Proyecto Final de Ingeniería

Aplicación móvil para medir “sensaciones” a partir de redes sociales

Autor

Mariano Ezequiel Steininger

Ingeniería Informática

Tutor

Marcelo Alfonso Castro

Fecha

Septiembre de 2016

UADE

UNIVERSIDAD ARGENTINA DE LA EMPRESA

FACULTAD DE INGENIERÍA Y CIENCIAS EXACTAS

Equipo de Trabajo

Alumno(s)

Mariano Ezequiel Steininger, DNI 35.017.039, Legajo 132573, Ingreso 2007

Tutor(es) de Proyecto Final

Marcelo Alfonso Castro, FAIN, marcecastro@uade.edu.ar

Resumen

La masificación de internet, la utilización de los smartphones y el uso excesivo de las redes sociales hizo que en la web se encuentren diversos tipos de opiniones, por ejemplo, respecto a marcas, productos o empresas.

Todas estas opiniones pueden ser utilizadas para la toma de decisiones, desde el punto de vista del consumidor, para comprar o no determinado producto, para elegir una determinada empresa proveedora de un servicio (internet, telefonía, etc.); o bien, desde el punto de vista de las empresas pueden ser utilizadas, por ejemplo, para conocer la reputación de un producto o línea de productos y así poder decidir respecto a las nuevos ofertas y/o lanzamientos.

Actualmente hay empresas que contratan recursos y servicios para averiguar y tener conocimiento de su imagen, esto conlleva una gran inversión de tiempo y dinero. Tener una herramienta que extraiga las opiniones, las procese y tenga la capacidad de determinar de forma automática si son positivas o negativa, resultaría en una reducción de tiempos y costos.

Ahora bien, no es para nada sencillo que una herramienta informática pueda interpretar un texto, es por ello que nació lo que se conoce como procesamiento del lenguaje natural (PLN). Esto se debe a que cuando un ser humano lee un texto u opinión, hay muchos mecanismos mentales que hacen que la persona que está leyendo lo interprete y entienda, en cambio, una herramienta no puede hacer eso por sí sola directamente.

Particularmente, para determinar la polaridad de una opinión existen diversos modelos y estrategias que se verán a lo largo del presente trabajo: aprendizaje supervisado, aprendizaje no supervisado, reglas ad-hoc, léxicos de opinión, basada en regiones y el modelo spin.

El presente trabajo tiene como objetivo desarrollar una herramienta informática basada en un modelo, que también se desarrollará, para determinar la polaridad de la opinión en forma cuantitativa y que tenga apertura a diversos idiomas.

Para poder cumplir con lo dicho, se propondrá una implementación utilizando los recursos de BabelNet, SentiWordNet y APIs de Java.

Inicialmente la implementación se realizará para el idioma español y el porcentaje de exactitud esperado para determinar la polaridad de las opiniones se establecerá superior al 70%.

Abstract

Internet has become very popular, a lot of people used smartphones and social networks. The consequences are that we can find to many opinions of different topics, for example brands, products or companies.

These opinions can be used to take decisions from the consumers point of view to buy some products or services (internet, phone, computer brands etc.); the companies can also search about products reputation. This information could be relevant for future marketing decisions.

Currently there are companies that hiring employees and services to be aware of their image, but it takes greats investments in time and money. Having tools to automatically extract opinions in the web, processes them and detect if there are positive or negative will result useful to minimize times and costs.

To develop a software that translate texts and detect positive or negative opinion is not as simple as it sound. That's why the natural processing language (NLP) born. When a man reads a text there are many mental mechanisms that influence in who is reading, that's what a software can´t do on its own.

Particularly, to determine the polarity in opinions there are several models and strategies that will be develop on this report like supervised learning, unsupervised learning, using ad-hoc rules, used lexicon opinion based on regions and spin model.

This report aims to develop a software tool that will determine the polarity of opinion (negative or positive) quantitatively and in several languages.

To complete the tool development will propose an implementation using Babelnet, SentiWordNet and Java APIs resources.

The implementation will be done for Spanish language and the accuracy rate expected for the polarity of opinions is set higher than 70%.

Tabla de Contenidos

1. INTRODUCCIÓN

La masificación del uso de internet y los Smartphone provocó un crecimiento exponencial del uso de las redes sociales, donde las personas publican diariamente sus experiencias y opininiones. Mucho de todo este contenido puede expresar las sensaciones de las personas respecto a algo (empresas, productos, personas, etc), entendiéndose como sensaciones a las emociones (enojo, alegría, tristeza, orgullo, etc), que a su vez se las puede clasificar en positivas o negativas. Todas estas sensaciones podrían ser utilizadas por las empresas para obtener:

- Imagen de la empresa, marcas y/o productos, dadas por publicaciones de:
  - Experiencias por la interacción directa con la empresa.
  - Opiniones sobre la empresa.
  - Experiencias y opiniones sobre líneas de productos.
  - Experiencias y opiniones con productos específicos.
- Oportunidades generadas por nuevas necesidades o carencias de los productos en el mercado.

Dado que las opiniones expresadas en las redes sociales pueden influir en la elección de productos o servicios, el conocimiento de las sensaciones por parte de las empresas empieza a tener aún más importancia. Por lo mencionado, uno de los principales beneficios es obtener mayor cantidad de clientes.

Para poder explotar y obtener beneficios de toda la información mencionada, es necesario realizar un procesamiento computacional que no sólo recolecte las opiniones en las redes sociales, sino que también procese, entienda y muestre el resultado del procesamiento de forma tal que le permita a las empresas tomar decisiones. El entender la información recopilada requiere un profundo conocimiento lingüístico que en ocasiones puede requerir un elevado coste computacional, es por ello que nació lo que se conoce como minería de opiniones. Muchos sistemas de minería de opiniones utilizan:

- **Lingüística computacional**: Trata de construir modelos de lenguaje entendibles por las computadoras.
- **Procesamiento del Lenguaje Natural (PLN)**: Es la aplicación de los modelos definidos por la lingüística computacional, ocupándose de aspectos más

técnicos y de los algoritmos con el fin de estructurar los textos y extraer la información que hay en ellos.

1.1 DEFINICIÓN DEL PROBLEMA

En el presente trabajo se tomaran las sensaciones como positivas o negativas, sin entrar en el detalle de la emoción que expresa cada opinión. Cuando se quiere que una maquina determine la polaridad de un texto de forma automática es necesario que se resuelvan principalmente las siguientes cuestiones:

- **Negaciones**: Son un modificador que invierte la polaridad de expresiones polares, es decir que estas expresiones polares pasan a tener la polaridad opuesta. La principal problemática es identificar dónde termina el ámbito de acción de la negación.
- **Intensificaciones**: Son expresiones que pueden modificar (tanto aumentando como disminuyendo) la intensidad de las palabras que están dentro de su ámbito de acción. La problemática es, como en las negaciones, identificar donde termina el ámbito de acción y además en cuanto aumenta o disminuye la intensidad de las palabras que conforman el ámbito de acción.
- **Modalidad**: Si bien no hay una única definición, a nivel general todas las definiciones de la modalidad intentan diferenciar entre hechos o eventos que sucedieron de otros que tienen una cierta probabilidad de que sucedan.
- **Sarcasmos, ironías y sátiras**: Ponen en evidencia la clara diferencia entre el significado literal y la intención de un texto, mientras el significado literal depende del significado de las palabras de una frase, la intención está asociado a qué es lo que realmente quiso decir la persona
- **Orientación e importancia de las palabras**: Este aspecto es uno de los más importantes a la hora de decidir cómo se va a determinar la polaridad de un texto. Las preguntas a responder son:
    - Para una misma palabra que puede tener diferentes categorías léxicas, ¿la orientación y peso es el mismo en todas sus categorías léxicas? Por ejemplo “camino”, puede ser un verbo o un sustantivo. Se conoce como ambigüedad léxica.
    - ¿Todos los sentidos de una palabra tienen la misma orientación y peso? Por ejemplo “cura” puede referirse al sacerdote o a la medicina.

A las palabras mencionadas se las denomina polisémicas, y generan lo que se llama ambigüedad semántica.

- Hay una tercera ambigüedad, la ambigüedad sintáctica. Ésta se da cuando una oración puede ser interpretada de más de una forma, es por ello que se considera que es la que menos se da cuando lo que se quiere es determinar la polaridad de un texto.

Las cuatro primeras cuestiones se verán en detalle, incluyendo la forma en que se ha intentado resolverlas, en el Estado de arte. La orientación e importancia de las palabras se considera que es una cuestión más de diseño del modelo que uno desee implementar y por lo tanto se irá detallando durante todo el trabajo las decisiones tomadas, a modo introductorio, se menciona que se decidió hacer una desambiguación léxica, mientras que para la ambigüedad semántica se ha utilizado un promedio entre todos los significados.

1.2 OBJETIVOS

El presente trabajo tiene como objetivo proponer un modelo que permita determinar la polaridad de las opiniones aplicable a diversos idiomas. Adicionalmente se hará la implementación de dicho modelo para la polarización de opiniones en español extensible a otros idiomas. En cuanto al alcance, se excluyó la detección de sarcasmos y la modalidad. Se definió como meta tener un porcentaje de opiniones polarizadas correctamente superior al 70%.

2. ESTADO DEL ARTE

2.1 INTRODUCCIÓN

En el presente capitulo se describirán los problemas y dificultades para determinar la polaridad de una opinión, la tarea para realizar la desambiguación léxica (etiquetado gramatical), diferentes recursos que se utilizan para la polarización y finalmente una posible taxonomía de los métodos de polarización junto con la explicación de los más importantes.

2.2 PROBLEMAS Y DIFICULTADES

Para determinar la polaridad de las opiniones, es importante tener en cuenta ciertas palabras o situaciones que pueden afectar directamente a la polaridad de la opinión, ya que si no se tienen en cuenta es posible obtener resultados fallidos.. Hasta el presente, entre los más tenidos en cuenta se encuentran:

- La negación
- Los intensificadores

Entre los que no son muy tenidos en cuenta encontramos:

- La modalidad
- Sarcasmos, ironías, sátiras

2.2.1 Negación

La negación es considerada un elemento lingüístico para, valga la redundancia, negar una palabra, parte de la oración o una oración completa utilizando un sema lexicalizado, una palabra, normalmente adverbio, o una locución.

Algo que no está muy claro al momento de procesar una opinión es el alcance de la negación, para ello existen dos estrategias diferentes, la primera es considerar que la negación abarca todas las palabras entre la señal de negación y el primer signo de puntuación (Pang, Lee, y Vaithyanathan, 2002); la segunda estrategia es considerar un número fijo de palabras después de la señal de negación (Hu y Liu, 2004). Sin embargo, más allá de estas dos estrategias, al presente hay pocos trabajos que intentan delimitar el ámbito de acción de la negación.

Se considera que las negaciones son un modificador que invierte la polaridad de expresiones polares, es decir que estas expresiones polares pasan a tener la

polaridad opuesta. Las palabras que son afectadas por las negaciones se las conoce como constituyentes.

Según Wilson, Wiebe y Hoffmann (Wilson et.al, 2005; 2009), si se trata la negación como un modificador de la polaridad se puede modelar como 3 tipos de grupos:

- **Características de negación**: Representan la existencia de una negación, entre las que se pueden diferenciar: la negación local ("no es bueno"), negación del sujeto ("nadie piensa que es bueno") o negaciones que afectan varias palabras, por ejemplo la negación de la proposición ("no se ve muy bien")
- **Modificadores de características**: Hacen referencia a la presencia de diferentes modificadores de la polaridad y son considerados más débiles que las expresiones de negación ordinarias. Se pueden distinguir tres tipos: modificadores de polaridad generales, los de polaridad positiva y los de polaridad negativa. La principal diferencia entre estos tipos es que el primer tipo invierte la polaridad como las negaciones, mientras que los otros dos solo invierten un tipo de polaridad particular, por ejemplo un modificador de polaridad positiva puede modificar expresiones negativas o también puede expresar polaridad positiva.
- **Características de modificación de la polaridad**: En contraste con las caracteristicas de negación, hay palabras que no son explícitamente negaciones pero modifican las expresiones polares, por ejemplo "desilusionado."

Estos tres grupos son tenidos en cuenta y utilizados para determinar la polaridad de los términos. La polaridad de cada constituyente se obtiene de un lexicón afectivo y aplicando reglas de inferencia se deriva la polaridad total a partir de las polaridades individuales.

Para resolver la polaridad de una oración que tiene términos negativos se puede decir que consta de dos etapas: determinar los constituyentes una vez encontrada la negación siguiendo alguna estrategia y modificar la polaridad de los constituyentes.

2.2.2 Intensificadores

En la mayoría de los trabajos se define a los intensificadores a los términos o expresiones que pueden modificar (tanto aumentando como disminuyendo) la intensidad de las palabras que están dentro de su ámbito de acción. Dado que hay términos que aumentan y otros que disminuyen la intensidad Quirk, Greenbaum, Leech y Svartvik (Quirk et.al,1985) clasificaron los intensificadores en dos grupos:

- **Amplificadores**: Aumentan la intensidad de las palabras que están dentro de su ámbito de acción, por ejemplo “tan”, “muy” y “bastante”
- **Decrementadores o disminuidores**: Disminuyen la intensidad de las palabras que están dentro de su ámbito de acción, por ejemplo “poco”, “en absoluto”

El enfoque más utilizado para detectar los intensificadores es tener una lista con los cuantificadores y tenerlos en cuenta si son utilizados como modificadores adverbiales y adjetivales.

Una vez detectados los cuantificadores se les asigna un valor dependiendo de cómo modifiquen la polaridad, para ello hay diferentes estrategias. Una de las estrategias, utilizada por Kennedy y Inkpen (Kennedy y Inkpen, 2006) y por Polanyi y Zaenen (Polanyi y Zaenen, 2006), consiste añadir un valor a los intensificadores usando sumas y restas: para adjetivos positivos el valor de la intensidad de la expresión polar es de 2, si el adjetivo es amplificado tendrá un valor de 3 y si es disminuido un 1; para los adjetivos negativos se les asigna un valor de -2, -3 si son amplificados y -1 si son disminuidos. Esta estrategia tiene principalmente dos inconvenientes:

1. No todos los cuantificadores intensifican o disminuyen de la misma manera, no es lo mismo “extraordinariamente” que “bastante”, el primero es más fuerte que el segundo; en otras palabras, es necesario tener en cuenta diferentes intensidades para determinar el impacto real de cada cuantificador sobre los términos que están dentro de su ámbito de acción.
2. No se tiene en cuenta que hay términos más fuertes que otros, no es lo mismo decir “realmente fantástico” que “realmente bien”, ya que “fantástico” es más fuerte que “bien” y la intensificación debe depender también del término que se intensifica.

Para finalizar, cabe destacar que hay otras formas de enfatizar opiniones, por ejemplo usando mayúsculas o signos de exclamación, aun así se les puede aplicar la misma estrategia que el resto de los intensificadores.

2.2.3 Modalidad

Si bien no hay una única definición, a nivel general todas las definiciones de la modalidad intentan diferenciar entre hechos o eventos que sucedieron de otros que tienen una cierta probabilidad de que sucedan. Por ejemplo:

- Polanyi y Zaenen (Polanyi y Zaenen, 2006) afirman que los operadores modales permiten distinguir entre eventos reales y eventos irreales; entre los eventos reales, podemos encontrar lo que ha ocurrido, ocurre o va a ocurrir, mientras que los irreales se refiere a los que podría ocurrir o lo que debería suceder.
- Carter y McCarthy (Carter y McCarthy, 2006) definieron la modalidad desde el punto de vista lingüístico, y consideraron importantes dos significados. El primero se refiere a la probabilidad, posibilidad y certeza, evaluaciones y deducciones de posibles hechos o eventos. El segundo hace referencia a los eventos o situaciones realizadas, toma de control del curso de dichos eventos o situaciones, generalmente expresando cuándo algo es obligado, necesario, deseable, debe estar prohibido o permitido.
- Lakoff (Lakoff, 1973) la denomina hedging y la define como “palabras cuyo trabajo es hacer las cosas más o menos difusas”.
- Sauri y Pustejovsky (Sauri y Pustejovsky, 2009) la denominan como factuality y la definen como “información que determina cuándo un evento mencionado en el texto se corresponde con una situación real o, por el contrario, con una situación con cierto grado de incertidumbre”.

Polanyi y Zaenen (Polanyi y Zaenen, 2006) consideran que dado que la modalidad puede neutralizar la polaridad de las expresiones polares que se encuentran dentro de su ámbito de acción, es importante tenerla en cuenta en el análisis emocional. Si bien la incertidumbre expresada mediante términos modales puede modificar el significado emocional, como han mencionado Baker, Bloodgood, Dorr, Callison-Burch, Filardo, Piatko, Levin y Lori y Miller Scott (Baker et al., 2012), en general la incertidumbre no modifica el significado emocional ya que las

expresiones polares positivas o negativas lo siguen siendo independientemente de si el evento o situación es un hecho o no, por ejemplo la oración “Te quiero matar” expresa que la acción no ha sido realizada, pero no quedan dudas de que es negativa. Desde la perspectiva del análisis afectivo, se considera que las formas modales que expresan deseos, necesidades o quejas (segundo significado de Carter y McCarthy (Carter y McCarthy, 2006)) generalmente modifican la polaridad e intensidad de las expresiones polares que están dentro de su ámbito de acción.

Hasta el presente hay pocos estudios respecto a cómo resolver el efecto de las formas modales en sistemas de análisis sentimental. En este sentido, un estudio para mencionar es el de Brooke (Brooke, 2009), donde hace foco en neutralizar las expresiones polares de los fragmentos de texto que contengan términos irrealistas, es decir que se anula toda expresión polar que este dentro del ámbito de acción de los modales que expresen eventos y situaciones no reales. Más allá del estudio mencionado y de otros estudios realizados, en ninguno se ha estudiado como impacta realmente la modalidad en las diferentes tareas del análisis sentimental ni cuál es el efecto de las diferentes formas modales en las emociones y sentimientos dentro del ámbito de acción de cada forma modal, sino que más bien se han hecho a un nivel global sin profundizar en cada forma modal en particular.

2.2.4 Sarcasmo

Como se lo describió Jaffe (Jaffe, 2014) tanto la ironía como el sarcasmo ponen en evidencia la clara diferencia entre el significado literal y la intención de un texto, mientras el significado literal depende del significado de las palabras de una frase, la intención está asociado a qué es lo que realmente quiso decir la persona. Es necesario mencionar que a los humanos no nos es fácil identificar las ironías: de acuerdo al pensamiento científico, el cerebro obtiene el significado literal de una frase, analiza la situación actual (podemos decir que sería el contexto en que fue dicha la frase) y se calcula una desconexión e infiere la intensión irónica, por ejemplo si alguien dice “Que rápida que funciona la computadora!” mientras observamos que la realidad es que anda muy lenta, el cerebro inicialmente registra que realmente funciona de forma rápida, al registrar que en realidad funciona lenta termina infiriendo que era una ironía, si la frase hubiera sido “Que lenta que anda la computadora!” el procesamiento e interpretación hubiera sido más rápido. Ahora bien, este tipo de ironía es sencilla, pero la realidad es que hay ironías más difíciles de captar.

Jaffe (Jaffe, 2014) también ha mencionado un experimento reciente, dirigido por la psicóloga Ruth Filik de la universidad de Nottingham que consistía en escuchar grabaciones de escenas irónicas (algunas contenían frases irónicas familiares típicas, como por ejemplo el “Veo que estás estudiando mucho” dicho por las madres a los hijos que están jugando videojuegos en lugar de estudiar) y chistes irónicos. El resultado obtenido fue que las personas tardaron más en reconocer las ironías no familiares, mientras que las familiares las captaron tan rápido como si no lo fueran, en conclusión si estamos acostumbrados a escuchar ciertas frases siempre con intenciones irónicas, el cerebro las almacenara tales y serán captadas más rápidamente.

