidores de las empresas sin realizar ningún filtrado:

- Feel like cyrus #tongue #girl #petite #snapchat #girly #100happydays #day40 #hello #lover #sunrise

- Llegaron mis regalitos #100happydays #natura #ciclo9 #bienestarbien
- Con mis gorditas lindas #100happydays #bf
- Para endulzar la tarde y ser feliz #100happydays

3.4. Identificación de características

A continuación se realizó la identificación de características las cuales permitirían clasificar los tweets con la polaridad adecuada, para ello se seleccionó las siguientes características:

- Matriz TF–IDF (Term Frecuency – Inverse Document Frecuency), es una medida estadística la cual ayuda a evaluar cuán importante es una palabra a un documento en una colección o corpus.[6]
- Matriz de conteo de categorías gramaticales mediante POS (Part–of–Speech), en cada documento se realiza un conteo de cada categoría gramatical existente y transportándola a una matriz.
- Uso de polaridades de términos, para estas características son identificadas en base al contexto del tweet (positivo o negativo) el cual es asignado así por la presencia o ausencia de negaciones en el tweet[7] y mediante ello se identificó 4 características, los cuales son:
  - Suma de puntajes de términos positivos en contextos positivos (TPCP).
  - Suma de puntajes de términos negativos en contextos positivos (TNCP).
  - Suma de puntajes de términos positivos en contextos negativos (TPCN).
  - Suma de puntajes de términos negativos en contextos negativos (TNCN).

En la Tabla 3.6 se muestra un ejemplo del vector de características generado sobre el tweet:

> No hay nada como recibir un mensaje cargado de tanta buena onda en una fecha como esta #PeruTeLlama @MovistarPeru http://t.co/3rSCbUVlrl

[6] http://www.tfidf.com
[7] Se considera términos de negación a: nunca, jamás, no, nada, ninguno, ninguna

Tabla 3.6 Ejemplo de vector de características

Característica	Atributos				
Matriz TF-IDF	@movistarperu	buena	fecha	mensaje	tan
	8.517867	0 7.233166	9.947312	6.856755	9.644464
Matriz de conteo de categorías gramaticales	Adjetivo	Sustantivo	Adverbio		
	1	3	1		
Suma de polaridades basada en polaridad	TPCP	TNCP	TPCN	TNCN	
	0	0.62	0	0	

3.5. Clasificación de tweets

Para la clasificación se utilizó un 70 % del conjunto del total de tweets de la primera etapa para la fase de entrenamiento, los cuales fueron previamente clasificados de forma manual. Posteriormente se utilizó el algoritmo de SVM para el entrenamiento de este conjunto de datos, para ello se utilizó 4 diferentes tipos de kernel para poder identificar cual era mejor para el modelo. En la Tabla 3.7 se observa los resultados de la clasificación.

Tabla 3.7 Tabla de discrepancias entre asignación de sentimiento con SVM (kernel varios) y la asignación manual

			Tweets con diccionario SWN			Tweets con diccionario EPA		
			neutral	**negativo**	**positivo**	**neutral**	**negativo**	**positivo**
Calificación con SVM y kernel:	**lin**	**neutral**	7108	1514	1493	7124	1462	1499
		negativo	314	1770	321	290	1808	289
		positivo	340	387	6404	348	401	6430
	pol	**neutral**	7429	2616	2656	7534	2677	3450
		negativo	7	158	25	21	290	62
		positivo	326	897	5537	207	704	4706
	rad	**neutral**	7149	1477	1416	7152	1335	1405
		negativo	286	1786	321	285	1933	334
		positivo	327	408	6481	325	403	6479
	sig	**neutral**	7118	1563	1480	7120	1481	1502
		negativo	283	1687	311	281	1758	313
		positivo	361	421	6427	361	432	6403

La Tabla 3.7 representa una tabla de discrepancia donde se indica cuantos de los tweets calificados mediante el algoritmo SVM y el kernel señalado están calificados de forma correcta, por ejemplo los hubo 6481 tweets correctamente calificados como positivos mediante el uso del algoritmo SVM con kernel radial y $1416 + 321$ mal calificados como neutrales o negativos respectivamente. Se remarcan con color amarillo aquellos que tuvieron mayor número de aciertos que el resto, lográndose identificar que el kernel radial da mejores resultados y que con el diccionario EPA se logran mejores resultados en la clasificación especialmente con los tweets negativos.

Después de haber analizado los resultados obtenidos con el conjunto de datos de prueba se observó que con el uso de SentiWordNet se obtuvo una exactitud en general del 78.45 % y con el diccionario EPA un 79.20 %, posterior a este análisis se procedió a realizar una predicción del sentimiento de los tweets del segundo bloque seleccionado de los tweets de los seguidores de las empresas de telefonía celular (los tweets extraídos con límites 13 octubre de 2014 al 13 de noviembre 2014). De este análisis se obtuvo una lista con los tweets negativos, de los cuales se muestra un sub conjunto en la Tabla 3.8.

Tabla 3.8 Muestra de tweets negativos obtenidos en el segundo bloque extraído

Tweets con polaridad Negativa
@soportemovistar y encima me llaman y me tratan mal!!! cc. @OSIPTEL
@soportemovistar es el colmo la atencin que brindan. Tengo un reporte de hace ms de 24 horas y vienen a mi casa sin avisarme. @OSIPTEL
@ClaroPeru porque no esta funcionando la atencin al cliente por va telefonica??? @ClaroPeru por favor se los pido, den un buen servicio. Su servicio es psimo. Me arruina el da cada da.
@MovistarPeru minimo con el pesimo servicio que dan.
RT @lucvarsar: @ClaroPeru pesimo servicio. El 4g solo funciona en 3 distritos y no te lo dicen. 35 llamadas al 123 y no solucionan nada.#e

En la Figura 3.3 se muestra las primeras 100 palabras negativas más frecuentes donde se observa los términos causantes de la generación de tweets negativos. Se evidencia que algunos de los términos los cuales los seguidores utilizan para quejarse *claroperu*, *movistar*, *soportemovistar*, *internet*, etc., siendo de este gráfico donde se puede identificar aquellos servicios los cuales ocasionan incomodidad entre los seguidores de las empresas de telefonía celular; estando entre ellos lo relacionado a *seguridad*, *internet* y *servicio* entre otros.

Figura 3.3 Nube de *tags* de tweets negativos del segundo bloque extraído

Para poder complementar lo visto en la nube de palabras negativas se realizó un análisis sobre el porcentaje de tweets de polaridad negativa que afectan a los servicios de cada empresa de telefonía, la misma se muestra en la Tabla 3.9, en la cual se puede observar que el mayor descontento se da directamente hacia la empresa propiamente dicha (Entel, Claro y Movistar), como es el trato y la forma de expresar su imagen. Adicionalmente se muestra que entre los servicios que se brindan, el mayor descontento se da entre los servicios de Internet, seguridad y los equipos en sus diferentes planes; dejando de lado los servicios tradicionales, como: servicio de llamadas, mensajería y otros servicios tradicionales los cuales al no se nombran individualmente debido a su bajo porcentaje de presencia en los tweets, por ello se incluyeron en la categoría "Servicio en general".

Tabla 3.9 Porcentaje de descontento entre empresa de telefonía y servicios

Empresa VS Servicio	**Empresa**	**Servicios en General**	**Equipo celular**	**Seguridad**	**Servicios adicionales**	**Internet**
Entel	52.72	10.32	13.04	10.33	8.70	4.89
Claro	43.29	13.34	16.25	11.88	9.40	5.83
Movistar	40.21	15.38	15.97	9.01	8.36	11.06

En la Tabla 3.10 se muestra el porcentaje de sentimiento en los tweets del segundo bloque de tweets extraídos, observándose que el 60.88 % de tweets son de carácter neutral, tipo noticias o mensajes auto generados por aplicaciones como YouTube[8], como se muestra en la Tabla 3.11.