A diferencia de lo que sucedía con las negaciones, los intensificadores y los modales, en los que podemos identificar palabras que nos dan indicios de que están presentes en la frase, en la ironía o sarcasmo esto no sucede y ahí es dónde radica la principal dificultad para el procesamiento automático. A esta dificultad se le suma el hecho de que para que los humanos captemos la ironía, como se mencionó anteriormente, a veces es necesario ver el contexto en que nos encontramos y cuando se intenta procesar de forma automática una determinada frase no se cuenta con el contexto visual que permita captar la ironía. Podríamos encontrar el contexto necesario para interpretar la ironía en textos, donde la frase u oración está dentro de un contexto literario, aun así es posible no tener mucha precisión; igualmente en general las ironías las encontramos en redes sociales, donde toda la frase puede ser sarcástica, sin tener un contexto que ayude a captar dicha ironía, es por ello que es necesario un enfoque diferente, es decir, teniendo en cuenta que no hay palabras que indiquen la presencia de la ironía y sin la existencia de un contexto que ayude a captarla. En la comunidad científica no fue hasta hace pocos años que se comenzó a estudiar y tratar de captar las ironías en textos.

Uno de los trabajos para destacar es el de Barbieri, Ronzano y Saggion (Barbieri, Ronzano y Saggion, 2015), dónde proponen un modelo computacional en el que cada mensaje es representado por un conjunto de rasgos diseñados para detectar el estilo satírico y no el contenido. Este modelo lo aplicaron a Twitter, por lo que cada mensaje sería un tweet. El modelo consiste en caracterizar cada tweet en 7 clases de características: frecuencia, ambigüedad, etiqueta gramatical (mas conocida como part of speech), sinónimos, sentimientos, caracteres (tienen en cuenta cantidad de mayúsculas, caracteres especiales como “.”, “!”, “?”, “$”, “%”, “&”, “+”, “-”, “=”) e improperios. Por otro lado tiene otro conjunto de características

basadas en las palabras (lema, bigramas y salto 1/2/3 gramas) utilizado para entrenar la línea base del sistema con el fin de realizar una evaluación comparativa con los otros grupos de características; cada conjunto de entrenamiento tiene las 1000 palabras más frecuentes para cada característica. Con el modelo propuesto por ellos obtuvieron alrededor de 75% de exactitud.

Más allá del trabajo mencionado y de los existentes, aún queda mucha investigación y trabajo por realizar, para mencionar un ejemplo, en una ocasión se le ha pedido a Siri (aplicación de Apple para responder preguntas y realizar acciones mediante el lenguaje coloquial, sin ingresar nada desde teclado) que reproduzca una canción y al reproducir una canción que no era la pedida se le dijo sarcásticamente “Eres fantástica”, la respuesta de Siri fue “Gracias”. Si bien el ejemplo en una interacción humano-maquina, esto es aplicable para sistemas de análisis emocional u otro tipo de sistemas.

2.3 ETIQUETADO GRAMATICAL

El etiquetado gramatical (part-of-speech tagging, POS tagging o POST) es proceso que consiste en asociar a cada palabra del texto que se está analizando una etiqueta que indique su categoría léxica, teniendo en cuenta el contexto dentro del texto en que fue usada cada palabra. La principal complejidad del etiquetado gramatical consiste en la ambigüedad gramatical, por ejemplo la palabra curva puede ser un sustantivo o adjetivo dependiendo de cómo se use:

“Cuidado con la curva que está a 100 metros” (sustantivo)

“Por favor dibuja una línea curva” (adjetivo)

La mayoría de las implementaciones están basadas en el aprendizaje automático, es decir que se utiliza un corpus anotado para aprender y otro para etiquetar. Una vez entrenado el etiquetador, éste va a tener información acerca de cada palabra, la etiqueta asignada y su contexto para poder etiquetar otros textos.

2.3.1 Etiquetas

Generalmente la etiqueta gramatical o part-of-speech (POS) se ha basado en funciones sintácticas y morfológicas; desde el punto de vista sintáctico se agrupan en clases que funcionan de forma similar respecto al contexto en que se encuentran; por otro lado, desde el punto de vista morfológico se agrupan respecto a los afijos que contienen. Si bien hay muchas formas de categorizar las etiquetas gramaticales, es posible identificar dos grandes categorías o clases, por un lado las clases cerradas y por otro las clases abiertas. Las clases cerradas son las que tienen un conjunto miembros relativamente fijo, en esta clase podemos encontrar por ejemplo las preposiciones ya que es un conjunto cerrado de palabras y rara vez se agregan nuevas preposiciones; en esta clase también se encuentran las palabras funcionales (por ejemplo “de”) que en general son muy cortas y tienen usos estructurales en la gramática. En cambio, las clases abiertas son las que no tienen un conjunto de miembros fijo y varían constantemente, por ejemplo los verbos y sustantivos, ya que continuamente se agregan nuevos o eliminan. Entre las clases abiertas encontramos:

- **Sustantivos**: Son la clase sintáctica que expresa personas, lugares o cosas. Dado que las clases sintácticas (entre ellas los sustantivos) son definidas sintáctica y morfológicamente en lugar de semánticamente hay palabras para personas, lugares y cosas que no pueden ser sustantivos y a la inversa, hay sustantivos que pueden no ser palabras para personas, lugares o cosas. Entre las principales características se encuentran la capacidad para ocurrir con determinantes, tomar posesivos y ocurrir en forma plural (la mayoría pero no todos). Son agrupados en dos grupos:
    - **Sustantivos propios**: Son nombres de personas o entidades, generalmente empiezan en mayúscula.
    - **Sustantivos comunes**: En algunos idiomas, como el inglés se los clasifica en:
        - **Sustantivos contables**: Permiten establecer un número en unidades y generalmente posee forma singular o plural, por ejemplo remera/s, monitor/es
        - **Sustantivos incontables**: No es posible determinar un número de unidades, por ejemplo sal, azúcar
- **Verbos**: Son una clase de palabra que expresan acciones o procesos. Tienen formas morfológicas, algunas de ellas son de tiempo, modo, persona,

regularidad. Puede concordar en género, persona y número con algunos argumentos o complementos (conocidos como sujeto, objeto, etc)

- **Adjetivos**: Son palabras que expresan cualidades o propiedades y si bien tienen al igual que los sustantivos tienen género y número hay algunos que tienen un único género, por ejemplo amable y grande. El género y número dependen del sustantivo que los acompañe. Si bien la clasificación de adjetivos es más amplia, los más relevantes y utilizados entre los adjetivos de clase abierta son los adjetivos calificativos, que añaden cualidades al sustantivo y se dividen en:
    - **Especificativos**: Concretan el significado del sustantivo y suelen aparecer después del sustantivo, por ejemplo "Quiero un auto rojo"
    - **Explicativos o epítetos**: Indican cualidades que el sustantivo lleva de por sí y suelen aparecer antes del sustantivo, por ejemplo "blanca nieve" o "verde hierba".
    - **Adverbios**: Son definidos como modificadores del verbo, adjetivo o de otro adverbio. Se pueden clasificar en adverbios de:
        - **Lugar**: aquí, allí, ahí, allá, acá, arriba, abajo, cerca, lejos, encima, debajo, delante, detrás, enfrente, alrededor, atrás, etc.
        - **Tiempo absoluto**: pronto, todavía, aún, ya, ayer, hoy, mañana, siempre, nunca, jamás, enseguida, ahora, mientras, próximamente, prontamente, tarde, temprano, anoche.
        - **Modo**: bien, mal, peor, mejor, regular, así, tal, como, aprisa, adrede, despacio, deprisa, todas las que se formen con las terminaciones "mente".
        - **Cantidad o grado**: tan, tanto, todo, nada, aproximadamente, muy, poco, muy poco, mucho, bastante, más, menos, algo, demasiado, casi, sólo, solamente.

Entre las clases cerradas encontramos:

- **Preposiciones**: Son enlaces formados por una o varias palabras que unen componentes de una oración para brindarles sentido. La significación responde a circunstancias de movimiento, lugar, tiempo, modo, causa, posesión, pertenencia, materia y procedencia

- **Determinantes**: Ocurren con sustantivos y en general marcan el inicio de una frase sustantiva, por ejemplo “a”, “el”, “ese”
- **Verbos auxiliares**: Brindan información gramatical (de modo, tiempo, persona, número y formas impersonales) y semántica adicionales al otro tipo de verbos mencionado. Por ejemplo “Mañana habremos llegado a casa”
- **Conjunciones**: Son utilizadas para unir frases, clausulas o sentencias. Se dividen en:
    - **Coordinantes**: Unen elementos de igual estado, por ejemplo “y” y “o”
    - **Subordinativas**: Se utilizan cuando uno de los elementos es de algún estado integrado, por ejemplo “Me molestó que hayas roto el vidrio”
- **Numerales**: También conocidos como determinantes numerales, generalmente acompañan o están acompañados por sustantivos y expresan cantidades, posiciones, etc. Los más importantes son:
    - **Cardinales**: Indican número o cantidad de elementos, por ejemplo “Bajó siete pisos”
    - **Ordinales**: Expresan el orden de un sustantivo dentro de un grupo, por ejemplo, “Salió primero en la maratón”
    - **Múltiplos**: Expresa las veces que se contiene una cantidad, por ejemplo “Quiero el doble de porción de nueces”
    - **Partitivos**: Expresan las partes en que se puede dividir un objeto, por ejemplo “Quiero la mitad de la pizza”, “Nos vemos en media hora”
- **Pronombres**: Son formas que en general actúan como una clase de atajo para referirse a un evento, entidad o frase sustantiva. Se dividen en:
    - **Personales**: Hacen referencia a personas o entidades, por ejemplo “yo”, “tú”, “él”
    - **Posesivos**: Generalmente indican una relación abstracta entre una persona y un objeto, por ejemplo “mío”, “tuyo”, “mi”.

2.3.2 Conjunto de etiquetas

Al día de hoy no existe un conjunto de etiquetas o tagset adecuado. Hay tanto grandes tagsets como otros más acostados, los primeros generalmente brindan descripción sintáctica más específica. La elección del tagset a utilizar depende de las necesidades de detalle lingüístico. Generalmente y como era de esperarse los tagsets más acotados están incluidos en los más grandes ya que las

etiquetas más específicas pueden ser convertidas en etiquetas de menor especificidad (perdiendo el detalle). Por otro lado, también es posible convertir etiquetas de menor detalle en etiquetas de mayor detalle. Por ejemplo, Penn Treebank tiene el siguiente tagset:

Tabla I: Tags Treebank

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense

Number	Tag	Description
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Fuente: Alphabetical list of part-of-speech tags used in the Penn Treebank Project

2.3.3 Métodos de etiquetado

Como lo mencionan Jurafsky y Martin (Jurafsky y Martin, 2009), la entrada para todo algoritmo de etiquetación automática es una cadena de palabras junto con un tagset (conjunto de etiquetas). Generalmente los signos de puntuación también son etiquetados, es por ello que es necesario que sean separados de las palabras. El proceso para separar los signos de puntuación de las palabras se denomina tokenización y puede realizarse de forma previa o durante el etiquetado. Durante la tokenización también se hace la desambiguación del fin de la oración de un signo de puntación (como las abreviaciones, por ejemplo “etc.”).

El mayor problema del etiquetado gramatical consiste en resolver las ambigüedades, las mismas se dan cuando una palabra puede tener diferentes etiquetas y es necesario analizar el contexto en que es utilizada para determinar la etiqueta a asignar a la palabra.

Jurafsky y Martin (Jurafsky y Martin, 2009) identificaron 3 clases de algoritmos de etiquetado automático:

1. **Etiquetadores basados en reglas**: Cómo su nombre lo indica, son etiquetadores que tienen un gran conjunto de reglas para desambiguar palabras, dichas reglas son realizadas de forma no automática.
2. **Etiquetadores basados en probabilidad o estocásticos**: Son etiquetadores que fueron entrenados utilizando corpus de entrenamiento para que aprendan a desambiguar y etiquetar palabras. El aprendizaje

consiste en extraer información de la probabilidad de que a una palabra le corresponda una etiqueta determinada según el contexto en que aparece.

3. **Etiquetadores basados en la transformación**: Es un combinación de los otros dos, los etiquetadores de esta clase utilizan reglas para desambiguar las palabras pero dichas reglas son aprendidas de forma automática a partir de corpus de entrenamiento etiquetados previamente.

Además de estas 3 clases, hay otros paradigmas utilizados, por ejemplo modelos de máxima entropía, árboles de decisión, RNA, SVM, algoritmos evolutivos, etc

2.3.4 Etiquetadores basados en reglas

De acuerdo a lo mencionado por Paniagua, García, y Gallardo-Paúls (Paniagua, García, y Gallardo-Paúls, 2005) el funcionamiento de estos etiquetadores consta de dos etapas:

1. Para cada palabra se determinan las posibles etiquetas, a partir de un diccionario o analizador morfológico
2. Se aplican las reglas contextuales que, teniendo en cuenta el contexto en que fue utilizada cada palabra descarta etiquetas hasta seleccionar que mejor se adecue al contexto.

Como todo enfoque basado en reglas, las principales ventajas es el diseño flexible y potente, por ejemplo el etiquetador utilizado el analizador EngGC para el inglés (Karlsson et.al, 1995) tiene una exactitud del 99%; entre las desventajas se encuentra el tiempo que requiere diseñar las reglas, la complejidad de mantenimiento que crece a medida que crecen la cantidad de reglas, por ejemplo, según para el idioma ingles se manejan 3800 reglas y un diccionario de etiquetas con 56000 entradas.

2.3.5 Etiquetadores basados en probabilidades

De acuerdo a Jurafsky y Martin (Jurafsky y Martin, 2009) para el etiquetado de cada palabra, los etiquetadores de esta clase seleccionan la etiqueta más probable en términos estadísticos. El modelo se basa en la probabilidad de la coaparición de categorías en los textos, es decir que para etiquetar una determinada palabra analiza la probabilidad de una determinada etiqueta posible dadas las

etiquetas de palabras anteriores. En esta clase de etiquetadores, se pueden destacar dos:

- **Modelo de N-gramas**: El modelo se supone que solo unas pocas palabras condicionan la probabilidad de aparición de la siguiente palabra. El "n" indica el número de etiquetas anteriores a tener en cuenta, los valores más usados son 1, 2 y 3, denominados unigrama, bigrama y trigrama respectivamente. La creación de los trigramas y todo n-grama, independientemente del valor de "n" es utilizando corpus de entrenamiento y registrando cada uno de los tríos o cualquier otra unidad (como se mencionó, dado por el valor de "n") que aparezca en el texto y además, se debe calcular la probabilidad de aparición de cada trío o unidad elegida utilizando la formula $P(U_i \mid U_{i-2}\ U_{i-1})$. No es el más utilizado pero es el que mejores resultados ha obtenido ya que teniendo en cuenta parte del contexto local mejora la fiabilidad de la estimación.
- **Modelo oculto de Markov**: También conocido como HMM, fue descripto por Baum y Petrie (Baum y Petrie 1966). Para comprenderlo, es necesario explicar que es una cadena de Markov. Una cadena de Markov es un proceso aleatorio que tiene un conjunto de estados (espacio de estados) y consiste en las transiciones de un estado a otro, donde la probabilidad de pasar de un estado a otro depende únicamente del estado en que se encuentra el proceso y no de los estados por los que ha pasado anteriormente (propiedad de Markov). En otras palabras, es una sucesión de variables aleatorias (X_1, X_2,.. X_n) que satisface la propiedad de Markov:

$$\mathbf{Pr}(\boldsymbol{X_{n+1} = x} \mid \boldsymbol{X_1 = x_1, X_2 = x_2, .., X_n = x_n}) = \mathbf{Pr}(\boldsymbol{X_{n+1} = x} \mid \boldsymbol{X_n = x_n}) \quad (1)$$

Los valores que toman las variables aleatorias son los estados de la cadena. Las cadenas de Markov se las suele representar como un autómata finito probabilista. Cabe agregar que la suma de las probabilidades de todas las posibles transiciones a realizar desde el estado actual es igual a 1. El modelo HMM asume que se está modelando una cadena de Markov donde los estados son desconocidos (están ocultos, de ahí el nombre). En la tarea de etiquetado, el conjunto de etiquetas son los estados ocultos y las palabras de la frase a etiquetar son el conjunto de observaciones y el objetivo del modelo HMM es encontrar la sucesión de etiquetas que maximice la probabilidad de encontrar esa sucesión de palabras asumiendo que:

  - La probabilidad de ocurrencia de una palabra es independiente de las palabras que se encuentran antes y después de esta.

- La probabilidad de una etiqueta solo depende de la etiqueta inmediatamente anterior (propiedad de Markov).

2.4 STEMMING

El stemming es un método para obtener las raíces de las palabras, por ejemplo la raíz de “comiendo” es “com”. Entre los métodos automáticos, se destacan:

1. **Búsqueda en tabla**: Fue ideado por Lovins (Lovins, 1968) y consiste en tener en tabla todas las combinaciones de posibles sufijos.
   - Ventajas:
     - Es sencillo
     - Usando un árbol-B o una dispersión estas búsquedas serán muy rápidas.
   - Desventajas
     - Construcción de la tabla, especialmente para el español, ya que muchas palabras presentan muchas formas morfológicas.
     - Difícil para palabras específicas a un dominio.
     - Las tablas pueden llegar a ser muy extensas.
     - No existen tablas estándar y si existieran dada la extensión de la tabla, sería para ciertas temáticas, por lo tanto, para nuevas temáticas, habría que generar nuevas tablas.

Para visualizar más las desventajas, se presentará como sería la tabla para el término “presentar”, la cual tiene 25 formas morfológicas:

Tabla II: 25 formas morfológicas del termino presentar

Palabra	Combinaciones de sufijos
Presentable	able
Presentables	ables
Presentación	ación
Presentaciones	aciones
Presentado	ado
Presentador	ador

Palabra	Combinaciones de sufijos
Presentadores	adores
Presentándomos	ándonos
Presentar	ar
Presentáramos	áramos
Presentaríamos	aríamos
Presentarla	arla
Presentarlas	arlas
Presentarle	arle
Presentarles	arles
Presentarlo	arlo
Presentarlos	arlos
Presentarse	arse
Presentase	ase
Presentásemos	ásemos
Presente	e
Presentémonos	émonos
Presentismo	ismo
Presento	o

Fuente: Panessi y Bordignon, 2011, página 73

2. **Eliminación de afijos**: Este tipo de algoritmos elimina los sufijos y/o prefijos de las palabras obteniéndose la raíz de la misma, por ejemplo se usa para eliminar los plurales. El algoritmo más utilizado es el Porter (Porter, 1980) que solo elimina sufijos. En general solo se eliminan los sufijos, dado que hay más sufijos que prefijos y es más sencillo.
   - Ventajas:
     - Con una cantidad pequeña de reglas se puede obtener gran eficiencia
     - Ante una nueva palabra se puede sacar su raíz fácilmente
   - Desventajas
     - Dependen del idioma
     - El conjunto de reglas impacta directamente en la calidad del stemmer.