[8]http://youtube.com/

Tabla 3.10 Distribución de Sentimiento en Tweets en el segundo bloque extraído

Tweets totales	192786
Tweets positivos	29.97 %
Tweets negativos	9.15 %
Tweets neutrales	60.88 %

Tabla 3.11 Muestra de tweets neutrales obtenidos en el segundo bloque extraído

Tweets con polaridad neutral
RT @telefonicaperu: #Movistar present novedades en “La Zona del Saber” con ’V-Learning’, curso de ingls online para clientes Movistar. ht
RT @MovistarPeru: #Samsung lanzara Alpha su nueva gama de smartphones! Tendra carcasa metlica para competir con el #iPhone6 http://t.co
Mi diario est disponible! http://t.co/uQRzPK9As3 Gracias a @entel_empresas
Cmo influye internet en tu vida? http://t.co/zpGShwffks
Me ha gustado un vdeo de @YouTube de @drawcidix (http://t.co/L1Qh7gPnxi - Mejorar la velocidad de tu internet l 3G y WiFi l Zte
Vuelta a clase, la nueva promocin de Apple con descuentos y tarjetas regalo http://t.co/dVkJ00jmyh
ltimos das para acceder al precio de promocin, slo hasta el 01 de Julio! http://t.co/9tXkCa9L3u

3.5.1. Modulación de los parámetros del clasificador

Para optimizar el algoritmo SVM se hizo variar las constantes de regularización (*cost*) del algoritmo desde una constante igual a 1 hasta 100, obteniéndose lo mostrado en la Figura 3.4, donde se observa lo siguiente:

- Para los tweets neutrales se logró el mayor número de tweets correctamente calificados cuando la constante de regularización es igual a 1.
- Para los tweets negativos se logró el mayor número de tweets correctamente calificados cuando la constante de regularización es igual a 34.
- Para los tweets positivos se logró el mayor número de tweets correctamente calificados cuando la constante de regularización es igual a 14.

- Sumarizando todos los tweets calificados, se observó que se logra la mayor cantidad de tweets correctamente calificados cuando la constante de regularización es igual a 14, lográndose un 79.26 % de exactitud en la calificación con el diccionario EPA. Como se observa en la Figura 3.4 si se hubiera colocado la constante de regularización en 1 o mayor a 60 la cantidad en tweets correctamente calificados hubiera disminuido notablemente.

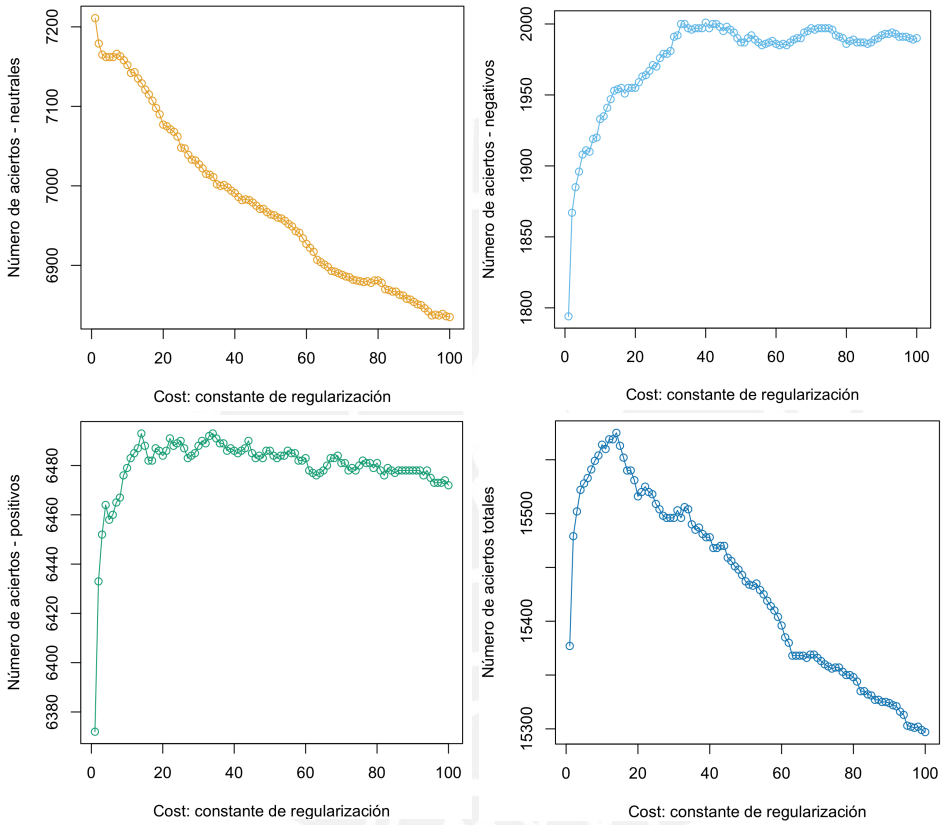

Figura 3.4 Gráficos de líneas entre los tweets aceptados y la constante de regularización

3.6. Análisis del comportamiento de los usuarios

Así mismo se observó cual era el comportamiento de los usuarios en general con respecto a sus tweets, para ello en la Figura 3.5 se generó un diagrama de dispersión de las polaridades de los usuarios, los cuales fueron calificados como usuarios positivos, negativos o neutrales dependiendo de si tenían más tweets calificados como positivos, negativos o neutros.

En la leyenda de la Figura 3.5 se identifica la polaridad con la que los usuarios fueron calificados, así mismo se muestra que existen usuarios que publican mucho y de forma positiva incluso solo tweets con esta polaridad, por ejemplo el seguidor más positivo posee

276 tweets positivos, 0 tweets negativos y 2 tweets neutrales. Así también, hay usuarios que realizan la acción contraria como por ejemplo el seguidor más negativo posee 20 tweets positivos, 59 tweets negativos y 51 tweets neutrales, como se observa en la figura la mayor cantidad de usuarios publica poco y de forma variada.

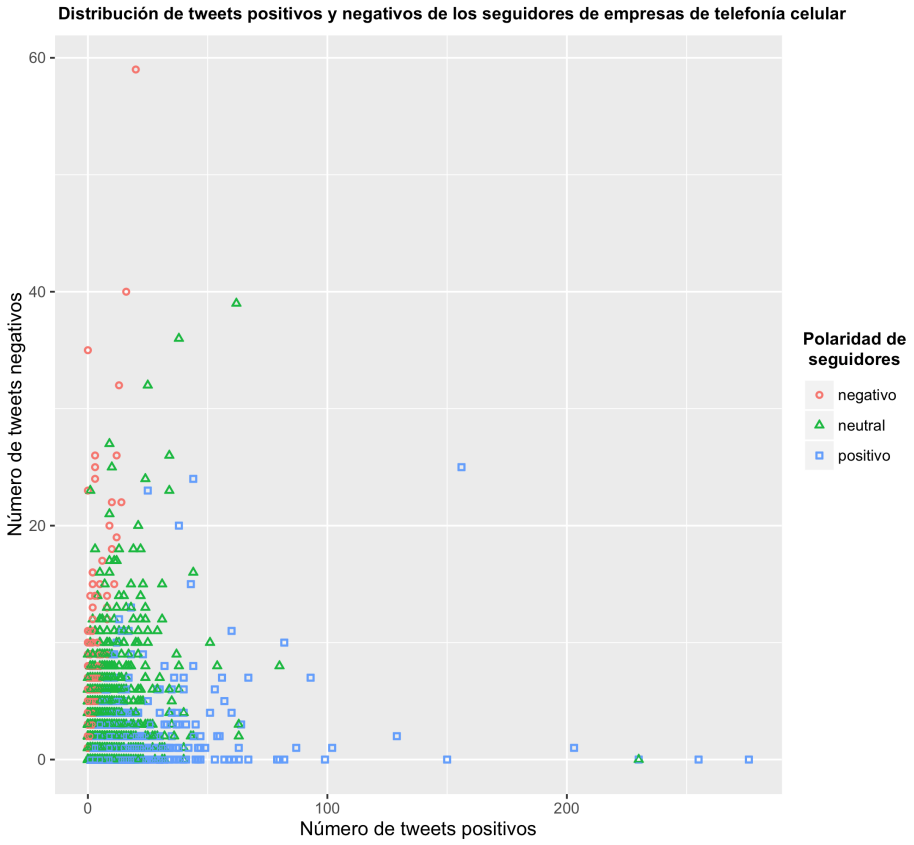

Figura 3.5 Relaciones entre tweets positivos, negativos y neutrales de los tweets de los seguidores de las empresas de telefonía celular

Por otro lado, se realizó un análisis acerca de tweets y seguidores de cada una de las empresas de telefonía celular la cual en términos porcentuales se refleja en la Tabla 3.12 donde se observa lo siguiente:

Tabla 3.12 Tabla comparativa entre seguidores de las empresas de telefonía celular y sus tweets