- Se debe construir la tabla de reglas

3. **Variedad de sucesores**: Consiste en agrupar palabras que tengan la misma raíz, por ejemplo, todas las palabras de la tabla anterior (Tabla II) estarían en el mismo grupo y se eliminarían los sufijos. La variedad de sucesores se refiere a la cantidad de caracteres diferentes que forman diferentes sufijos ya que cada uno va a tener una longitud diferente.
   Una vez que se tiene la variedad de sucesores para una determinada palabra se debe definir el método para segmentar la palabra. Como se describe en el sitio Lematización (Universidaddad de Costa Rica) los métodos pueden ser 4:
   - **Método del valor de corte**: Se selecciona un valor de corte para las variedades de sucesores y se identifica un límite cada vez que se alcanza ese valor de corte.
   - **Método de los picos y valles**: Se hace el corte de segmento después de los caracteres cuya variedad de sucesores excede a la del carácter que los precede y a la del que lo sigue.
   - **Método de palabra completa**: Se hace el corte después de un segmento si éste es una palabra completa en el corpus.
   - **Método de la entropía**: Aprovecha la distribución de las variedades de sucesores. Usando su ecuación se calculan las entropías de una palabra, se selecciona un valor de corte y se identifican los límites de segmento cuando se supera este valor de corte.

- Ventajas:
  - Sencillo
  - Permite la lematización de forma automática.
- Desventajas:
  - Solo elimina sufijos de las palabras
  - Un valor de corte pequeño provoca cortes incorrectos y con valores de corte grandes se pierden cortes correctos

2.5 HERRAMIENTAS

2.2.4.1. Babelnet

Babelnet es un diccionario enciclopédico multilingüe muy amplio y una red semántica que une conceptos y entidades nombradas en una red de relaciones

semánticas, creados mediante la integración de Wikipedia y Wordnet, además de otros recursos léxicos como Wiktionary, OmegaWIki, Wikidata y Open Multilingual Wordnet e integrado mediante un algoritmo de enlazado automático, completado léxicamente mediante traducciones automáticas. Está compuesto por cerca de 14 millones de entradas, denominadas synsets. Los synsets (también llamados BabelSynsets) agrupan palabras que son sinónimos y representan un mismo concepto o entidad.

Babelnet ofrece dos formas para integrarla con las aplicaciones, la primera ofreciendo acceso Resource Description Framework (RDF) a la Linguistic Linked Open Data cloud; la otra es descargar la enciclopedia completa y realizar consultas de forma local. Las principales clases de la API de Babelnet son:

- **Babelnet**: Es el punto de entrada a todo el contenido de Babelnet. Implementa el patrón singleton para asegurar que haya una sola instancia de la misma.
- **BabelSynset**: Como se ha mencionado anteriormente, representa un conjunto de lexicalizaciones multilingües que son sinónimos y expresan un concepto o entidad. Se lo puede ver como un contenedor de BabelSenses. Está compuesto principalmente por:
  - BabelSenses
  - BabelPOS: Los posibles valores son NOUN, ADJECTIVE, VERB, ADVERB, INTERJECTION, PREPOSITION, ARTICLE, DETERMINER, CONJUNCTION, PRONOUN
  - BabelGloss: Es una definición del concepto en un determinado idioma.
  - BabelExample: Es una frase de ejemplo del significado del synset
  - BabelImage: Es una imagen que representa el concepto
  - BabelSynsetIDRelation: Permite conectar semánticamente un synset con otro.
- **BabelSense**: Es una palabra o expresión en una lengua determinada asociada a un cierto synset, en consecuencia, cada ocurrencia de un BabelSense en diferentes synsets es un BabelSense diferente. Está compuesto principalmente por:
  - BabelSynset: Representa el synset al que pertenece el BabelSense
  - BabelPOS: Los posibles valores son NOUN, ADJECTIVE, VERB, ADVERB, INTERJECTION, PREPOSITION, ARTICLE, DETERMINER, CONJUNCTION, PRONOUN

- Lema: La lexicalización del sentido
- BabelSensePhonetics: Son las pronunciaciones escritas y de audio del sentido.
- BabelSenseSource: Es la fuente del sentido, por ejemplo Wikipedia. Wordnet, Wiktionary u otro de los mencionados anteriormente.

Toda la información expresada en esta sección fue obtenida de la página de Babelnet.

2.2.4.2. WordNet

Como mencionan Miller, Beckwith, Fellbaum, Gross, Derek y Miller, Katherine J. (Miller et. al, 1990), WordNet es una base de datos léxica de Ingles que contiene synsets (agrupaciones de sinónimos cognitivos de sustantivos, verbos, adjetivos y adverbios). La cantidad de palabras, synsets y pares de palabras es la siguiente:

Tabla III: Cantidad de información de WordNet

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Sustantivos	117798	82115	146312
Verbos	11529	13767	25047
Adjetivos	21479	18156	30002
Adverbios	4481	3621	5580
Totales	155287	117659	206941

Fuente: *WNSTATS(7WN)*

Al igual que Babelnet, cada synset representa conceptos distintos, cada uno está formado por un conjunto de palabras que son sinónimos (si una palabra tiene varios significados va a pertenecer a varios synsets) y por la definición de glosses (definición y/o frases de ejemplo). Dentro de un synset, las palabras están ordenadas de acuerdo a la frecuencia de uso. Los synsets están vinculados entre sí por medio de relaciones semánticas, conceptuales o léxicas. De acuerdo a Miller, Beckwith, Fellbaum, Gross, Derek y Miller, Katherine J. (Miller et. al, 1990), entre las relaciones encontramos las siguientes:

- **Sinónimos**: Dentro de la misma categoría sintáctica, una palabra es sinónima de otra si se pueden sustituirse en un mismo contexto, por ejemplo “Se edificó su propia casa” y “Se construyó su propia casa”.
- **Antónimos**: Una palabra es antónima de otra si tienen significados opuestos, por ejemplo claro/oscuro, alto/bajo.
- **Hiponimia**: Es una relación entre significados de las palabras, se dan únicamente por los significados de las palabras. “X” es un hiponimo de “Y” y “X” es un tipo de “Y”. Por ejemplo: Girasol es un hiponimo de flor y girasol es un tipo de flor.
- **Hiperonimia**: Es la relación inversa de la hiponimia: “Y” es un hiperonimo de “X” si “X” es un tipo de “Y”: Flor es un hiperonimo de girasol, y girasol es un tipo de flor.
- **Meronimia**: Una palabra “X” es meronima de “Y” si “X” es una parte de “Y”, por ejemplo: embrague y frenos son meronimos de automovil, ya que embrague y frenos son partes del automóvil.
- **Holonimia**: Esta relación es la inversa de la meronimia: “Y” es un holonimo de “X” si “X es una parte de “Y”, siguiendo el ejemplo de la meronimia, automóvil es holonimo de embrague y frenos ya que los mismos son partes del automóvil.
- **Troponimia**: Relaciona verbos y es el equivalente de la relación de hiponimia para los nombres.
- **Entailment**: En esta relación un término implica al otro. Por ejemplo, divorcio/matrimonio, graduarse/estudiar.

2.2.4.3. Lexicones emocionales

2.2.4.3.1. General Inquirer

Es primer recurso lingüístico, diseñado por Stone en Hardvard en 1966 (Stone et. Al, 1966). Su versión original contiene 1915 palabras que denotan positividad y 2291 que denotan negatividad. Adicionalmente diferencia entre subcategorias de palabras positivas y negativas, por ejemplo Fuerte vs. Debil, Activa vs. Pasiva, Placer vs. Dolor. Las principales desventajas son:

- no contempla múltiples conceptos de los términos,

- solo tiene unigramas (una sola palabra), provocando que haya conceptos que no son contemplados, ya que algunos solo pueden ser expresados mediante términos de más de una palabra.

Por lo mencionado, es útil para combinar con otros recursos lingüísticos, pero no como único recurso. Está en inglés y es gratis para su uso en investigación.

2.2.4.3.2. SentiWordNet

Es un recurso lingüístico diseñado especialmente para la minería de opiniones, descrito en detalle por Denecke (Denecke, 2008) y Baccianella, Esuli y Sebastiani (Baccianella, Esuli y Sebastiani, 2010). Para su desarrollo se han tomado todos los synsets de WordNet 3.0 y se le han anotado un grado de positividad, negatividad y objetividad (o neutralidad), pudiendo tener un synset valores positivos y negativos La forma en que se han anotado los synsets es mediante un algoritmo semiautomático. A grandes rasgos, el algoritmo consiste en:

1. Expandir un léxico semilla anotado a través de las relaciones de WordNet.
2. Entrenar un conjunto de clasificadores con diferentes algoritmos y conjunto de entrenamiento para que anote en “Positiva”, “Negativa” y “Objetiva”
3. Utilizando el clasificador, anotar las palabras que se vayan leyendo a través de un corpus.

Recientemente se ha añadido otro algoritmo para la redefinición de puntuaciones, las mismas son supervisadas por humanos.

Se encuentra en inglés y se distribuye bajo una licencia: "ShareAlike" de Creative Commons, que permite su uso comercial siempre y cuando se mencione a los autores.

2.2.4.3.3. Linguistic Inquiry and Word Count (LIWC)

Es un recurso lingüístico creado por James W. Pennebaker, Roger J. Booth, y Martha E. Francis que se encuentra descrito en Tausczik, 2009. Tiene 2300 palabras y más de 70 clases (o categorías). Es un recurso que sirve para calcular el porcentaje de palabras que denotan emociones positivas o negativas en un determinado texto, para lograr esto incluye diccionarios para contar palabras que pertenecen a una determinada categoría de significado psicológico. Entre las categorías más relevantes se encuentran “Procesos afectivos” que tiene como

subcategorias “Emociones positivas” y “Emociones negativas”; “Procesos cognitivos” que incluye como subcategorias “Tentativo” o “Inhibición”

El principal valor agregado de este recurso es que la investigación fue realizada por psicologos que tienen en cuenta el uso de pronombres de primera persona en singular que implican que la persona que escribe la opinión tiene un determinado estado psicológico (sea depresión, alegría, etc).

Ventajas:

- Asegura una alta precisión en el espectro cubierto por su contenido ya que está creado por psicólogos
- Tiene clasificaciones complejas
- Soporta varios idiomas, entre ellos el español

Desventajas:

- No es gratuito, tiene dos versiones con costos diferentes.
- Al igual que el General Inquirer solo tiene unigramas

2.2.4.3.4. MPQA Subjectivity Cues Lexicon

El Multi-Perspective Question Answering (MPQA) es un lexicón creado por Theresa Wilson, Janyce Wiebe, y Paul Hoffmann en 2005 que contiene una lista de términos que denotan subjetividad. De acuerdo a Wilson, Wiebe y Hoofman (Wilson, Wiebe y Hoofman, 2005), contiene 2718 palabras en la categoría “positivas” y 4912 en la categoría “negativas”. Para cada término aporta:

- La polaridad, es decir si es positiva o negativa
- La categoría gramatical del termino
- La intensidad de la subjetividad, que puede ser “fuerte” o “débil”
- La longitud de las palabras de la expresión
- Si el término es representado por su raíz o en su totalidad. Si está representado por su raíz, todos los términos que tengan esa raíz tienen la polaridad indicada en ese recurso para esa raíz.

El presente lexicón junto con los distintos corpus está en inglés y es posible usarlos para investigación, ya que se distribuyen bajo licencia GNU GPL.

2.2.4.3.5. EffectWordNet

Al igual que SentiWordNet, es una expansión de Wordnet. Como mencionan Choi y Wiebe (Choi y Wiebe, 2014), para crearlo se anotaron todos los synsets de WordNet con etiqueta positiva y negativa, luego se construyó un grafo a partir de un léxico base, utilizando las relaciones entre synsets del WordNet y se agregaron de forma semisupervisada etiquetas positivas y negativas a otros términos. Para diferentes conceptos de una misma palabra puede haber una etiqueta diferente. A los términos que no tengan polaridad se les añade una etiqueta null. Posee 3298 sentidos de términos con etiqueta positiva, 2427 con etiqueta negativa y 5296 con etiqueta nula. Cada sentido tiene adjuntado una frase que expresa ese sentido junto con una lista de significantes que expresan el significado. Es posible usarlos para investigación, ya que se distribuyen bajo licencia GNU GPL.

2.2.4.3.6. Emotinet

Ibañez, Serrano y García (Ibañez, Serrano y García, 2009) y Balahur (Balahur, 2011) la describen como una ontología diseñada para detectar emociones basadas en el conocimiento de sentido común que representa:

- el sentido común que se conoce de los conceptos,
- la interacción entre los conceptos
- la consecuencia afectiva provocada por la interacción
- las acciones realizadas por un agente que puede repercutir en una emoción
- relaciones entre emociones

Es útil para un modelo de análisis de opiniones y detección de emociones pero no para la polarización de las mismas ya que no aporta información relativa a la polaridad de los términos.

2.2.4.3.7. Bing Liu Opinion Lexicon

De acuerdo a lo descrito en Ding, Liu y Yu (Ding, Liu y Yu, 2008), consiste una en una lista de términos que tiene 2006 términos en la categoría “positivas” y 4783 en la categoría “negativas”. Además, incluye términos expresados incorrectamente y variantes morfológicas de los mismos.

Entre las desventajas, se encuentra que al igual que el General Inquirer solo tiene unigramas y que no contempla múltiples conceptos de los términos.

Es útil para los modelos cuantitativos por su sencillez, y la forma en que resuelven la desventaja que no contempla múltiples conceptos de los términos, esto lo hacen encontrando dependencias entre la aparición de un término léxico polarizado y otro considerando un conjunto suficientemente grande de textos anotados. Se encuentra en inglés.

2.2.4.3.8. Comparación de lexicones emocionales

Para finalizar la sección de lexicones emocionales, es útil hacer una comparación entre los mismos. Potts (Potts, 2001) los compraró y se encontró que:

- hay coincidencias entre los lexicones,
- no todos tienen la misma polaridad para una misma palabra

Los niveles de desacuerdo encontrados fueron:

Tabla IV: Niveles de desacuerdo entre lexicones

	MPQA	Opinion Lexicon	Inquirer	SentiWordNet	LIWC
MPQA	–	33/5402 (0.6%)	49/2867 (2%)	1127/4214 (27%)	12/363 (3%)
Opinion Lexicon		–	32/2411 (1%)	1004/3994 (25%)	9/403 (2%)
Inquirer			–	520/2306 (23%)	1/204 (0.5%)
SentiWordNet				–	174/694 (25%)
LIWC					–

Fuente: Sentiment Symposium Tutorial: Lexicons

Como se puede observar, exceptuando SentiWordnet que es con el que mayor nivel de desacuerdo se da, no hay grandes diferencias entre los lexicones. El estudio realizado por Potts consistió en contar cuantas veces aparecía la palabra “bad” en críticas de cinematográficas en el sitio web IMDB discriminando por la cantidad de estrellas de la crítica (entre 1 y 10). El resultado obtenido se resume en el siguiente gráfico:

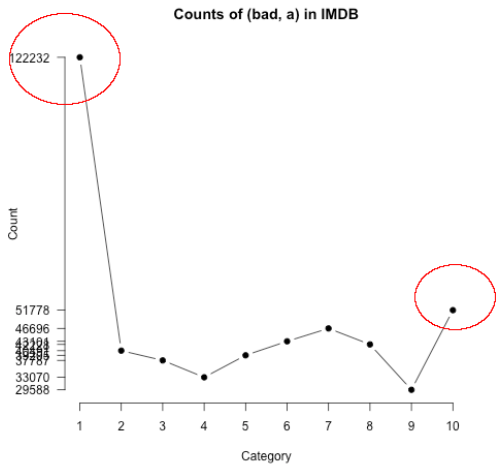

Figura 1: Cantidad de veces que aparece la palabra bad por categoría

Fuente: Sentiment Symposium Tutorial: Lexicons

Se pueden observar principalmente dos cosas:

- Lo esperable: las críticas de una estrella es donde aparece la palabra “bad” más veces.
- Lo inesperado: La cantidad que veces que aparece “bad” va disminuyendo hasta las 4 estrellas, luego sube y vuelve a bajar hasta su punto mínimo en las criticas de 9 estrellas, pero lo más sorprendente es que la segunda mayor cantidad de apariciones de la palabra se da en las críticas con 10 estrellas.

Para solucionar el problema descripto en lo inesperado, utilizó el estimador de máxima verosimilitud, es decir que dividió los totales por la cantidad de palabras en dicha categoría y como resultado, tomando las frecuencias relativas obtuvo el siguiente resultado:

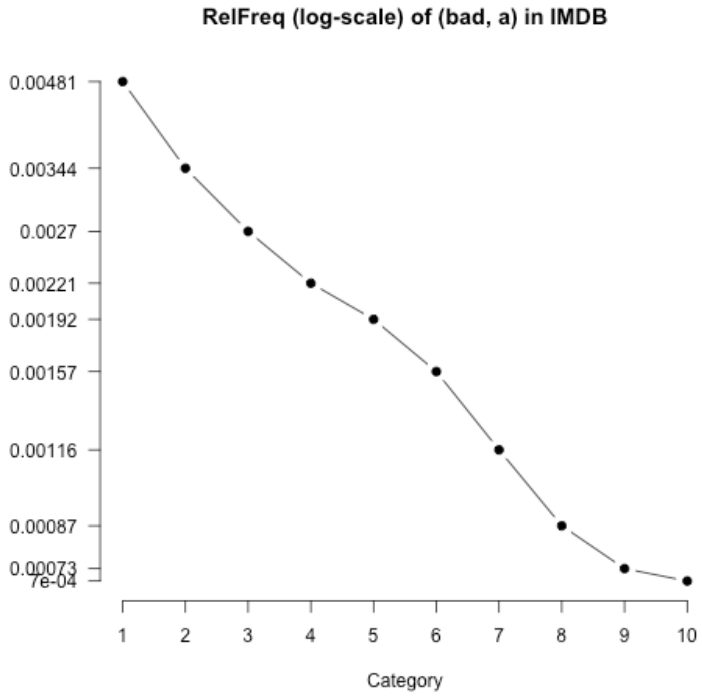

Figura 2 Frecuencia de la palabra bad por categoría

Fuente: Sentiment Symposium Tutorial: Lexicons

Potts identificó que no sólo los adjetivos dan información sobre la polaridad de un texto, sino que hay expresiones más complejas o palabras que parecen irrelevantes que en realidad pueden ser significativas, por ejemplo, encontró que las negaciones “no”, “not”, “n't”, “never” en las criticas cinematográficas de IMDB son mucho más frecuentes en críticas con pocas estrellas que en las de muchas.

En conclusión y para finalizar, las diferencias entre los lexicones están dados por cómo fueron consideradas todas las palabras y en el contexto que fueron tomadas, el estudio realizado por Potts deja en evidencia esto ya que la palabra “bad” fue encontrada tanto en términos negativos como positivos, es decir que al crearse los lexicones, en alguno puede tomarse una palabra como muy negativa (o muy positiva) y en otros no tanto.

2.6 MÉTODOS EXISTENTES

Para determinar la polaridad de una opinión existen principalmente tres enfoques, haciendo que haya 3 tipos de algoritmos.

- Algoritmos de clasificación basados en reglas Ad-Hoc
- Algoritmos de clasificación a través de aprendizaje supervisado
- Algoritmos de clasificación a través de aprendizaje no supervisado

2.6.1 Algoritmos de clasificación basados en reglas Ad-Hoc

El método consiste en escribir manualmente un conjunto de reglas que determinen la clase a la que pertenece una determinada opinión, en este caso, positiva o negativa. Es un método simple que si se refinan las reglas puede obtenerse una buena precisión. Las principales desventajas del método son que se necesita un conocimiento profundo del dominio, es difícil de mantener ya que la cantidad de reglas puede ser muy alto y es poco escalable.