	Claro	Movistar	Entel	Claro %	Movistar %	Entel %
Seguidores totales	188026	195017	10146			
Seguidores activos	19048	17154	1846	100.00 %	100.00 %	100.00 %
Seguidores positivos	6181	6270	536	32.45 %	36.55 %	29.04 %
Seguidores negativos	1709	1562	156	8.97 %	9.11 %	8.45 %
Seguidores neutrales	11158	9322	1154	58.58 %	54.34 %	62.51 %
Tweets totales	75737	60982	6786	100.00 %	100.00 %	100.00 %
Tweets positivos	16675	16293	1557	22.02 %	26.72 %	22.94 %
Tweets negativos	2967	2572	272	3.92 %	4.22 %	4.01 %
Tweets neutrales	56095	42117	4957	74.07 %	69.06 %	73.05 %

1. En promedio, los seguidores de las 3 empresas de telefonía celular publican mayor cantidad de tweets neutrales en un promedio de 58.48 % observándose que la mayoría de tweets son publicaciones acerca de promociones, noticias o mensajes auto–generados, como se muestra en la Tabla 3.11.

2. La empresa de telefonía celular la cual posee dentro de sus seguidores el mayor porcentaje de usuarios, que en su mayoría, publica mayor número de tweets positivos es Movistar Perú con un 36.55 % de seguidores calificados como positivos.

3. Las 3 empresas poseen menos del 10 % de seguidores que publican mayor cantidad de tweets negativos lo que se entendería que esta red social no es muy utilizada para demostrar el descontento de los usuarios para con sus empresas proveedoras de servicios.

4. Los términos positivos, más utilizados entre las publicaciones de los usuarios se muestran en la parte izquierda de la Figura 3.6 donde podemos resaltar que la palabra más utilizada es *Dios* y también *Claro*. Así mismo, los términos más negativos se muestran en la parte derecha de la Figura 3.6 donde *triste* es la palabra más utilizada en las publicaciones de los usuarios, cabe resaltar que estos términos están fuertemente conectados con los mostrados en la Figura 3.3.

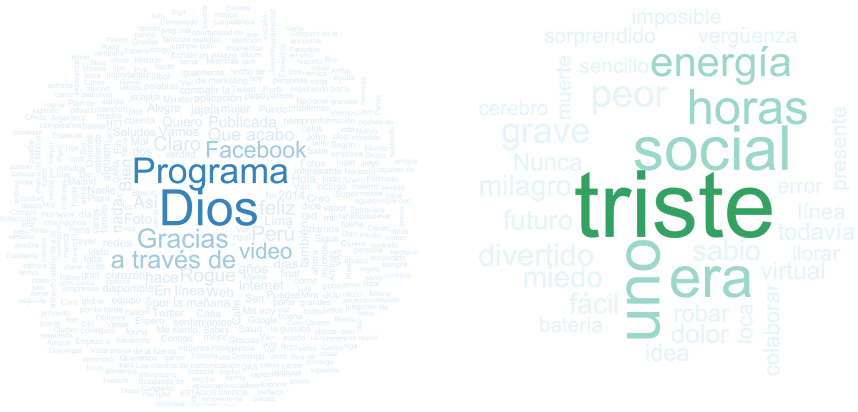

Figura 3.6 Nube de Palabras Positivas y Negativas más utilizadas

Así mismo, en la Figura 3.7 se muestra un análisis del comportamiento de los seguidores de cada empresa de telefonía celular de forma diaria, para esta figura se observa el análisis realizado al primer bloque de tweets correspondiente entre las fechas 2014-06-25 al 2014-07-25, correspondiente al primer mes de Entel Perú en Twitter. Como se observa el comportamiento de los tweets negativos de las empresas es siempre bajo en comparación con los tweets positivos y neutrales.

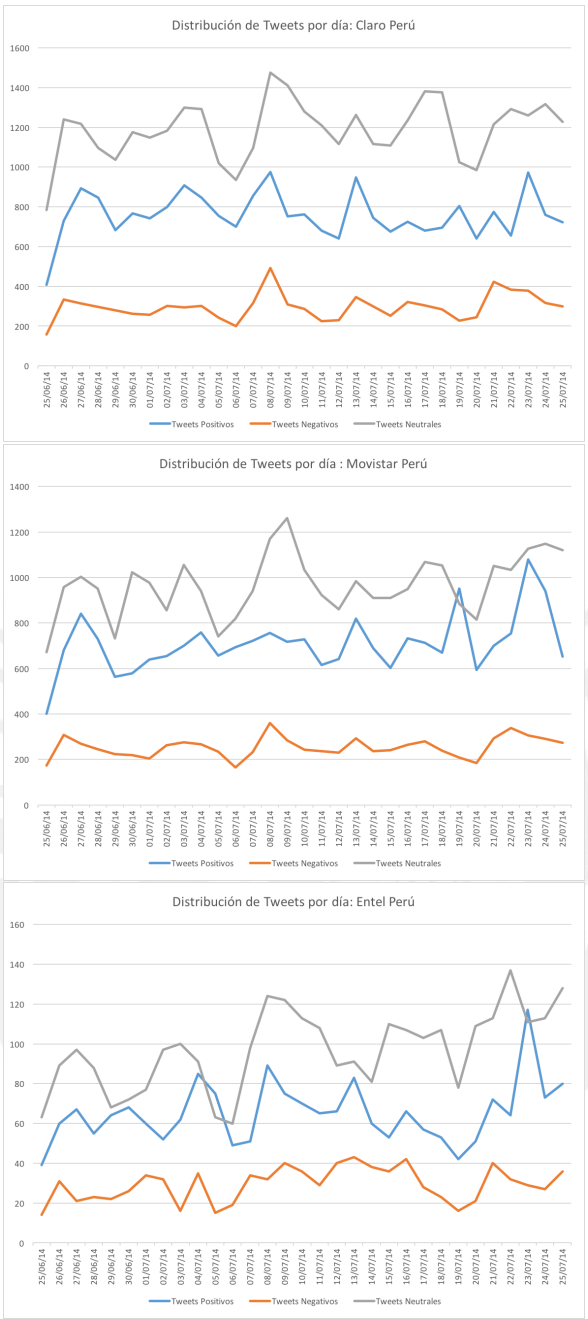

Figura 3.7 Comportamiento diario de los seguidores de las empresas de telefonía celular en el Perú: Bloque 1

De forma complementaria se realizó un análogo para con los tweets de las empresas de telefonía celular en las fechas 2014-10-13 al 2014-11-13, la cual se muestra en la Figura 3.8, aquí lo más resaltante es desde que Entel Perú comenzó a operar, el número de tweets negativos proveniente de los usuarios hacia Movistar y Claro se incrementó casi al 100%, mientras que la taza de tweets positivos de Entel Perú se incrementaba de forma constante,

los cuales en el transcurso de los días comenzó a fluctuar de forma más constante como se observa en la Figura 3.7 y Figura 3.8.