2.6.2 Algoritmos de clasificación a través de aprendizaje supervisado

Se pueden reconocer dos fases en estos algoritmos: en la primera fase se tienen dos conjuntos; el de muestras ya clasificadas para construir el clasificador, denominado de entrenamiento y el de muestras a clasificar, denominado de test o validación. El objetivo de esta fase es construir un modelo o regla para poder clasificar las opiniones. La segunda fase consiste en clasificar las muestras de las que se desconoce la clase.

Para el caso de la polarización de opiniones, las clases van a ser dos: positivas y negativas. Los algoritmos más utilizados son Naive-Bayes y Support Vector Machines (SVM).

2.3.2.1. Naive-Bayes

Este clasificador está basado, como no podía ser de otra manera, en el Teorema de Bayes (Bayes y Price, 1763). Recordemos que dicho teorema permite calcular la probabilidad de un suceso habiendo sucedido otro que influye en el anterior, la fórmula que representa esto es:

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \quad (2)$$

Ahora, cuando se habla del clasificador, lo que se intenta decir es que se desea determinar la clase a la que pertenece un ejemplo X, para poder clasificar utilizando este teorema, debemos asumir que:

- El ejemplo X está representado por k valores de atributos, siendo el conjunto de valores $\{a_1, a_2, a_3, \ldots, a_k\}$
- Los valores de los atributo de los ejemplos son condicionalmente independientes dado el valor de la clase.
- Hay un conjunto finito de clases (V) integrado por j clases, $V = \{v_1, v_2, v_3, \ldots, v_j$

La clase a la que pertenece el ejemplo X va a estar determinada por:

$$v_{max} = argmax_{v_j \in V} \left(P(v_j | a_1, \ldots, a_n) \right) (3)$$

Aplicando Bayes:

$$v_{max} = argmax_{v_j \in V} \left(\frac{P(a_1, \ldots, a_n | v_j) * P(v_j)}{P(a_1, \ldots, a_n)} \right) (4)$$

Como los valores de atributo son condicionalmente independientes, podemos decir que:

$$P(a_1, \ldots, a_n | v_j) = \prod_i P(a_i | v_j) (5)$$

Reemplazando 3 en 4, queda que:

$$v_{max} = argmax_{v_j \in V} \left(P(v_j) \prod_i P(a_i | v_j) \right) (6)$$

Para su uso hay que tener en cuenta que se puede aplicar cuando:

- Se dispone de conjuntos de entrenamiento mediano o grande.
- Los valores de los atributo de los ejemplos son condicionalmente independientes dado el valor de la clase.

2.3.2.2. Support Vector Machines

Antes de comenzar a explicar concretamente SVM es necesario mencionar la representación matemática realizada por Schölkopf (Schölkopf, 1999) de la clasificación con múltiples atributos, entendiéndose como múltiples atributos los valores que toman los diferentes atributos de los elementos a clasificar (asignarles una clase). La representación del problema de clasificación con múltiples atributos se puede expresar como:

"Se quiere estimar una función de decisión:

$$f: R_n \rightarrow \{\pm 1\} (7)$$

empleando elementos de un conjunto de entrenamiento para entregar a la "máquina de clasificación", entendiéndose como máquina de clasificación por ejemplo, una red neuronal, algoritmo genético, etc."

Los elementos del conjunto de entrenamiento se los identifica como $\{ x_i \}$, $x_i \in R^n$, $i \in \{ 1 ,... l \}$ y se los genera utilizando una distribución de probabilidad desconocida P(x, y) siendo y_i el conjunto de etiquetas o clases asociadas. La salida de la clasificación sería una determinada etiqueta “y” para un x_i, entonces el conjunto de datos considerados sería:

$$(x_1 , y_1) ,..., (x_l , y_l) \in R^n \times \{ \pm 1 \} \quad (8)$$

El objetivo es que la función f clasifique correctamente (f(x)=y) ejemplos generados por la misma distribución de probabilidad (P(x, y)) que la utilizada para generar el conjunto de entrenamiento.

El error de entrenamiento se define como:

$$R_{ent}[\alpha] = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{2}|f(x_i,\alpha) - y_i| \quad (9)$$

El riesgo o error de test esperado para una máquina de entrenamiento se define como:

$$R[\alpha] = \int \frac{1}{2}|f(x,\alpha) - y| dP(x,y) \quad (10)$$

La pérdida se encuentra representada por:

$$\frac{1}{2}|f(x,\alpha) - y_i| \quad (11)$$

SVM fue inicialmente fue diseñada para clasificar entre dos clases (clasificador binario). Si bien el método se puede utilizar para K clases, dado que el presente trabajo apunta a la clasificación de Positiva – Negativa el método se explicará con dos clases, es decir cómo se definió inicialmente. A diferencia del enfoque estadístico, en el que se asume que los datos son generados por una distribución de probabilidad que nos es desconocida y a partir de la cual diseñamos el clasificador, el enfoque propuesto por Vapnik y Chervonenkis (Vapnik y Chervonenkis, 1974) apunta a diseñar el clasificador durante el entrenamiento, es decir a partir de los datos del corpus de entrenamiento, mediante determinados algoritmos basados en la Teoría del Aprendizaje. A grandes rasgos, en el entrenamiento se aprende o establecen parámetros desconocidos del modelo: a partir de los datos de entrenamiento el algoritmo de aprendizaje selecciona la función de decisión que mejor se ajuste a las entradas y salidas esperadas entre las todas las posibles.

La idea de SVM consiste en separar los datos en el espacio de entrada con un hiperplano lineal y cuando no es posible hacerlo se traslada los vectores de entrada a un nuevo espacio de dimensión más alta (sus propiedades garantizaran la

alta generalización de la máquina de aprendizaje, principalmente debido a la maximización del margen de separación entre los vectores de las dos clases) mediante una aplicación no lineal. Una vez que se tiene el hiperplano, los puntos que se encuentran de un lado del hiperplano van a ser etiquetados con una clase y los que están del otro lado con otra, en otras palabras, el hiperplano va a separar los puntos que pertenecen a una clase de los puntos que pertenecen a otra.

Para la polarización de opiniones el conjunto de entrenamiento estaría formado por opiniones, los cuales están expresadas mediante los vectores $\{ x_i \}$, $i \in \{ 1 ,... l \}$ y el conjunto de etiquetas posibles, positiva y negativa, están expresadas como $y_i \in \{ -1 , 1 \}$, el -1 representando las negativas y el 1 las positivas. Asumiendo que utilizando un hiperplano lineal es posible separar el conjunto de entrenamiento, los puntos x que pertenecen al hiperplano satisfacen la ecuación:

$$w \cdot x + b = 0 \quad (12)$$

Siendo: w un vector normal al hiperplano

$|| \cdot ||$ la norma euclidea

b/w la distancia perpendicular del hiperplano al origen

Durante el aprendizaje lo que se busca es el hiperplano óptimo entre todos los hiperplanos capaces de separar datos. Se conoce como hiperplano óptimo a aquel que es capaz de separar los puntos con el mayor margen de separación entre cada elemento del conjunto de entrenamiento y el hiperplano. Siguiendo esta idea, Vapnik y Chervonenkis (Vapnik y Chervonenkis, 1974) diseñaron un algoritmo de aprendizaje para datos separables que plantea resolver el siguiente problema:

$$\text{“}\textit{Encontrar } w \in \mathcal{R}^n \; y \; b \in \mathcal{R}$$

$$\textit{que minimicen } \tau(w) = \frac{1}{2} \|w\|^2$$

$$\textit{sujeto a } y_i(w^t \cdot x_i + b) \geq 1 \; \forall \; i = 1, \dots, l \text{“}$$

Una vez obtenidos w y b es posible realizar la clasificación de la opinión utilizando el signo de (11), es decir que se clasifica el elemento de entrenamiento x con la formula $\text{sign}(w^t \cdot x_i + b)$ y el error de clasificación va a estar dado por $R_{ent}(w,b)$. Por otro lado, de la inecuación $y_i(w^t \cdot x_i + b) \geq 1$ se pueden obtener los vectores soporte, que van estar formados por los puntos que verifican la igualdad en la misma y están representados por las ecuaciones $w^t \cdot x_i + b = -1$ y $w^t \cdot x_i + b = 1$. Estos vectores, pertenecerán a uno de los dos posibles hiperplanos óptimos de separación.

Si se intenta aplicar el algoritmo para datos no separables la función objetivo crece excesivamente. Para resolverlo se incorporan variables nuevas de valor pequeño. El problema queda de la siguiente forma:

$$\text{"Encontrar } w \in \mathcal{R}^n \quad b \in \mathcal{R} \; y \; \varepsilon_i i = 1, .., l$$

$$\text{que minimicen } \tau(w) = \frac{1}{2} \|w\|^2 + C \sum_{l=1}^{l} \varepsilon_i$$

$$\text{sujeto a } y_i(w^t \cdot x_i + b) \geq 1 - \varepsilon_i \; y \; \varepsilon_i > 0 \; \forall \; i = 1, \ldots, l \text{ "}$$

siendo C un parámetro que será estimado por el clasificador.

Para la SVN no lineal se proyectan las variables de entrada en un espacio mayor (generalmente de dimensión infinita) que al que pertenecían para luego aplicar SVM en el nuevo espacio, comúnmente conocido como espacio de características. Así, la SVN no lineal es capaz de separar los elementos del conjunto de entrenamiento con una probabilidad de error dada por $R_{ent}(w,b)$. La proyección se realiza aplicando funciones kernel, que pueden ser Polinomial-homogénea, Perceptron, Función de base radial Gaussiana o Sigmoid

2.6.3 Algoritmos de clasificación a través de aprendizaje no supervisado

Al igual que los algoritmos supervisados, se utiliza un conjunto de entrenamiento para entrenar al clasificador, pero la diferencia entre un conjunto de entrenamiento y otro, es que para los algoritmos supervisados, el conjunto de entrenamiento tiene los datos y los resultados correctos (clase a la que pertenece), mientras que en los algoritmos no supervisados no se tienen los resultados correctos. El entrenamiento consiste en encontrar similitudes, patrones o estructuras entre los datos que permitan agruparlos sin tener conocimiento de la clase a la que pertenecen ya que como se mencionó, el conjunto de entrenamiento no brinda esa información.

Dado que hay diversos trabajos que optan por una clasificación con aprendizaje no supervisado, se han seleccionado algunos de los más representativos.

2.3.3.1. Clasificación basada en léxicos de Opinión

Los clasificadores de este grupo se basan en tener términos semilla cuya polaridad es conocida. La clasificación se realiza calculando la distancia o similitud

de los términos de la opinión respecto a las semillas. En este grupo se pueden mencionar principalmente dos trabajos:

- Sidorov, Miranda-Jiménez, Viveros-Jiménez, Gelbukh, Castro-Sánchez, Velásquez, Díaz-Rangel, Suárez-Guerra, Treviño y Gordon (Sidorov et al., 2012) definieron un léxico de más de 2000 palabras en español, un conjunto de emociones positivas y negativas (alegría, ira, miedo, tristeza, sorpresa y disgusto) y la distancia de cada palabra del léxico a cada emoción. El algoritmo que calcule la polaridad de una frase calcularía la distancia de cada término a las emociones definidas.
- Turney (Turney, 2002): Es el caso extremo de este grupo, dónde utiliza solo dos términos como semillas, "excellent" y "poor". Para clasificar los textos calcula la distancia de cada término a las semillas. Esta distancia representa la orientación semántica de la palabra que se está analizando. El algoritmo que propone consiste en 3 pasos:
    1. Se utiliza una herramienta de POS Tagging y se extraen las frases del documento que tengan adjetivos o adverbios.
    2. Estimar la distancia (orientación semántica) de cada frase extraída utilizando el algoritmo Pointwise Mutual Information - Information Retrieval (PMI-IR)
    3. Asignar las clases "recomendado" o "no recomendado" al documento en base al cálculo de la distancia promedio de todas las frases extraídas.

El algoritmo PMI-IR consiste en calcular la distancia u orientación semántica mediante la ecuación:

$$PMI(t_1, t_2) = log_2\left(\frac{p(t_1 \wedge t_2)}{p(t_1) \wedge p(t_2)}\right) \quad (13)$$

Siendo $p(t_1 \wedge t_2)$ la probabilidad de que los términos 1 y 2 (t_1 y t_2 respectivamente) co-ocurran en el corpus de búsqueda. Si los términos t_1 y t_2 son independientes desde el punto de vista estadístico de la ecuación anterior, si modificamos $p(t_1 \wedge t_2)$ por $p(t_1 * t_2)$ el resultado que se obtiene sería el grado de dependencia estadística entre los t_1 y t_2. Para calcular la cantidad de información que se obtiene respecto de una palabra cuando se observa la otra, a la división se le debe calcular el $\log_2$, entonces, en otras

palabras, para determinar cuan seguido aparece t_1 con t_2 se utiliza la siguiente ecuación:

$$PMI(t_1, t_2) = log_2\left(\frac{p(t_1 \wedge t_2)}{p(t_1) * p(t_2)}\right) (14)$$

Uno de los operadores más utilizados para realizar búsquedas en el corpus de búsqueda es el operador "NEAR". El operador "NEAR" es un operador binario que devuelve resultados cuando ambos los términos t_1 y t_2 aparecen a 10 o menos palabras de distancia.

Aplicando el algoritmo al clasificador de Turney (Turney, 2002), t_1 sería una frase del documento y t_2 una semilla (excellent o poor), la ecuación para calcular la orientación semántica (SO) de t_1 sería:

$$SO(t_1) = PMI(t_1,\text{excelent}) - PMI(t_1,\text{poor}) (15)$$

Aplicando lo explicado, la ecuación quedaría de la siguiente manera:

$$SO(t_1) = log_2\left(\frac{hits(t_1 \wedge \text{"excellent"}) * hits(poor)}{hits(t_1 \wedge \text{"poor"}) * hits(excelent)}\right) (16)$$

Siendo la probabilidad de ocurrencia representada por la cantidad de hits en el corpus, en otras palabras, es la cantidad de resultados devuelta por el operador "NEAR" para la consulta t_1 NEAR t_2.

Una frase es positiva si el resultado es positivo y negativa si es negativo.

Adicionalmente y para cerrar la explicación del algoritmo, si la cantidad de semillas tanto positivas como negativas fuera mayor, la ecuación quedaría de la siguiente manera:

$$SO(t_1) = \sum_{p \in posWords} PMI(t_1,\text{p}) - \sum_{n \in negWords} PMI(t_1, n) (17)$$

Siendo p una semilla positiva y n una semilla negativa. Aplicando las operaciones algebraicas explicadas:

$$SO(t_1) = log_2\left(\frac{\prod_{p \in posWords} hits(t_1 \wedge \text{p}) * \prod_{n \in \text{neg}Words} hits(n)}{\prod_{n \in \text{neg}Words} hits(t_1 \wedge \text{n}) * \prod_{p \in posWords} hits(p)}\right) (18)$$

En estas dos formas de determinar la polaridad de la opinión para obtener resultados más precisos es necesario aplicarlo a un dominio en particular o dominios similares, ya que por ejemplo, hay términos que depende del dominio puede ser positivos o negativos, por ejemplo "frío" está asociado a emociones negativas para comidas, pero para bebidas, por ejemplo la cerveza, es positiva o

“añejo” que para los vinos está asociada a emociones positivas mientras que para comida o productos electrónicos está más asociada a emociones negativas.

2.3.3.2. Clasificación basada en regiones

El método propuesto por Kim y Hovy (Kim y Hovy, 2004) parte de la idea de que las opiniones tienen una región que expresan los sentimientos del tópico de la opinión, en otras palabras, no toda la opinión aporta en la determinación de la polaridad y está basado a grandes rasgos en tópicos, entidades y la asignación de un valor positivo y negativo a cada palabra para luego combinar esos valores y así determinar la polaridad de la opinión.

Cada tópico tiene un conjunto de textos acerca de él y se seleccionan las oraciones que tengan expresiones asociadas al tópico y a entidades candidatas. Para cada oración se selecciona la entidad más cercana a la expresión del tópico utilizando un etiquetador de nombres de entidades. Una vez seleccionada la entidad se determina el tamaño de la región, dónde como se mencionó, se encuentra el sentimiento expresado por la entidad. Para determinar el tamaño de la región hay varias estrategias:

1. La oración completa
2. Las palabras que se encuentran entre la entidad y la expresión que define al tópico
3. 2 palabras antes y después de la entidad
4. Las palabras desde la entidad hasta el final de la oración.

Una vez definida la región, se determina la polaridad del sentimiento en la región combinando las polaridades de cada palabra de la región.

Para determinar la polaridad de cada palabra de la región se arma una lista de palabras positivas y otra de palabras negativas. Para construir las listas se utilizan palabras semillas y se expanden utilizando palabras y relaciones de WordNet, con la premisa de que los sinónimos tienen la misma polaridad y los antonimos la polaridad opuesta, por lo tanto la lista de palabras positivas se va a conformar por las semillas, los sinónimos y los antonimos de las palabras de la lista de palabras negativas y la lista de palabras negativas se va a conformar por las semillas, los sinónimos y los antonimos de las palabras de la lista de palabras positivas. Además de las palabras mencionadas, a cada lista se le agrega las palabras extraídas del glosario de WordNet. Una vez armadas estas listas se

eliminan las palabras que pertenecen a las dos listas, con el objetivo de evitar la presencia de palabras ambiguas.

Una vez obtenidas las dos listas de palabras, se procede a calcular la polaridad de cada palabra de la región, para ello se obtienen todos los sinónimos en WordNet y se determina cómo interactúan con la lista de sentimientos, esto se puede expresar de la siguiente forma:

$$argmax_c P(c|w) = argmax_c P(c|syn_1, syn_2 \dots syn_n) \quad (19)$$

siendo c el sentimiento (positivo o negativo), w la palabra a clasificar y los syn son los sinónimos de la palabra obtenidos de WordNet. Existen dos modelos para calcular la ecuación, el primer modelo se deriva de la clasificación documentos y se expresa de la siguiente manera:

$$argmax_c P(c|w) = argmax_c P(c)P(w|c) = argmax_c P(c)P(syn_1, syn_2 \dots syn_n|c) = argmax_c P(c) \prod_{k=1}^{m} P(f_k \vee c)^{count(f_k, synset(w))} \quad (20)$$

siendo:

- f_k la k-ésima palabra la lista de palabras de la clase c (sentimiento positivo o negativo) que a su vez pertenece al conjunto de sinónimos de w.
- count(f_k,synset(w)) la cantidad de ocurrencias de la palabra f_k en el conjunto de sinonimos de w
- P(f_k|c) la división entre la cantidad de veces que aparece f_k en la lista de palabras de la clase c y la cantidad total de palabras de la dicha lista.
- P(c) la división entre la cantidad de palabras de la lista de la clase c y la cantidad de total de palabras de todas las clases.

La idea del segundo método es que cuantos más sinónimos de w se encuentren en la lista de palabras de c es mayor la probabilidad de que la palabra pertenezca a esa clase y se expresa de la siguiente manera:

$$argmax_c P(c|w) = argmax_c P(c)P(w|c) = argmax_c P(c) \frac{\sum_{i=1}^{n} \text{count}(syn_i, c)}{count(c)} \quad (21)$$

siendo:

- count(c) la cantidad de palabras que se encuentran en la lista de palabras de la clase c
- count(syn_i,c) la cantidad de ocurrencias de syn_i en c

Para ambos métodos, la polaridad de la palabra va a ser la que haya dado mayor resultado.

Una vez obtenida la polaridad de cada palabra, se obtiene la polaridad de la región, para ello existen 3 modelos:

1. Considerar sólo las palabras de forma binaria y no el valor obtenido, se expresa de la siguiente manera:

$$polaridad(s) = \prod_{i=1}^{|s|} signo(w_i) \quad (22)$$

siendo s la región, |s| es la cantidad de palabras de la región y signo(w_i) es 1 si la polaridad de la palabra w_i es positiva y -1 si es negativa. En este modelo se incluyen palabras modificadoras de la polaridad y se basa en que los negativos se cancelan unos a otros (esto se puede deducir de la multiplicatoria), por ejemplo si se está negando un hecho negativo la esa parte de la región va a ser positiva.

2. El segundo modelo calcula la media armónica de los valores de cada palabra para determinar la polaridad de la región, se expresa de la siguiente manera

$$P(c|s) = \frac{1}{n(c)} \sum_{i=1}^{n} p(c|w_i) \ , si \ arg_j \max \ p(c_j|w_i) \quad (23)$$

siendo n(c) la cantidad de palabras que pertenecen a la clase c dentro de la región. En este modelo importa tanto la cantidad de palabras de una determinada clase como el valor calculado para cada palabra de la región.

3. El tercer modelo se está basado en la media geométrica:

$$P(c|s) = \ 10^{n(c)-1} \prod_{i=1}^{n} p(c|w_i) \, , si \ arg_j \max \ p(c_j|w_i) = c \quad (24)$$

Resumiendo, éste método aborda tanto la identificación de la entidad que emite la opinión como la región que expresa el sentimiento de la opinión. La mayor deficiencia es que asume que solo hay una entidad y en consecuencia una sola opinión (no siempre es así) debido a que una vez detectado el tópico, sólo utiliza la entidad que se encuentra más cerca y la región asociada al mismo.

2.3.3.3. Clasificación basada en el modelo Spin

El modelo spin pertenece a la teoría física y consiste en un arreglo con N electrones, cada uno tiene un spin que puede tomar valor +1 o -1 que indican una dirección, arriba y abajo respectivamente. Además, se considera que los electrones cercanos tienden a tener el mismo spin por cuestiones energéticas. La configuración mínima se alcanza aplicando iterativamente la función de energía al modelo spin.

Takamura, Inui y Okumura (Takamura, Inui y Okumura, 2005) proponen usar este modelo para determinar la polaridad de las palabras, dónde cada palabra sería un electrón y la polaridad estaría representada por el spin asociado. La representación del modelo es mediante un grafo, dónde las palabras van a estar conectadas si una aparece en el glosario de otras, esto parte de la idea de que la polaridad de las palabras del glosario tiende a ser igual a la polaridad de la palabra a la que pertenece dicho glosario. Para construir el grafo, definen dos conjuntos:

- GL + (t): Conjunto de palabras del glosario de la palabra t, excluyendo las palabras que dependen sintácticamente de la negación
- GL – (t): antónimos de t y las palabras que fueron excluidas de GL + (t), es decir, las palabras sintácticamente dependientes de una negación.

La matriz de adyacencia W = {w_{ij} } se define como:

$$w_{ij} = \begin{cases} 1 \; si \; t_i \epsilon \; GL_+(t_j) \acute{o} \; t_j \epsilon \; GL_+(t_j) \\ -1 \; si \; t_i \epsilon \; GL\_(t_j) \acute{o} \; t_j \epsilon \; GL\_(t_j) \\ 0 \; en \; otro \; caso \end{cases} \quad (25)$$

De modo general, se puede decir que dado un t_1 y t_2, obtienen un peso de w_{12} = 1 si en el glosario de t_1 aparece t_2 o a la inversa, caso contrario obtienen el peso de w_{12} = -1.

A grandes rasgos, para determinar la polaridad de las palabras se aplica la función energía, mencionada cuando se explicó el modelo spin. Para ello definen palabras semillas para transmitir su polaridad a las demás.

Para la construcción del grafo utilizaron el diccionario japonés Iwanami (Nishio et al., 1994) y el sistema de análisis morfológico para el japonés (Matsumoto et.al., 2002). Solo tuvieron en cuenta los sustantivos, adjetivos, verbos, adverbios y algunas palabras de negación. Las palabras que se encuentran precedidas por negaciones fueron consideradas dependientes sintácticamente de la negación y se eliminaron las palabras que no están relacionadas con ninguna otra. El resultado es un grafo de 58185 palabras y al igual que Turney, 2002 utilizaron como semilla las palabras "good" y "bad".

La evaluación del método consistió en anotar manualmente la polaridad de 9790 palabras extraídas del diccionario (2491 positivas, 3141 negativas y 4158 neutras). Dado que el modelo no incluye la polaridad neutra, ésta fue excluida, se realiza una evaluación binaria. Para los sustantivos se obtuvo una precisión de 81,2%, para los verbos 76,2%, para los adjetivos 74,5% y para otras 77,7%. El

principal problema de método es que solo acepta dos valores (1 y -1), quedando excluido el valor neutro.

2.7 CONCLUSIÓN

Como se ha visto lo largo del capítulo para cada etapa del análisis de texto para determinar la polaridad hay diferentes estrategias y herramientas a utilizar, cada una con sus ventajas y desventajas. Si bien hay muchas otras estrategias, se ha intentado hacer un resumen con las estrategias principales para mostrar la variedad de las mismas.

3. DESARROLLO - CAPÍTULO 1

3.1 INTRODUCCIÓN

En el presente capítulo se propondrá un modelo general independiente de la implementación. Dicho modelo tiene como objetivo describir el “qué” del sistema que determine la polaridad de las opiniones.

3.2 DESARROLLO

El método elegido para determinar la polaridad de las opiniones es cuantitativo, basado en el puntaje de cada palabra dentro del contexto en que se encuentra, por lo tanto podemos dividir la polarización en dos etapas:

1. **Preprocesamiento**: El objetivo de esta etapa es analizar la opinión y recopilar todo lo necesario para determinar la polaridad de la opinión. A grandes rasgos, las dos actividades principales son el postagging y la búsqueda de lemas y sinónimos.
2. **Polarización**: El objetivo de esta etapa, como su nombre lo indica, es determinar la polaridad de la opinión utilizando la opinión y todo lo obtenido en la etapa anterior obteniendo un puntaje para cada palabra y aplicando reglas para resolver las negaciones, intensificadores, modalidades y sarcasmo.

El modelo propuesto para determinar polaridad de las opiniones, teniendo en cuenta las etapas mencionadas, es el siguiente:

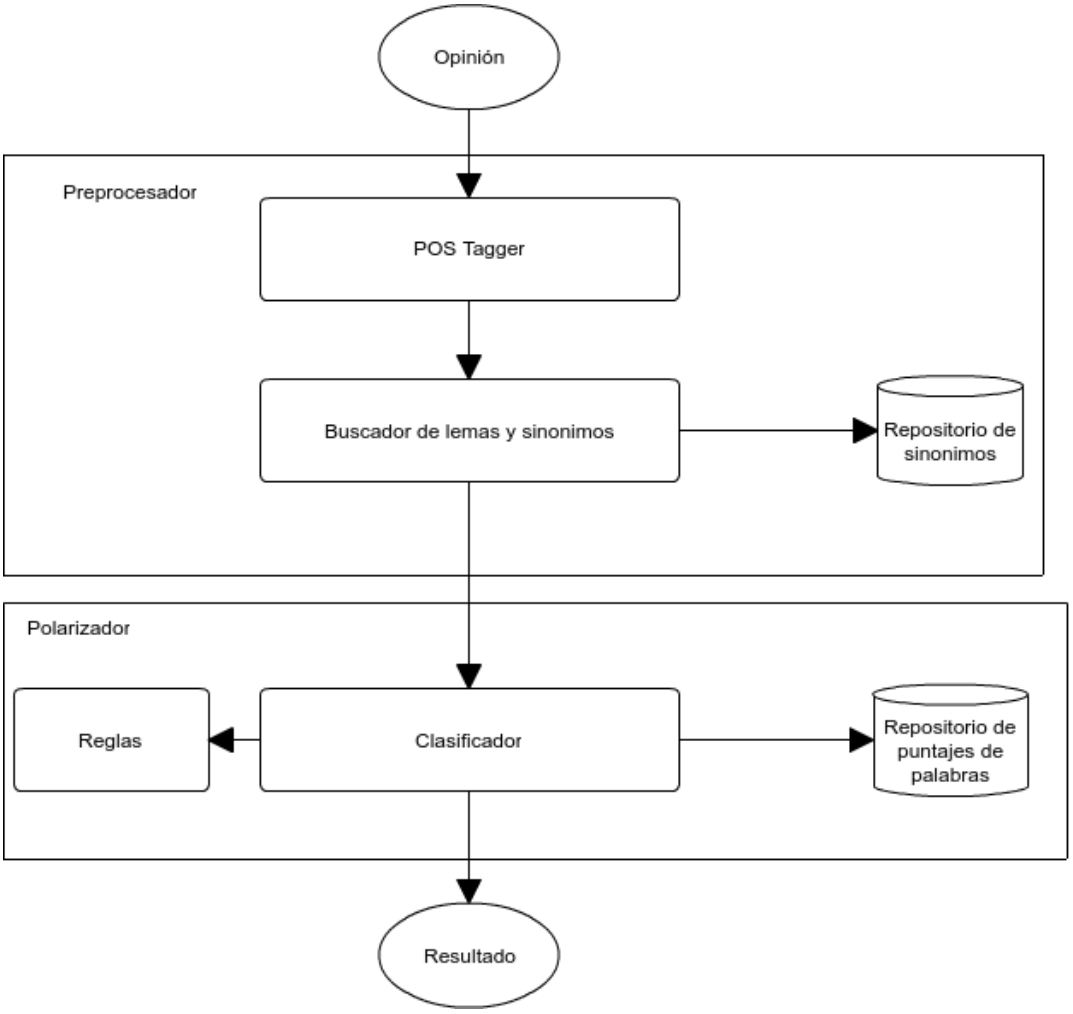

Figura 3: Modelo propuesto para determinar la polaridad

Los elementos se pueden clasificar en dos tipos: módulos (representados por rectángulos) y repositorios (representados por cilindros); finalmente, las elipses representan las entradas y salidas. A continuación se dará una breve explicación de cada elemento:

- **POS Tagger**: Su objetivo, como su nombre lo indica y de acuerdo a lo descripto anteriormente, es determinar la categoría léxica de cada palabra de la opinión teniendo en cuenta el contexto de cada palabra dentro de la misma.
- **Buscador de lemas y sinónimos**: Su objetivo principal es encontrar el lema de cada palabra de la opinión (si es verbo el infinitivo, si es un adjetivo o sustantivo el singular) y los sinónimos con el fin de proveer al polarizador alternativas de búsqueda.

- **Clasificador**: Su objetivo determinar la polaridad de la opinión a partir de los puntajes de cada palabra, de acuerdo al contexto en que se encuentra.
- **Rules**: Es un módulo de soporte para el clasificador que contiene reglas para cada idioma y tienen como objetivo resolver problemáticas generales, como las negaciones, intensificadociones, modalidades y sarcasmo.
- **Repositorio de sinónimos**: Contiene los sinónimos para cada palabra de acuerdo a su categoría léxica. Si se quiere que la implementación sea utilizada para soportar múltiples idiomas debe usarse un repositorio acorde a ello, y lo mejor hoy en día es que brinde los sinónimos en ingles ya que al presente para dicho idioma los lexicones, ontologías, etc están más completos y hay mayor cantidad.
- **Repositorio de puntajes de palabras**: Contiene los puntajes de cada palabra de acuerdo a su categoría léxica.

La principal razón por la que se escogió un método cuantitativo fue el objetivo de buscar un método que permita determinar la polaridad de las opiniones independientemente de la temática de la opinión y el tipo de opinión. Dado que los métodos de aprendizaje automático necesitan ser entrenados por lo general son aplicados a un cierto tipo de texto (o tópico) y el corpus de entrenamiento utilizado es aplicado a ese tipo, por ejemplo si se quiere polarizar criticas cinematográficas, el corpus de entrenamiento va a estar formado por criticas cinematográficas. Ahora bien, si se quiere polarizar cualquier tipo de texto siguiendo este enfoque, se necesita de un corpus de entrenamiento muy extenso y si a eso se le agrega que se busca polarizar textos en diferentes idiomas, el tamaño del conjunto de entrenamiento se volvería aún mucho mayor. En este sentido, la ventaja del método elegido es que se puede aplicar para cualquier tipo de texto y cualquier idioma de los aceptados por el repositorio de sinónimos y el resto de la implementación.

Si bien el modelo propuesto no está basado íntegramente en reglas para determinar la polaridad, una desventaja es (como todo enfoque basado en reglas) el tiempo en el diseño y mantenimiento de las reglas, sin embargo, como se verá en siguientes capítulos es posible encontrar una cantidad razonable de reglas que abarquen diversas situaciones. De hecho, las reglas sólo aumentan la precisión de la implementación del modelo, se pueden no implementar reglas que igualmente es posible determinar la polaridad de la opinión. Por estos motivos se considera que es

una desventaja no tan importante en comparación a un modelo basado íntegramente en reglas.

Un tema a tener en cuenta es que el puntaje de las palabras almacenada en el repositorio de puntajes de palabras no es 100% confiable, sea porque fueron generados de forma automática o si fue realizado de forma manual, el factor subjetividad es muy importante. Aun así no se considera una desventaja del modelo propuesto ya que los conjuntos de entrenamiento también son creados con cierta subjetividad.

Otra ventaja del modelo propuesto es que se puede definir una escala de valores, formada por ejemplo por “Muy negativa”, “Negativa”, “Neutral”, “Positiva”, “Muy positiva” y de acuerdo al puntaje obtenido se determina en que parte de la escala se encuentra la opinión. Lo mencionado es posible hacerlo con métodos de aprendizaje automático, pero el corpus de entrenamiento debe ser mayor, con el método propuesto no se necesita hacer ningún agregado, más que definir los rangos de valores para cada elemento de la escala.

3.3 CONCLUSIÓN

El método propuesto permite una implementación abierta a cualquier idioma, los cuales no son excluyentes, es posible realizar una implementación que soporte diversos idiomas dónde cada uno va a tener reglas asociadas que permitan resolver las negaciones, intensificadores, modalidades y sarcasmo. Como todo método el propuesto tiene sus ventajas y desventajas.

4. DESARROLLO - CAPÍTULO 2

4.1 INTRODUCCIÓN

En el prensente capitulo se verá a nivel global cómo se ha decidido implementar el modelo propuesto para el idioma español.

4.2 DESARROLLO

Con el objetivo de facilitar el entendimiento de la implementación, se presentará el nivel 0, que como se puede observar, cada módulo implementado corresponde con un módulo del modelo, y lo mismo ocurre con los repositorios (los repositorios Babelnet y Wordnet representan la implementación del elemento “Repositorio de sinónimos”).

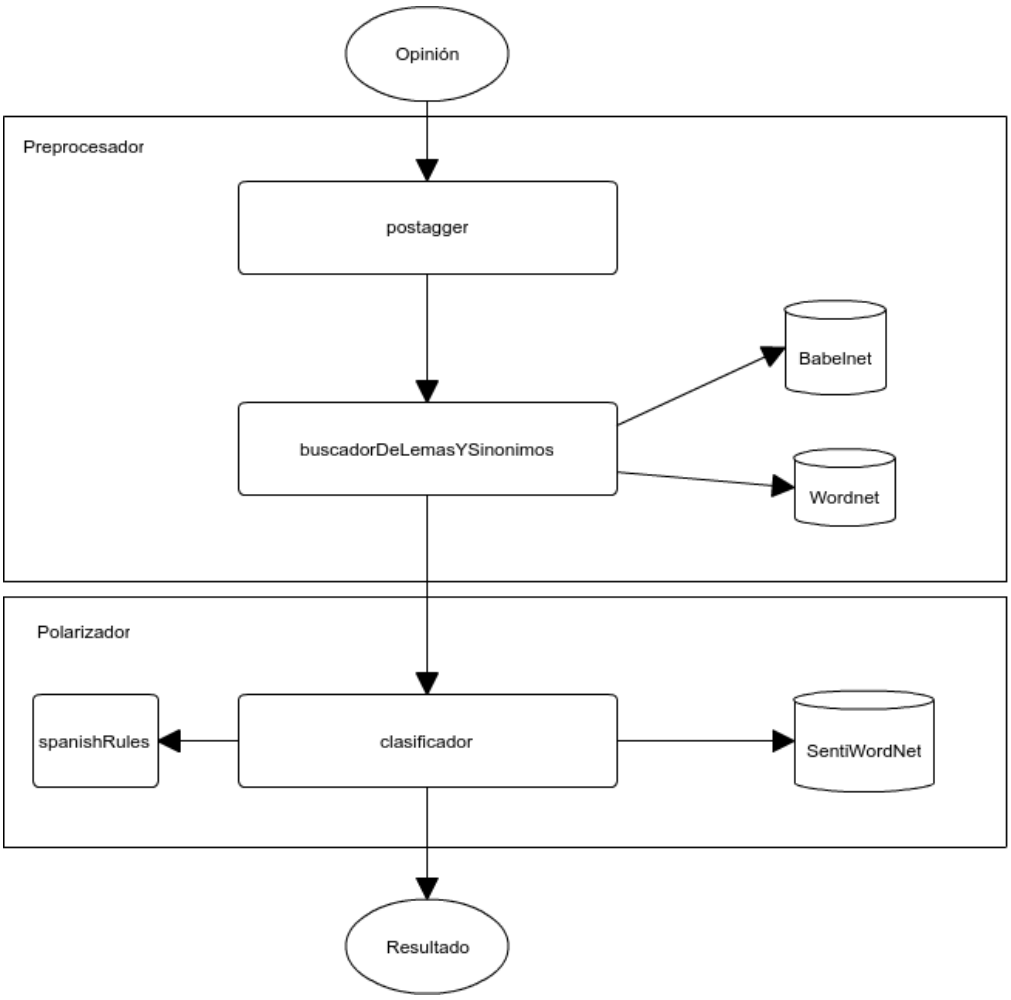

Figura 4: Implementación del modelo

En los próximos capítulos se explicará cada módulo junto con los submódulos que integran cada uno y cada repositorio utilizado en la implementación.

4.3 CONCLUSIÓN

Se ha presentado cómo se implementará el modelo pudiéndose observar a nivel general los modulos implementados y los repositorios que se utilizaron.

5. DESARROLLO - CAPÍTULO 3

5.1 INTRODUCCIÓN

En el prensente capitulo se verá el módulo postagger utilizado en la implementación

5.2 DESARROLLO

Como se ha mencionado anteriormente, este módulo es el encargado de determinar la categoría léxica de cada palabra de acuerdo al contexto en que se encuentra. Antes de describir su implementación, es necesario mencionar que la corrección ortográfica de la opinión no se realiza antes del proceso de postaggeo, sino que se realiza después ya aunque se realice la corrección ortográfica antes del postaggeo nada asegura que sea la palabra correcta, en consecuencia, realizar el proceso de tagging con una palabra mal escrita o con una palabra que no es correcta dentro del contexto de la opinión es lo mismo, ya que en ambos casos la opinión va a estar mal escrita, sea por una razón u otra.

Este módulo está principalmente integrado por una API denominada TreeTagger, desarrollada por Helmut Schmid en el proyecto de cooperación técnica en el Instituto de Lingüística Computacional de la Universidad de Stuttgart. Ha sido aplicado de forma exitosa en diversos idiomas, entre los principales, alemán, inglés, francés, italiano, español, chino y es adaptable a otros idiomas si se encuentran disponibles un léxico y un corpus de entrenamiento etiquetado manualmente.