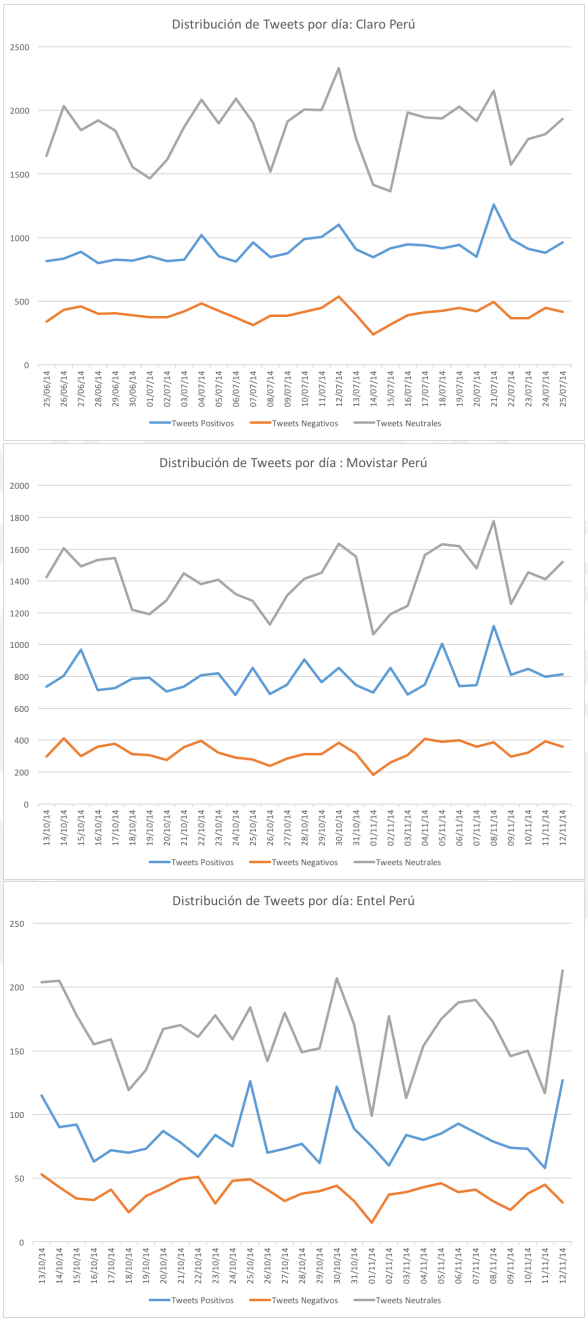

Figura 3.8 Comportamiento diario de los seguidores de las empresas de telefonía celular en el Perú: Bloque 2

Capítulo 4

Conclusiones y Trabajos Futuros

4.1. Conclusiones

En la presente investigación se llego a concluir las siguientes cosas con respecto al desarrollo experimental:

- Para el Objetivo Específico 1 se logró identificar que los tweets provenientes de los usuarios de telefonía celular no solo contenían información de telecomunicaciones sino también en otros contextos, para lo cual fue necesario desarrollar una estrategia que permita realizar la limpieza de estos tweets mediante un análisis de frecuencia de términos que permitiera identificar si un tweet era de telecomunicaciones. Este proceso permitió reducir de 26 millones de tweets a 2 millones, lo cual demuestra que aproximadamente el 92 % de tweets son poco relevantes.

- Para el Objetivo Específico 2 se identificó que la no eliminación de los emoticones permitió fortalecer el sentimiento del tweet que lo contenía. para ello se etiquetaron en total 191 emoticones los cuales estuvieron presentes en el 41.68 % de los tweets.

- Así también, se creó un diccionario propio basado en el uso de diccionarios estándares (SentiWordNet y EPA). Los mejores resultados se consiguieron combinando el diccionario EPA con los *hashtags* que poseían una polaridad calculada por *Label Propagation*.

- Para el Objetivo Específico 3 se observó que el mayor descontento de los usuarios de telefonía es acerca de su empresa proveedora y los servicios tradicionales, siendo en segunda instancia los servicios de Internet, seguridad y los equipos que brindan en sus diferentes planes, los términos que son más utilizados para identificar los tweets negativos hacia cada empresa.

- Para el Objetivo Específico 4 se verifica que la mayor cantidad de usuarios activos la posee la empresa Claro, seguida por Movistar y finalmente Entel. Esta última presenta un mayor porcentaje de seguidores neutros (62.51 %). La empresa Movistar presenta el mayor porcentaje de seguidores positivos (36.55 %) y negativos (9.00 %). Finalmente la empresa Claro presenta un balance entre seguidores positivos (32.45 %), negativos (8.97 %), y neutros (58.58 %). Esto nos indica que el mayor número de seguidores de estas empresas son de comentarios neutros referidos especialmente a noticia y propagandas de las empresas.

- Así también, se observó que desde que Entel Perú comenzó a operar en el mercado de telecomunicaciones el número de tweets negativos hacia Movistar y Claro se incrementó casi al 100 % mientras que la taza de tweets positivos de Entel Perú se incrementaba de forma constante, los cuales en el transcurso de los días posteriores a su inicio de operaciones comenzó a fluctuar de forma más constante.

4.2. Trabajos Futuros

Se plantea como trabajo futuro el analizar más a fondo los algoritmos de propagación de etiquetas tomando como base el *Label Propagation*, ya que como se evidenció en el presente trabajo el uso de un diccionario diferente al tradicional, como lo es EPA, mejoró el porcentaje de aciertos en la clasificación a pesar de tener menor cantidad de palabras conocidas y con valores en su puntaje de polaridad, esto debido a que durante el proceso de propagación de etiquetas para brindarle un puntaje de polaridad a los *hashtags* también se propagó a palabras que no se conocía.

Así como se logró identificar mediante la propagación de puntajes de polaridad, algunos servicios que eran de común descontento entre los seguidores de las empresas de telefonía celular sería interesante el poder extraer el tipo de seguidores que son los que se expresan de forma negativa de estos servicios y así identificar los grupos etarios donde se puede reforzar las promociones o beneficios del uso de servicios.