TreeTagger, como los etiquetadores de n-gramas convencionales (Church 1988 y Kempe, 1993) modela la probabilidad de una secuencia de palabras etiquetando de forma recursiva mediante:

$$p(w_1 w_2 \ldots w_n, t_1 t_2 \ldots t_n) := p(t_n / t_1\, t_2)\, p(w_n / t_n)\, p(w_1 w_2 \ldots w_{n-1}, t_1 t_2 \ldots t_{n-1}) \quad (26)$$

En lo que difieren los TreeTagger con los etiquetadores de n-gramas, es en la forma en que se estima la probabilidad de transición. Los etiquetadores de n-gramas utilizan la siguiente formula basada en la estimación de máxima verosimilitud (MLE):

$$p(t_n / t_{n-2}\, t_{n-1}) = \frac{F(t_{n-2} t_{n-1} t_n)}{F(t_{n-2} t_{n-1})} \quad (27)$$

siendo:

$F(t_{n-2}t_{n-1}t_n)$ el número de ocurrencias del trigrama en el corpus,

$F(t_{n-2}t_{n-1})$ el número de ocurrencias del bigrama.

En cambio TreeTagger estima la probabilidad de transición con un árbol de decisión binaria. La probabilidad de un trigrama dado se determina siguiendo el camino correspondiente a través del árbol hasta que se alcanza una hoja. El árbol de decisión se construye de forma recursiva a partir de un conjunto de entrenamiento de trigramas utilizando una versión modificada del algoritmo ID3 (Quinlan, 1983). Cada paso de la recursión consiste en:

1. Dividir el conjunto de muestras de trigramas en dos subconjuntos con máxima distinción con respecto a la distribución de probabilidad de la tercera etiqueta (prevista)
2. Realizar una prueba que consiste en examinar una de las dos etiquetas anteriores y comprobar si es lo mismo a etiquetar t, la prueba que se realiza es:

   $$\text{tag}_{-i}=t;i \in 1,2;t \in T \quad (28)$$

   siendo T el conjunto de etiquetas posibles
3. Adjuntar al nodo actual la prueba de mayor rendimiento de la información. El criterio utilizado para comparar todas las pruebas posibles “q” consiste en determinar cuál es la prueba que aporta mayor ganancia de información respecto a la tercera etiqueta. Maximizar la ganancia de información es equivalente a minimizar la cantidad media de información “I_q“ que todavía se necesita para identificar la tercera etiqueta. Una vez que el resultado de la prueba de “q” se conoce, la fórmula para obtener Iq es:

   $$I_q = -p(C_+/C)\sum_{t\in T} p\,(t/C_+)\log_2\big(p(t/C_+)\big) - p(C_-/C)\sum_{t\in T} p\,(t/C_-)\log_2\big(p(t/C_-)\big) \quad (29)$$

   siendo:

   C el contexto que corresponde al nodo actual

   C_+es igual a C más la condición de que la prueba “q” tenga éxito

   C_-es igual a C mas la condición de que la prueba “q” falle

   $p(t/C_+)$ la probabilidad de la tercer etiqueta si la prueba tuvo éxito

   $p(t/C_-)$ la probabilidad de la tercer etiqueta si la prueba falló

Las probabilidades mencionadas se estiman a partir de las frecuencias con MLE:

$$p(C_+/C) = \frac{f(C_+)}{f(C)} \quad (30)$$

$$p(C_-/C) = \frac{f(C_-)}{f(C)} \quad (31)$$

$$p(t/C_+) = \frac{f(t,C_+)}{f(C_+)} \quad (32)$$

$$p(t/C_-) = \frac{f(t,C_-)}{f(C_-)} \quad (33)$$

4. Expandir de forma recursiva el nodo con cada uno de los subconjuntos de entrenamiento que conformaban la prueba y adjuntarlos como dos subárboles (recordar que es un árbol binario). Si la próxima prueba generaría al menos un subconjunto de trigramas cuyo tamaño está por debajo de un umbral predefinido (f (C_+) < Umbral o F (C_-) < Umbral) las probabilidades de etiqueta p(t/C) para la tercera etiqueta son estimadas utilizando todos los trigramas que han pasado a la presente recursión y que son almacenados en el nodo actual. La fórmula para calcular la probabilidad es:

$$p(t/C) = \frac{f(t,C)}{f(C)} \quad (34)$$

siendo:

f (C) es el número de trigramas en el conjunto de entrenamiento actual.

f (t, C) es el número de trigramas cuya tercer etiqueta es t.

Una vez terminada la recursión se puede considerar terminado el árbol inicial, es considerado inicial porque luego de ser construido es podado. La poda consiste en ir analizando nodo a nodo e ir eliminando los subnodos que cumplan dos condiciones:

1. Los dos subnodos son hoja
2. La ganancia de información ponderada en el nodo está por debajo de cierto umbral. La ganancia de información ponderada G se define como:

$$G = f(C)\left(I_0 - I_q\right) \text{ con } I_0 = \sum_{t \in T} p\,(t/C) \log_2\left(p(t/C)\right) \quad (35)$$

siendo I_o la cantidad de información que se necesita para eliminar la ambigüedad en el nodo actual; I_q es la cantidad de información que todavía se necesita después de que el resultado de la prueba “q” se conoce.

La razón por la que primero se construye el árbol y luego se poda es que si se aplica el criterio de ganancia de la información durante la construcción puede que haya nodos que no cumplan con el criterio y estos no se construyan, cuando en realidad debería construirse porque todos sus subnodos cumplen el criterio.

Una vez construido y podado el árbol, este está listo para ser utilizado para determinar la mejor secuencia de etiquetas para una determinada secuencia de palabras, para ello TreeTagger utiliza el algoritmo Viterbi (Viterbi 1967).

TreeTagger tiene un léxico que contiene las probabilidades a priori para cada palabra, es similar al léxico que fue utilizado por Cutting (1992) y consta de 3 partes:

1. **Un léxico fullform**, creado a partir de un corpus de entrenamiento etiquetado (cerca de 2 millones de palabras del Penn Treebank Corpus). A dicho corpus se le contó la cantidad de apariciones de cada par palabra/etiqueta y las etiquetas de cada palabra con una frecuencia relativa menor al 1% fueron eliminadas ya que fueron consideradas errores de etiquetado en la recopilación original.
2. **Un léxico sufijo**, está construido de forma automática desde el corpus de entrenamiento. Se organiza como un árbol donde todos los nodos, salvo el raíz, es marcado con un carácter y los vectores de probabilidades se encuentran adjuntados a los nodos hoja. Una vez construido el árbol, el mismo es podado utilizando la medida de información, I(S), de cada nodo. La medida de información se calcula de la siguiente manera:

$$I(S) = -\sum_{\text{pos}} P\,(\text{pos}/S)\log_2\big(P(\text{pos}/S)\big) \quad (36)$$

siendo S el sufijo del nodo actual y P(pos/S) la probabilidad de la etiqueta pos dada una palabra con el sufijo S

Luego para cada nodo hoja se calcula la ganancia de información ponderada G(aS), calculada de la siguiente manera:

$$G(\text{aS}) = \text{F(aS)}\big(I(S) - I(\text{aS})\big) \quad (37)$$

siendo S el sufijo del nodo padre, aS el sufijo del nodo actual y F(aS) es la frecuencia del sufijo aS.

Si la ganancia de información en alguna hoja del árbol de sufijos está por debajo de un umbral dado, se retira la hoja. Las frecuencias de la etiqueta de todos los subnodos borrados de un nodo padre se recogen en el nodo predeterminado del nodo padre. Si el nodo predeterminado es el subnodo único que queda, se elimina también. En este caso, el nodo padre se convierte en una hoja y también se comprueba si hay que borrarlo.

La búsqueda en el árbol de sufijos consiste en una búsqueda a lo largo del camino, donde los nodos están anotados con las letras

del sufijo en orden inverso. La búsqueda en el árbol retorna el vector de probabilidades asociado a la hoja a la que se llegó, se pueden llegar a los nodos hojas en dos situaciones:

1. se encontró el sufijo
2. no se llegó a un nodo hoja porque no existe el sufijo, pero fue posible llegar siguiendo por un nodo predeterminado (no siempre puede estar presente).

En caso de no llegar a un nodo hoja, la búsqueda falla.

3. **Una entrada por defecto**. Una vez podado el árbol de sufijos, se construye la entrada por defecto restado las frecuencias de todas las hojas del árbol resultante y la normalización de las frecuencias restantes.

La secuencia de búsqueda de una palabra en el léxico es la siguiente:

1. Se busca en el léxico fullform
2. Si no se encuentra, todas las mayúsculas se pasan a minúsculas y se vuelva a buscar en el léxico fullform
3. Si no se encuentra, se busca la palabra en el léxico sufijo
4. Si no se encuentra, se devuelve la entrada por defecto

La búsqueda se detiene en cualquier paso en caso de encontrarse la palabra, y se devuelve el vector de probabilidades de la etiqueta correspondiente.

Para finalizar, se mencionaran los test y resultados relevantes obtenidos que justifican la utilización de TreeTagger. La prueba consistió en comparar un etiquetador trigrama (Kempe, 1993) con diferentes versiones de TreeTagger:

1. versión bigrama,
2. versión trigrama reemplazando las frecuencias 0 por 0,1 antes de calcular las probabilidades de las etiquetas en las hojas del árbol de decisión
3. versión trigrama reemplazando las frecuencias 0 por $(10^{-10})^5$ antes de calcular las probabilidades de las etiquetas en las hojas del árbol de decisión. La razón de la presente versión es para ver que tan fuerte es la influencia de la elección de este parámetro en el etiquetado
4. versión quatrogram

Para las pruebas se utilizaron datos del corpus Penn-Treebank del cual 2 millones de palabras se utilizaron para el entrenamiento y 100000 palabras de una

parte diferente del corpus para la prueba. La diferencia entre TreeTagger y el etiquetador trigrama es que el primero tiene, como se explicó anteriormente, un lexicón de sufijos, mientras que el segundo no, pero al léxico fullform del segundo se le agregaron 170000 entradas adicionales que se crearon a partir de una lista de palabras con un analizador morfológico. Los resultados obtenidos fueron:

Tabla V: Resultados de la evaluación de las distintas versiones de TreeTagger

Método	Contexto	Exactitud
Trigram tagger	trigram	96,06%
TreeTagger	bigram	95,78%
TreeTagger (0.1)	trigram	96,34%
TreeTagger	quatrogram	96,36%
TreeTagger	trigram	96,32%

Fuente: Schmid, 1994. Página 16

Como se puede observar, en todas las versiones de TreeTagger, exceptuando la de bigramas, se obtuvo una exactitud mayor que el etiquetador trigrama estándar de Kempe (Kempe, 1993). La siguiente figura grafica la influencia del tamaño del corpus de entrenamiento en la calidad de etiquetado:

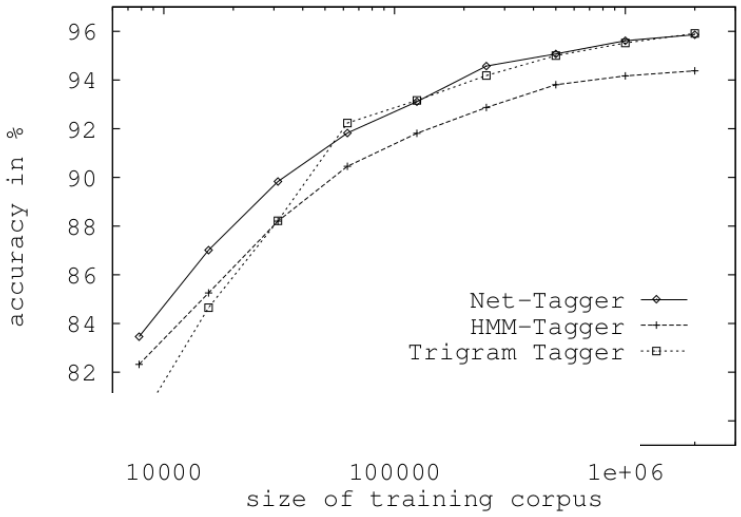

Figura 5:% exactitud por tamaño del corpus de entrenamiento

Se puede observar que en el etiquetador trigrama la curva no es del todo pareja, en consecuencia la exactitud del mismo depende mucho del tamaño del corpus de entrenamiento, en cambio para las versiones de TreeTagger (bigrama y trigrama) la exactitud se deteriora lentamente cuando se achica el corpus de entrenamiento. Además, para el valor mínimo que se muestra en el gráfico, TreeTagger da mejores resultados y si se continúa achicando el corpus, la diferencia con el etiquetador trigrama estándar va a ser mucho mayor. En resumen, TreeTagger es más robusto con respecto al tamaño del corpus de entrenamiento en contraste con el etiquetador trigrama estándar.

Toda la descripción de TreeTagger fue obtenida de Schmid, 1994.

5.3 CONCLUSIÓN

Cómo se vió a lo largo del capitulo, la implementación de TreeTagger ha dado buenos resultados en cuanto a precisión, por lo que se decidió utilizarlo en la presente implementación.

6. DESARROLLO - CAPÍTULO 4

6.1 INTRODUCCIÓN

En el prensente capitulo se verá el módulo buscadorDeLemasYSinonimos junto con todos sus submodulos.

6.2 DESARROLLO

El objetivo del módulo buscadorDeLemasYSinonimos es como su nombre lo indica, obtener los lemas y sinónimos de cada palabra de la opinón. Para poder cumplir con el objetivo de este módulo y facilitar la polarización, hay diversas problemáticas a resolver:

1. Errores ortográficos.
2. Los repositorios contienen los lemas de las palabras, no tienen los verbos conjugados, plurales, etc.
3. Los repositorios que contienen los puntajes de las palabras pueden no contener el lema de la palabra que está en la frase, pero tiene lemas de sus sinónimos.
4. Los principales repositorios están en inglés, mientras que las opiniones tenidas en cuenta para el trabajo presente son en español.

Para resolver dichas problemáticas se han implementado diversos submódulos y como se puede observar en el diagrama, el principal submódulo y orquestador es el denominado buscadorDeWords.

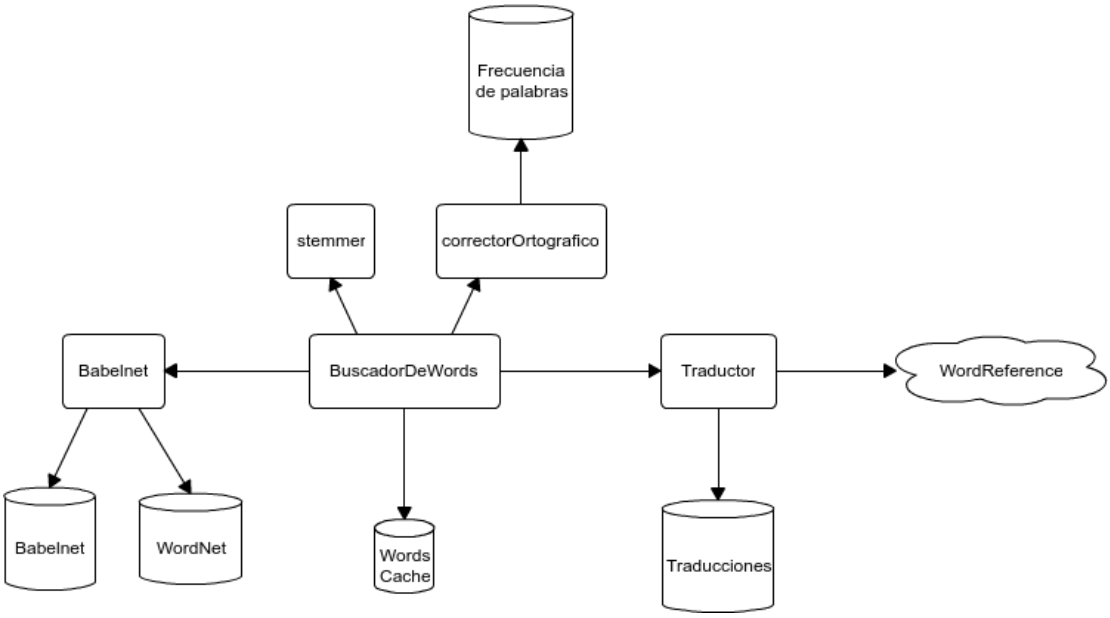

Figura 6: Submodulos y repositorios del modulo buscadorDeLemasYSinonimos

A continuación, en las siguientes secciones se describirá la implementación de cada submódulo junto con los repositorios que utiliza.

6.2.1 Submódulo correctorortografico

Como su nombre lo indica, su principal objetivo es corregir las palabras con errores ortográficos. Está compuesto principalmente por una API denominada LanguageTool y por un repositorio descargado del corpus de la Real Academia Española (RAE), al cual se lo denominó “Frecuencias de palabras”.

- **Frecuencias de palabras**: Este repositorio contiene las palabras (la RAE las denomina formas ortográficas) registradas en el Corpus de Referencia del Español Actual (CREA), con sus frecuencias absolutas y normalizadas, sin tener en cuenta las mayúsculas o minúsculas. A modo informativo, el CREA tiene como objetivo brindar información relativa a la lengua durante diferentes periodos de la historia almacenando las variedades relevantes de dicha lengua y está formado por diversos textos, por ejemplo la versión escrita contiene textos de otros corpus, prensa, libros, entre otros.
- **LanguageTool**: Es la API encargada de proporcionar posibles palabras correctas para las palabras mal escritas. Soporta diversos idiomas, entre los que se encuentra el español, inglés, alemán, polaco, japonés, francés, griego, italiano.

El funcionamiento del submódulo, es el siguiente:

1. Se obtienen las posibles correcciones provistas utilizando LanguageTool
2. De las correcciones provistas se toman las 3 primeras con mayor frecuencia de acuerdo al repositorio “Frecuencia de palabras”

6.2.2 Submódulo traductor

Dado que uno de los objetivos de la implementación es que soporte diversos idiomas y que los principales repositorios están en inglés, es necesario un submódulo que se encargue de realizar las traducciones necesarias y que sea independiente del idioma de las opiniones, para ello se ha desarrollado un submódulo que:

1. Mantiene el repositorio “Traducciones”, dicho repositorio es generado y actualizado por la implementación y almacena el par palabra/traducción de cada palabra que se traduce, con el objetivo de mejorar la performance de la implementación y no realizar traducciones que ya se realizaron.
2. Se comunica con WordRefence para realizar las traducciones. Actualmente, no hay servicios de traducción gratuitos para integrar a las aplicaciones, por lo que se decidió que la aplicación realice la traducción a través de consultas a http://www.wordreference.com/. Teniendo en cuenta esto, el repositorio “Traducciones” toma mucha más importancia, ya que si se está consultando constantemente a dicha página, se ha comprobado, después de una cierta cantidad de consultas seguidas en un periodo muy corto de tiempo deja de responder por un tiempo y esto va a afectar directamente a los resultados de la polaridad de las opiniones, ya que pueden no encontrarse sinónimos y no va a ser posible asignarle un puntaje a la palabra.

Para finalizar se explicará el funcionamiento del submódulo:

1. Ante una nueva palabra, se busca la misma en el repositorio “Traducciones”, de existir, el resultado de la traducción va a ser lo almacenado en el repositorio mencionado.
2. De no existir la traducción en el repositorio se consulta a WordReference por la traducción. Cabe mencionar que la consulta consiste en hacer un GET a la página mencionada, por lo tanto la respuesta que se obtiene es

un texto HTML y para extraer las traducciones se utilizan expresiones regulares.

3. De encontrarse resultados en la consulta a WordReference, se almacena el mismo en el repositorio y se retornan las posibles traducciones.

6.2.3 Submódulo babelnet

El objetivo de este submódulo es obtener los lemas y sinónimos en inglés de cada palabra de la opinión. Se puede afirmar que este submódulo permite que la implementación esté abierta a diversos idiomas, la principal razón de esta afirmación es que, como ya se ha mencionado anteriormente la mayoría de los tesauros están en inglés y si se los quiere utilizar, es necesario que las opiniones estén en inglés o que haya un paso intermedio para pasar del idioma de cada palabra de la opinión al inglés, este paso intermedio es la una de las funciones que cumple el presente submódulo, ya que permite obtener los synsets de una determinada palabra al idioma que se desee, en este caso, en inglés.

La implementación se encuentra íntegramente formada por la API de Babelnet, con minimas modificaciones para ajustarla al contexto en el que va a ser utilizada. Como se ha mencionado anteriormente Babelnet ofrece dos formas para integrarla con las aplicaciones, la primera ofreciendo acceso Resource Description Framework (RDF) a la Linguistic Linked Open Data cloud; la otra es descargar la enciclopedia completa y realizar consultas de forma local. Por cuestiones de performance y de no hacer cientos o miles de consultas al cloud que pueden repercutir en la prestación del servicio por parte del cloud, se decidió la segunda forma.

6.2.4 Submódulo stemmer

El objetivo de este submódulo es obtener los posibles lemas para cada palabra de la opinión. Para evitar todo tipo de confusión, cabe aclarar que el presente submódulo se denomina así porque el principal tarea para obtener los posibles lemas de las palabras es el stemming.

Pare realizar el stemming se utiliza la API denominada Snowball, desarrollada para diversos idiomas, entre los que se encuentran inglés, francés, español, portugués, italiano, rumano, alemán, danés, sueco, noruego y finlandés. Dicha API implementa el algoritmo creado por Porter (Porter, 1980), es decir que hace un stemming de la categoría “eliminación de afijos”, que como se mencionó en

el estado de arte, elimina sólo sufijos. La implementación para el español está basada en las reglas del idioma y antes de describir el algoritmo, es necesario mencionar algunas consideraciones y definiciones:

- Letras consideradas acentuadas: ***á***, ***é***, ***í***, ***ó***, ***ú***, ***ü***, ***ñ***
- Letras consideradas vocales: ***a***, ***e***, ***i***, ***o***, ***u***, ***á***, ***é***, ***í***, ***ó***, ***ú***, ***ü***
- R1 y R2:
  - R1 es la región después de la primera consonante después de una vocal, si no hay vocal, es considerada una región nula.
  - R2 es la región después de la primer consonante despues de una vocal dentro de la región R1, si no hay vocal, es considerada una región nula.
  - Ejemplo: para la palabra parlante la región R1 sería “lante” y al región R2 “te”.
- RV: Si la segunda letra es consonante la región RV es después de la siguiente vocal, si las dos primeras letras son vocales la región RV es después de la siguiente consonante y si empieza con consonante-vocal la región RV es después de la tercer letra. Si no se encuentran éstas posiciones, la región RV es el final de la palabra.
- Ejemplos: Para macho la región RV es "ho", para oliva es "va", para trabajo es "bajo" y para áureo es "eo"

Mencionadas las consideraciones y definiciones, se explicara el algoritmo:

1. Adjuntar pronombres

   1.1. Buscar el sufijo más largo entre ***me***, ***se***, ***sela***, ***selo***, ***selas***, ***selos***, ***la***, ***le***, ***lo***, ***las***, ***les***, ***los***, ***nos***

   1.2. Eliminar el sufijo si viene después de

   a) ***iéndo***, ***ándo***, ***ár***, ***ér***, ***ír*** en RV***.*** Además de la dicha eliminación se elimina el acento agudo, por ejemplo la palabra “haciéndola” quedaría como “Haciendo”)

   b) ando, iendo, ar, er, ir en RV.

   c) ***yendo*** precedido de una ***u.***En este caso, yendo debe estar en RV, pero la u anterior puede no estarlo.

2. Buscar el sufijo más largo y realizar la acción asociada

2.1. Sufijos: anza, anzas, ico, ica, icos, icas, ismo, ismos, able, ables, ible, ibles, ista, istas, oso, osa, osos, osas, amiento, amientos, imiento, imientos

Acción: Eliminar si está en R2

2.2. Sufijos: adora, ador, ación, adoras, adores, aciones, ante, antes, ancia, ancias

Acción: Eliminar si está en R2 y si está precedido por ***ic*** eliminarlo también si está en R2

2.3. Sufijos: logía, logías

Acción: Reemplazar por ***log*** si está en R2

2.4. Sufijos: ución, uciones

Acción: Reemplazar por ***u*** si está en R2

2.5. Sufijos: encia, encías

Acción: Reemplazar por ***ente*** si está en R2

2.6. Sufijos: ***amente***

Acción: Eliminar si está en R1, si es precedido por ***iv*** eliminar si está en R2 (y si es precedida ademas por ***a*** eliminar si está en R2), de lo contrario, si está precedido por ***os***, ***ic*** o ***ad*** eliminar si está en R2.

2.7. Sufijos: ***mente***

2.8. Acción: Eliminar si está en R2. Si está precedido por ***ante***, ***able*** o ***ible***, eliminar si está en R2.

2.9. Sufijos: ***idad***, ***idades***

Acción: Eliminar si está en R2. Si está precedido por ***abl***, ***ic*** o ***iv***, eliminar si está en R2.

2.10. Sufijos: ***iva***, ***ivo***, ***ivas***, ***ivos***

Acción: Eliminar si está en R2. Si está precedido por ***at*** eliminar si está en R2.

3. El presente paso es para los sufijos de los verbos, aun así, el paso 2.1 se realizara si en el paso 1 no se eliminó ejecutó ninguna acción de las mencionadas y el paso 2.2 se realizara solo si se ejecutó el paso 2.1 y no se han eliminado sufijos.

3.1. Para los verbos con sufijos que empiezan con ***y*** (***ya***, ***ye***, ***yan***, ***yen***, ***yeron***, ***yendo***, ***yo***, ***yó***, ***yas***, ***yes***, ***yais***, ***yamos***), buscar el más largo en RV y eliminarlo si se encuentra precedido por ***u*** (no es necesario que esté en RV).

3.2. Para otros sufijos verbales buscar el sufijo más largo en RV y realizar la acción asociada

a) Sufijos: ***en***, ***es***, ***éis***, ***emos***

Acción: Eliminar y si está precedido por ***gu*** eliminar la ***u*** (***gu*** no necesita estar en RV)

b) Sufijos: arían, arías, arán, arás, aríais, aría, aréis, aríamos, aremos, ará, aré, erían, erías, erán, erás, eríais, ería, eréis, eríamos, eremos, erá, eré, irían, irías, irán, irás, iríais, iría, iréis, iríamos, iremos, irá, iré, aba, ada, ida, ía, ara, iera, ad, ed, id, ase, iese, aste, iste, an, aban, ían, aran, ieran, asen, iesen, aron, ieron, ado, ido, ando, iendo, ió, ar, er, ir, as, abas, adas, idas, ías, aras, ieras, ases, ieses, ís, áis, abais, íais, arais, ierais, aseis, ieseis, asteis, isteis, ados, idos, amos, ábamos, íamos, imos, áramos, iéramos, iésemos, ásemos

Acción: Eliminar

4. Este paso está relacionado a los sufijos residuales. Buscar el sufijo más largo en RV y realizar la acción asociada

4.1. Sufijos: ***os***, ***a***, ***o***, ***á***, ***í***, ***ó***

Acción: Eliminar

4.2. Sufijos: ***e***, ***é***

Acción: Eliminar y si está precedido por ***gu*** eliminar la ***u*** si la ***u*** está en RV.

5. Se remueven los acentos agudos.

La descripción del algoritmo de stemming de Snowball fue obtenida de la página oficial (Spanish stemming algorithm, 2016).

Como se mencionó en la descripción del submódulo, el objetivo es obtener los posibles lemas de una palabra determinada, entonces, el funcionamiento del mismo para cumplir su objetivo es:

1. Se obtiene la raíz de la palabra utilizando Snowball
2. Se le agregan los posibles sufijos para completar la raíz:

2.1. Si la categoría léxica de la palabra en el opinión es un verbo se generan 3 posibles lemas (o mejor dicho, los posibles infinitivos) formados por la raíz + “***ar***”, raíz + “***er***” y raíz + “***ir***”.

2.2. Si la categoría léxica de la palabra en el opinión no es un verbo se generan dos posibles lemas formados por raíz + “***a***” y raíz + “***o***”. Se concatena “***a***” u “***o***” por el femenino o masculino respectivamente.

El conjunto resultado va a estar formando posibles lemas formados de acuerdo al algoritmo descripto y si la categoría léxica de la palabra no es verbal, se le agrega además la raíz de la misma.

Ejemplos del resultado del submodulo serían: para la palabra “comiendo” los posibles lemas serían “comar”, “comer” y “comir”; para la palabra “reloj” los posibles lemas serian “reloja”, “relojo” y “reloj”; para la palabra “rápidamente” los posibles lemas serian “rápid”, “rápido” y “rápida”.

6.2.5 Submódulo buscadorDeWords

Una vez explicado todos los submódulos de apoyo es necesario explicar cómo es que son orquestados para finalmente en su conjunto cumplir el objetivo del módulo buscadorDeLemasYSinonimos, en otras palabras, el presente submódulo es el responsable de obtener los lemas y sinónimos de las palabras de la opinión para ser utilizados en el cálculo de la polaridad de la opinión utilizando el resto de los submódulos. Antes de explicar el algoritmo, es necesario mencionar que el repositorio WordsCache tiene todos los pares palabra-categoría léxica que han sido procesados tal cual han aparecido en los textos analizados junto con los synsets principales y secundarios obtenidos de BabelNet; este repositorio fue creado para reducir los tiempos de procesamiento ya que se detectó que la consulta a BabelNet producía demoras. El algoritmo implementado por el presente submódulo es el siguiente:

Ante una opinión, sólo se procesan los sustantivos, adjetivos, verbos y adverbios ya que son los únicos que aportan al cálculo de la polaridad. Para cada palabra:

1 Se buscan los synsets principales y secundarios en el repositorio WordsCache utilizando el par palabra- categoría léxica, si se encuentran, se sigue con la siguiente palabra.

2 Si no se ha encontrado el par palabra-categoría léxica se realiza una búsqueda sin traducciones ni correcciones ortográficas:

2.1 Si es un verbo, se consulta al submódulo stemmer los posibles verbos infinitivos del mismo, se consulta uno a uno al submódulo babelnet los

posibles infinitivos, de encontrar synsets, no se sigue consultando. Por ejemplo para la palabra “comiendo”, el stemmer va a retornar “comar”, “comer” y “comir”, se va a consultar “comar”, como no se van a obtener resultados, se consulta por “comer” y al obtener resultados, no se sigue consultando.

2.2 Si no es un verbo, se hace una consulta al submódulo babelnet con la palabra tal cual se encuentra en la opinión.

2.2.1 Si no se obtienen synsets, se consulta al submódulo stemmer las los posibles lemas y se consulta uno a uno al submódulo babelnet los posibles lemas, de encontrar resultados, no se sigue consultando, siguiendo el mismo procesamiento que los posibles infinitivos explicado anteriormente

3 Si no se obtuvieron synsets, se traduce al inglés la palabra. Dado que pueden haber varias traducciones, se sigue la misma lógica mencionada anteriormente, se consulta al submódulo babelnet cada traducción hasta encontrar synsets, una vez encontrado, se deja de consultar.

4 Si no se obtuvieron synsets, se intenta obtenerlos utilizando el lema tentativo propuesto por el módulo postagger, siguiéndose los pasos 1 y 2.

5 Si no se obtuvieron synsets, se utiliza el submódulo correctorOrtografico para obtener las posibles correcciones, y se realiza el paso 1 y 2 para cada posible corrección hasta que se encuentren synsets, si se encuentran, no se procesan las restantes correcciones, es decir, se sigue la misma lógica de corte que en pasos anteriores.

6 Si se encontraron synsets:

6.1 se le asigna a la palabra los synsets encontrados como synsets principales

6.2 se le asigna a la palabra los synsets asociados a los BabelSense del synset como synsets secundarios

6.3 se agrega el par palabra-categoría léxica junto con los synsets principales y secundarios al repositorio WordsCache.

En caso de finalizar el procesamiento de la palabra y no encontrar synsets, la palabra no aportara puntaje en el cálculo de la polaridad. Es de importancia explicar las razones por las que el módulo postagger es fundamental y justifican su existencia:

- Las consultas a Babelnet se realizan teniendo en cuenta la categoría léxica de la palabra obtenida mediante el módulo postagger, si no realizaría dicha tarea se asignarían a las palabras sinónimos que en realidad no lo son, ya que hay palabras que pueden tener distintas categorías léxicas según el contexto en que se encuentran, por ejemplo la palabra “copa” puede ser un sustantivo o un verbo, dependiendo del contexto en que se encuentre.
- Las posibles palabras propuestas por el submódulo stemmer dependen de la categoría léxica, los posibles resultados de acuerdo a la categoría léxica fueron descritos al explicarlo dicho submódulo.

Para cerrar es pertinente mencionar que cada submódulo utilizado por el presente no brinda una posible palabra sino que al contrario, brinda varias ya que no es posible determinar cuál es la correcta con un 100% de precisión, por lo tanto se va a tomar como correcta la primera que es encontrada en Babelnet.

6.2.6 Comentarios finales

Al comenzar la descripción del módulo se enumeraron ciertas problemáticas a resolver, por lo tanto, a modo resumen y para cerrar el módulo se mencionará como se resolvió cada problemática.

1. **Errores ortográficos**. Se resolvió mediante el submódulo correctorOrtografico.
2. Los repositorios contienen los lemas de las palabras, no tienen los verbos conjugados, plurales, etc. Se resolvió mediante el submódulo stemmer.
3. **Los repositorios que contienen los puntajes de las palabras pueden no contienen el lema de la palabra que está en la frase, pero tiene lemas de sus sinónimos**: Para resolverla se creó el submódulo babelnet.
4. **Los principales repositorios están en inglés, mientras que las opiniones tenidas en cuenta para el trabajo presente son en español**: Es la principal problemática que llevó a utilizar Babelnet ya que como se mencionó, es un diccionario enciclopédico multilingüe que permite obtener de forma sencilla lemas y sinónimos en inglés (o cualquier otro idioma).

6.3 CONCLUSIÓN

A lo largo del capitulo se vieron las principales problemáticas y cómo fueron resueltas junto con el detalle del algoritmo implementado en cada submodulo y la implementación permite extenderlo a diversos idiomas, gracias Babelnet y las APIs utilizadas.

7. DESARROLLO - CAPÍTULO 5

7.1 INTRODUCCIÓN

En el prensente capitulo se verá el módulo clasificador, cuyo objetivo es, una vez hecho el postaggeo y búsqueda de lemas y sinónimos, determinar la polaridad de la opinión.

7.2 DESARROLLO

Como se mencionó anteriormente, se definió un método cuantitativo para determinar la polaridad de la opinión, por lo tanto, es necesario que el presente módulo utilice un repositorio para obtener los puntajes de las palabras; además para tratar las diferencias entre diferentes idiomas y principalmente las negaciones, intensificaciones, modalidad y sarcasmo se apoya en módulo que implementa las reglas, en el presente trabajo se implementaron para el idioma español en el módulo “spanishRules” y como se mencionó al inicio del trabajo, no se ha tratado ni la modalidad ni el sarcasmo.

Dado que el módulo buscadorDeLemasYSinonimos brinda lemas y sinónimos en inglés independientemente del idioma de la opinión y la gran cobertura de palabras en inglés se ha elegido como repositorio de puntajes de las palabras a “SentiWordNet”. El puntaje de la opinión está dado por:

$$p_t = \sum p_i \times m_i \quad (38)$$

siendo: P_t el puntaje total de la opinión

p_i el puntaje de la palabra

m_i el multiplicador de la palabra

Dado que SentiWordNet tiene varios synsets (o sentidos) para una misma palabra y categoría léxica, cada uno con puntajes diferentes, había dos estrategias posibles:

1. Desambiguar el significado de la palabra para determinar el puntaje del synset para el sentido de la palabra utilizada.
2. Utilizar un puntaje representativo de todos los synsets para el par palabra/categoría léxica, por ejemplo un promedio.

Ante estas dos estrategias se optó por la segunda, dónde el puntaje para cada par palabra/categoría léxica se calcula como el promedio de los puntajes de todos sus sentidos. Las razones de la elección fueron:

- En las redes sociales hay opiniones muy pequeñas que no aportan información que pueda ser tenida en cuenta para determinar el sentido de una determinada palabra.
- SentiWordNet fue construido de forma semiautomática, por lo tanto puede haber errores en los puntajes, que se pueden compensar, en menor medida, obteniendo un puntaje representativo para ese par palabra/categoría léxica. De todas formas, aunque hubiera sido construido de forma manual los puntajes van a depender de la subjetividad de las personas que asignen los valores y del idioma con el que fue construido, por lo tanto se considera que también debería obtenerse un puntaje representativo.
- Si se persigue el objetivo de una implementación que soporte varios idiomas, SentiWordNet puede no contener todos los significados para un par palabra/categoría léxica para un idioma diferente al inglés, además que SentiWordNet no abarca todo el idioma inglés dejando afuera sentidos de palabras.

El multiplicador de la palabra y a modo introductorio para el siguiente capítulo, es determinado por las reglas de cada idioma, el módulo responsable de ello es spanishStemmer. Por defecto el multiplicador es 1.

7.3 CONCLUSIÓN

En el presente capitulo detalló cómo se calcula la polaridad de las opiniones, indicando de dónde se obtienen los puntajes de las cada palabra y las decisiones de implementación tomadas.

8. DESARROLLO - CAPÍTULO 6

8.1 INTRODUCCIÓN

En el prensente capitulo se verán las reglas implementadas, así cómo la lógica utilizada para definir los multiplicadores de cada regla

8.2 DESARROLLO

El objetivo del módulo es definir los multiplicadores de cada palabra de la opinión. La razón por la que cada idioma debe tener sus reglas es que cada idioma tiene estructuras sintácticas diferentes, en consecuencia el análisis de las negaciones, intensificadores y demás es diferente.

La mayoría de las reglas implementadas surgieron del análisis de las negaciones y los intensificadores, explicados en anteriormente en el estado de arte. Para definir los términos de inicio del ámbito de acción de cada regla se tuvo en cuenta el concepto de n-gramas, ya que hay palabras que para que tengan un determinado sentido tienen que aparecer juntas, por ejemplo “sin embargo”, “así que”. Cada regla tiene asociado un multiplicador, y si una palabra queda afectada por dos o más reglas reglas lo que se hace es multiplicar el multiplicador que tenía asociado la palabra hasta el momento con el multiplicador de la regla que se está ejecutando. Para facilitar la interpretación, primero se explicara cada regla (agrupadas por situación que intenta resolver) y se mencionara el multiplicador asociado y luego se explicará la lógica utilizada para definir los multiplicadores de cada regla.