Bibliografía

Aisopos, F., Papadakis, G., Tserpes, K., y Varvarigou, T. (2012). Content vs . Context for Sentiment Analysis : a comparative analysis over microblogs. *L3Sde*, pages 187–196.

Alhumoud, S. O., Altuwaijri, M. I., Albuhairi, T. M., y Alohaideb, W. M. (2015). Survey on Arabic Sentiment Analysis in Twitter. 9(1):364–368.

Amiri, H. y Chua, T.-S. (2012). Mining slang and urban opinion words and phrases from cQA services. *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, page 193.

Bermingham, A. y Smeaton, A. F. (2010). Classifying sentiment in microblogs: Is brevity an advantage? In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1833–1836, New York, NY, USA. ACM.

Boldrini, E. y Balahur, A. (2010). EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. *Proceedings of the . . .* , (July):1–10.

Bosch, R. (2013). Sentiment Analysis : Incremental learning to build domain models. pages 1–42.

Bouillot, F., Ienco, D., Matwin, S., Poncelet, P., y Roche, M. (2012). Presidential election 2012: How French politicians tweet? pages 1–13.

Bravo-Marquez, F., Mendoza, M., y Poblete, B. (2013). Combining strengths, emotions and polarities for boosting Twitter sentiment analysis. *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '13*, pages 1–9.

Bukhari, A., Science, C., y Qamar, U. (2014). Critical Review of Sentiment Analysis Techniques. (September):15–16.

Councill, I. G., McDonald, R., y Velikovich, L. (2010). What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. *Proceedings of the ACL Workshop on Negation and Speculation in Natural Language Processing Uppsala Sweden*, (July):51.

Fayyad, U. M., Piatetsky-Shapiro, G., y Smyth, P. (1996). Advances in knowledge discovery and data mining. chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA.

Garza, E., Badii, M., y Abreu, J. (2008). Mejoramiento de la c alidad de servicios mediante el modelo de las discrepancias entre las expectativas de los clientes y las percepciones de la empresa (improvement of service quality through the discrepancy model between the expectations of the customers and the perceptions of the company). *Daena: International Journal of Good Conscience*, 3(1):1–64.

Gebreselassie, G. y Date, G. (2011). Sentiment Analysis of Twitter Posts About News. *Monographs of the Society for Research in Child Development*, 76(2):123.

Ghenname, M., Subercaze, J., Gravier, C., Laforest, F., Abik, M., y Ajhoun, R. (2014). A hashtags dictionary from crowdsourced definitions. *Proceedings - International Conference on Data Engineering*, pages 39–44.

Go, A., Bhayani, R., y Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12.

Hoey, J. (2015). Good News or Bad News : Using Affect Control Theory to Analyze Readers ' Reaction Towards News Articles. pages 1548–1558.

Hoey, J. y Schr, T. (2015). Bayesian Affect Control Theory. pages 529–536.

Jiang, L., Yu, M., Zhou, M., Liu, X., y Zhao, T. (2011). Target-dependent Twitter Sentiment Classification. *Computational Linguistics*, pages 151–160.

Karamibekr, M. y Ghorbani, A. A. (2013). Lexical-Syntactical Patterns for Subjectivity. pages 241–250.

Kiritchenko, S., Zhu, X., y Mohammad, S. M. (2014). Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research*, 50:1–15.

Li, D., Shuai, X., Sun, G., Tang, J., Ding, Y., y Luo, Z. (2012). Mining topic-level opinion influence in microblog. *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, page 1562.

Lim, K. W. (2014). Twitter Opinion Topic Model : Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Mejova, Y. (2009). Sentiment Analysis: An Overview. Comprehensive Exam Paper. *Computer Science Department*, pages 1–34.

Mejova, Y. A. (2012). *Sentiment Analysis Within and Across Social Media Streams*. PhD thesis, Iowa City, IA, USA. AAI3516660.

Pang, B. y Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

R-Moreno, M. D., Cuesta, A., y Barrero, D. F. (2013). Twitter stream analysis in Spanish. *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics - WIMS '13*, page 1.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., y Martin, M. (2004). Learning Subjective Language. *Computational Linguistics*, 30(3):277–308.