8.2.1 Negación

Para esta situación existe una única regla, cuyo único objetivo es invertir la polaridad todas las palabras que forman parte de su ámbito de acción, por lo tanto y ya que en SentiWordNet los sentidos positivos tienen puntaje positivo y los sentidos negativos tienen puntaje negativo, para invertir la polaridad el multiplicador de la regla es -1. Para identificar la presencia se la negación en una opinión se utiliza la lista de palabras consideradas indicadoras de negación en la lengua española, la

lista de palabras está integrada por “Jamás”, “Nada”, “Nadie”, “Negativamente”, “Ni”, “Ningún”, “No”, “Nunca”, “Rehúso”, “Tampoco”, “Sin”, “Ni siquiera”. Para determinar el ámbito de acción de la negación entre las dos estrategias mencionadas anteriormente, se optó por la estrategia de Pang, Lee, y Vaithyanathan (Pang, Lee, y Vaithyanathan, 2002) que consistía en tomar todas las palabras entre la señal de negación y el primer signo de puntuación pero adicionalmente se agregaron otros términos, dando como resultado una lista de caracteres y términos: ‘!’, ‘? ’, ‘. ’, ‘, ’, ‘; ’, ‘ (’, ‘) ’, “porque”. La razón por la que tienen en cuenta “porque” es porque su presencia indica que se va a dar las razones de lo mencionado previamente, por ejemplo la opinión “No es mala persona porque ayuda a los que más lo necesitan” es claramente positiva y si no se considera la palabra “porque” como fin de la negación posiblemente sea considerada una opinión negativa, va a depender mucho de los puntajes de las palabras utilizadas. Cabe mencionar que si no se encuentran los caracteres y términos de corte de la negación se va a aplicar la negación a toda la opinión.

8.2.2 Intensificadores

Para definir las reglas para implementar los intensificadores se utilizó como base la clasificación realizada en Quirk et.al. (1985), que como se mencionó, se clasificó los intensificadores en amplificadores y decrementadores. En la nueva clasificación propuesta, los amplificadores y decrementadores se dividen en lo incrementadores/decrementadores que afectan un número indeterminado de palabras y los incrementadores/decrementadores de la próxima palabra que afectan a la palabra posterior al intensificador.

8.2.3 Amplificadores

Entre las reglas que afectan a un número indeterminado de palabras se pueden identificar 2 reglas diferentes:

1. **Exclamaciones**: Es la forma por excelencia para poner énfasis en toda la opinión o parte de la misma. Para identificarla se busca la presencia del signo de exclamación (“!”) y van a verse afectadas las palabras que se encuentren entre los caracteres '?', '.', ',', ';', '¡' (si no se encuentran alguno de los caracteres de inicio, se toma desde el inicio) y ‘!’. A todas las palabras que estén entre el carácter de inicio y de fin se le asignara un multiplicador de 3.

2. **Otros incrementadores**: Se considera que hay situaciones como razonamientos o contrastes que deben tener una mayor influencia al momento de determinar la polaridad de la opinión ya que en general indican cual es la idea final de la opinión o que directamente indican que lo posterior debe tener más importancia que el resto de la opinión. Los términos que indican el inicio del ámbito de acción son:
    - “Pero” y “Sin embargo”: Son tomados como términos de contraste, cuando uno dice “Es bello pero no sirve para nada”, claramente por más que el objeto al que se hace referencia tiene algo positivo, la orientación de la opinión es negativa ya que el “pero” indica que, más allá de lo positivo, pesa más lo negativo.
    - “entonces” y “así que”: Son tomados como términos de razonamiento, indican el inicio de una conclusión, le procede lo que encierra la opinión o parte de la opinión, por lo tanto se considera que se le debe dar más importancia. Se tiene en cuenta que en general lo que precede a estos términos tienen la misma orientación que la conclusión, pero de no haber una clara polaridad en los términos precedentes por los términos utilizados en la opinión, los términos procedentes deberían contribuir a mejorar la exactitud, es por ello que se decidió aumentarles la importancia, o en otras palabras, incrementarla.

    A las palabras que estén en el ámbito de acción de la regla se le asigna un multilicador de 3.

Los dos grupos tienen como caracteres que indican el fin del ámbito de acción de los incrementadores los caracteres ‘!’, ‘?’, ‘.’, ‘,’, ‘;’, ‘ (‘, ‘)’.

Por otro lado, hay solo una regla del grupo de las reglas que afectan a la próxima palabra luego de los términos que indican amplificación. Las palabras que amplifican la siguiente palabra son “tan”, “muy”, “mucho”, “mucha”, “muchos”, “muchas” y el multiplicador asignado es 3. Dado que no es lo mismo decir “El producto es bueno” que “El producto es muy bueno” ya que más allá que las dos sean positivas, la segunda es mucho más positiva que la primera. Además si la opinión sería mucho más larga, ese “muy bueno” debería pesar más ya que dado que se ha elegido un método cuantitativo para determinar la polaridad es importante tener en cuenta estas cuestiones.

8.2.4 Decrementadores

Análogamente a los amplificadores, encontramos reglas que afectan a un número indeterminado de palabras, en este caso solo hay una regla para decrementar un conjunto indeterminado de palabras. Los términos que indican el inicio del ámbito de acción son “Mas allá” y “a pesar”, los caracteres que indican el fin del ámbito de acción son ‘!’, ‘?’, ‘.’, ‘,’, ‘;’, ‘ (‘, ‘)’ y el multiplicador que se le asigna a las palabras que forman parte del ámbito de acción un 0.25. Los términos mencionados como inicio del ámbito reducen la importancia en la polaridad de las palabras que afectan porque lo que se quiere decir utilizándolas es que hay cosas más positivas o negativas que lo que se menciona dentro del ámbito de acción, por ejemplo “Más allá de haber jugado un mal partido, es un muy buen equipo que está lleno de personas muy profesionales”, esta opinión es claramente positiva y queda en evidencia que lo que está luego del “más allá” es menos importante que lo que está después de la “,”.

Por otro lado, al igual que los amplificadores hay una regla que afecta a la próxima palabra luego de los términos que indican el decremento. Las palabras que decrementan la siguiente palabra son “poco, “poca”, “pocos”, “pocas” y el multiplicador asignado es 0.5. Como se explicó con el amplificador de la próxima palabra, no es lo mismo decir “El producto es un poco malo” que decir “El producto es malo”, ambas opiniones son negativas, pero la primera no es tan negativa como la segunda y si la opinión sería mucho más larga, ese “poco malo” debería ser considerado menos negativo.

8.2.5 Otras reglas

Se han creado dos reglas que no están relacionadas con ninguno de las situaciones mencionadas en capítulos anteriores. Las dos reglas diseñadas tienen como objetivo anular la polaridad de una parte de la opinión y en ambas el multiplicador asignado es 0. Una regla tiene como objetivo anular los nombres o títulos encerrados entre comillas (‘’) ya que lo que se encuentra dentro no debe ser tenido en cuenta para determinar la polaridad de la opinión, por ejemplo “'El mejor jugador del mundo' es una película muy mala.”, en este caso, el titulo sería “El mejor jugador del mundo”, y no debe ser considerada porque es el nombre de una película. Por otro lado, hay palabras que para SentiWordNet tienen un puntaje y se considera que no deben tenerse en cuenta para el español; las palabras que son anuladas son:

- **“Nunca”**: Es considerado un negador y en SentiWordNet tiene un puntaje negativo alto y se considera que debe anulase porque si la opinión es “Nunca brindaron dieron un mal servicio” claramente es una opinión positiva pero si el "Nunca" tiene mayor puntaje que la suma de las otras palabras se va tomar como una opinión negativa, cuando realmente no lo es.
- **“Gracias”**: en SentiWordNet tiene un puntaje positivo alto, pero se considera que debe anulase porque si la opinión es positiva el resto de la opinión va a tener palabras con puntaje positivo, es decir que no agrega nada al cálculo de la polaridad, en cambio si la opinión es “Gracias por brindar un mal servicio” claramente es una opinión negativa pero si es Gracias tiene mayor puntaje que la suma de las otras palabras se va tomar como una opinión positiva, cuando realmente no lo es.
- **“muy”**: se considera un intensificador que no debe pesar en el cálculo de la polaridad

8.2.6 Los multiplicadores

A modo resumen, las reglas implementadas son las siguientes:

Tabla VI. Reglas implementadas

Regla	Inicio	Corte	Palabras	Multiplicador
Negación	Jamás Nada Nadie Negativamente Ni Ningún No Nunca Rehúso Tampoco Sin Ni siquiera	! ? . , ; ( ) porque		-1
Anulación títulos	'	'		0
Anulación	No tiene	No tiene	Nunca Gracias Muy	0

Regla	Inicio	Corte	Palabras	Multiplicador
Exclamación	? . , ; i (	!		3
Otros incrementadores	pero entonces sin embargo asi que	! ? . , ; ( )		5
Decrementador	Mas allá, a pesar	! ? . , ; ( )		0,25
Incrementador proxima palabra	tan, muy, mucho, mucha, muchos, muchas	Siguiente palabra		3
Decrementador próxima palabra	poco, poca, pocos, pocas	Siguiente palabra		0,5

El multiplicador por defecto es 1 para que si la palabra no es afectada por ninguna de las reglas mencionadas el puntaje que aporta al momento de calcular la polaridad de la opinión es el que se encuentra en SentiWordNet; partiendo de esto, la negación tiene valor -1 porque como se mencionó el puntaje de la palabra es el mismo pero invertido. Por otro lado, para anular los puntajes de las palabras, debe utilizarse como multiplicador 0 para que no aporte puntaje en el cálculo final.

Entre los incrementadores, se consideró que los que los denominados “Otros incrementadores” deben tener un multiplicador mayor ya que, como se mencionó, indican la idea final de la opinión u oración, mientras que los otros, solo le dan más

importancia a una palabra o grupo de palabras. Para reflejar lo mencionado, se decidió que el multiplicador de los “Otros incrementadores” sea 5, y el resto de los incrementadores (exclamaciones e incrementadores de la próxima palabra) estén en un punto medio entre 1 y 5, es decir 3; si se seguía ésta lógica con valores de 3 y 2 respectivamente se considera que no refleja la intención de la regla, llevando además a posibles errores en la polarización, por ejemplo la opinión “Ese dispositivo es incómodo y feo, sin embargo es útil” es positiva, más allá de que el “incómodo y feo” reflejan algo negativo, la intensión de la opinión es que más allá de esas cosas negativas o “malas”, se tiene una percepción positiva del dispositivo por ser “útil”, ahora bien: independientemente del puntaje que tenga SentiWordNet para esos 3 adjetivos, pueden darse los siguientes casos:

- Si no existiese la regla, la correcta polaridad va a depender de los puntajes del lexicón emocional: si | incomodo + feo | < | útil | el resultado va a ser correcto, caso contrario, incorrecto.
- De existir la regla, asumiendo que “m” es el multiplicador asociado a la regla “Otros incrementadores”, los dos casos más relevantes a mencionar son:
  - Si “m” es pequeño, | incomodo + feo | > | m * útil |
  - Si “m” es grande, | incomodo + feo | < | m * útil |

Notar que en todas las situaciones, se depende del puntaje que tengan las palabras, pero más allá de eso, lo que se busca es reducir la probabilidad polarizar mal y entre los dos casos mencionados para cuando existe la regla, la probabilidad de error para un “m” pequeño es mayor que para un “m” grande. En el ejemplo, supongamos los puntajes:

incomodo: -0,3
feo: -0,4
útil: 0,2

Para m = 3, la sumatoria que da la polaridad sería -0,3 -0,4 + 3 * 0,2 = -0,1;
Para m = 5, la sumatoria que da la polaridad sería -0,3 -0,4 + 5 * 0,2 = 0,3;

Como se puede observar, cuanto mayor sea el multiplicador, menor va a ser la probabilidad de error. La razón por la que se eligió 5 y no un número mayor es que la frase u oración que tiene este tipo de incrementadores puede estar en una opinión mucho extensa, y si utiliza un multiplicador muy grande se corre el riesgo de cometer errores al momento de calcular la polaridad total de la opinión.

Los multiplicadores de los decrementadores, siguen una lógica similar a la de los incrementadores, se entiende de los decrementadores “más allá” y “a pesar” reducen la polaridad de las palabras que están dentro del ámbito de aplicación de los decrementadores, por ejemplo “Más allá de ser un dispositivo incómodo y feo, es útil”, se debe reducir el puntaje de “incómodo” y “feo” por las situaciones explicadas con los incrementadores, pero en el caso de los decrementadores, el multiplicador debe estar entre 0 y 1. Dado lo mencionado, para los multiplicadores de los decrementadores de la próxima palabra se le asignó un valor de 0,5 por una decisión de que valga la mitad, y para los otros decrementadores siguiendo la misma lógica que los incrementadores, se asignó el valor el valor medio entre 0 y 0,5, es decir 0,25.

8.3 CONCLUSIÓN

Se ha explicado en detalle las reglas definidas para el español y la lógica aplicada para determinar los multiplicadores de cada una. Se puede apreciar que no fue necesario implementar muchas reglas en comparación con un enfoque basado íntegramente en reglas.

9. DESARROLLO - CAPÍTULO 7

9.1 INTRODUCCIÓN

En el presente capitulo detallara el caso de uso junto con el diseño de la aplicación, incluyendo los módulos desarrollados, diagramas de clases y diagrama de secuencia del mismo.

9.2 DESARROLLO

9.2.1 Caso de uso

Caso de Uso ID:	001		
Caso de Uso Nombre:	Consulta de opiniones		
Creado por:	Mariano Steininger	Ultima actualización por:	
Fecha Creación:	05/05/2016	Fecha última actualización:	

Actor:	Usuario de la aplicación	
Descripción:	Se consultan opiniones en redes sociales y se determina la polaridad de cada opinión.	
Precondiciones:	-	
Post-condiciones:	-	
Prioridad:	Media	
Frecuencia de uso:	Cada vez que el usuario desee consultar las opiniones sobre una marca o empresa	
Flujo Normal:	Actor	Sistema
	1. El usuario ingresa al sistema para consultar opiniones sobre una marca o empresa	

		2. El sistema solicita la marca o empresa que se desea consultar
	3. El usuario ingresa la marca o empresa que se desea consultar	
		4. El sistema obtiene las opiniones de la marca o empresa ingresada, determina su polaridad y muestra los resultados
	5. El usuario sale de la aplicación (1)	
	Fin de caso de uso	
Flujo alternativo 1	Actor	Sistema
	6. El usuario selecciona la opción “Volver” para realizar otra consulta	
		7. Vuelve a 2
Flujo alternativo 2	Actor	Sistema
	8. El usuario selecciona la opción “Ver detalle” para ver los textos clasificados junto con el puntaje asignado	
		9. El sistema muestra los textos clasificados junto con el puntaje asignado
Includes:		
Requerimientos		

No Funcionales	
Notas :	

9.2.2 Aplicaciones y módulos

Se han desarrollado dos aplicaciones:

1. Aplicación web que puede ser accedida desde cualquier browser y que expone un servicio REST para consultar las opiniones. Se decidió hacer un servicio REST para que también pueda ser consumido desde la aplicación mobile
2. Aplicación mobile que como se mencionó, consume el servicio REST expuesto.

La aplicación web ha sido desarrollada de forma modular utilizando Maven dado que permite un fácil manejo y administración del repositorio de dependencias externas y módulos de la aplicación. Por otro lado, la aplicación mobile se desarrolló utilizando Apache Cordova ya que utilizando los mismos archivos del front-ent del proyecto web (archivos de estilo, javascript, imágenes, etc) permite generar instaladores para los diversos dispositivos mobiles, facilitando así la portabilidad y el desarrollo de la aplicación.

Las principales razones por las que se decidió hacer los módulos implementados son para seguir el diagrama explicado en secciones anteriores y para que sea fácil de modificar: con la implementación realizada es posible modificar el stemmer, postagger, las reglas para que sea con otro idioma u otro modulo y lo único que se necesita hacer es implementar las interfaces definidas y modificar las dependencias Maven.

Las dependencias entre los módulos de la aplicación web se muestran en el siguiente diagrama:

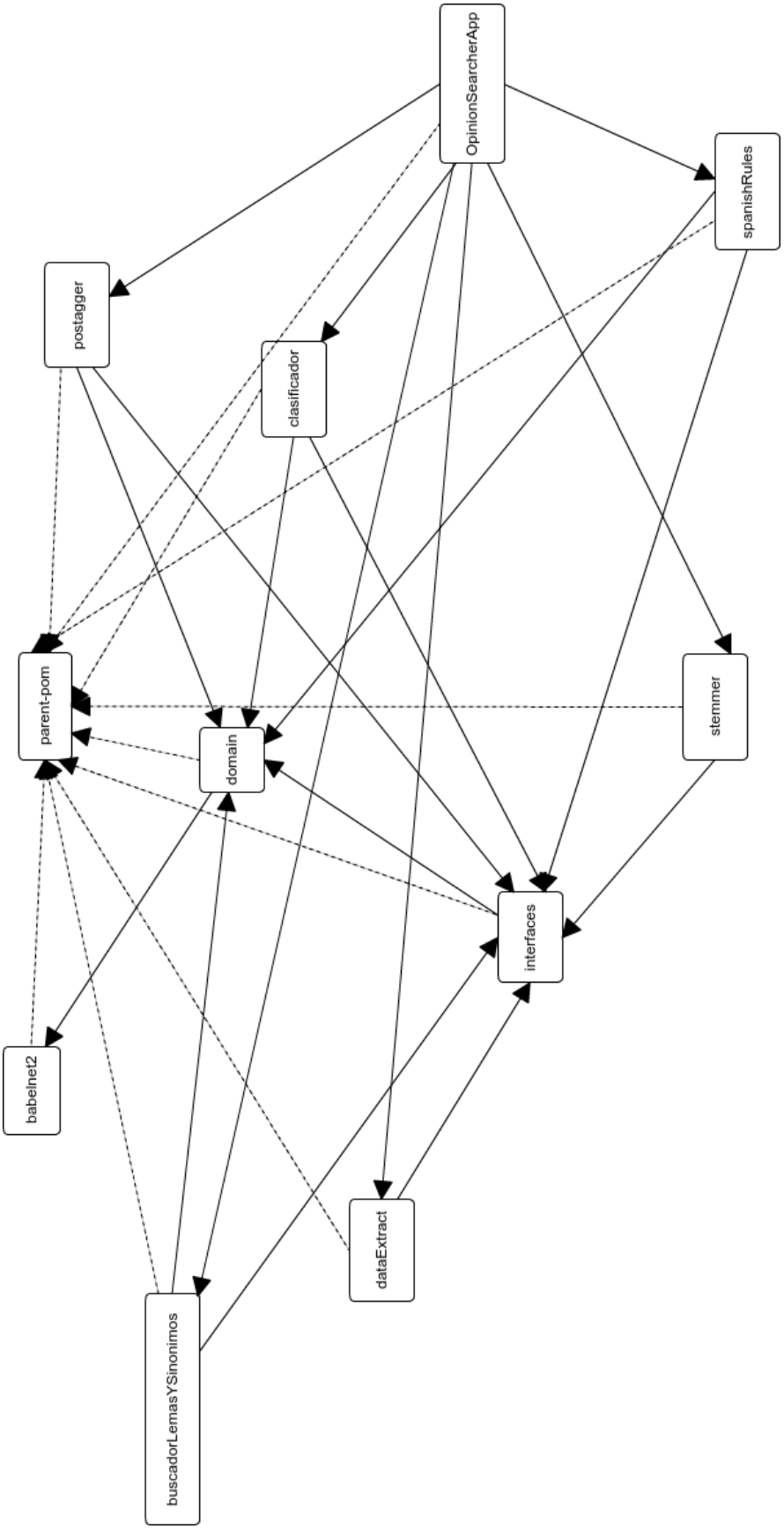

Figura 7: Diagrama de dependencias Maven

9.2.3 Diagramas de clase

Dado que son muchas interfaces y clases se expondrán diversos diagramas de clases:

- Del dominio general de la aplicación
- De las reglas implementadas
- Del modulo clasificador
- Del modulo para buscar los lemas y sinónimos (incluido el corrector ortográfico, traductor, etc)
- Que contenga la clase que recibe la petición del usuario y devuelve la respuesta, con el objetivo de indicar interfaces que utiliza y otras clases que no están en el resto de diagramas de clase).

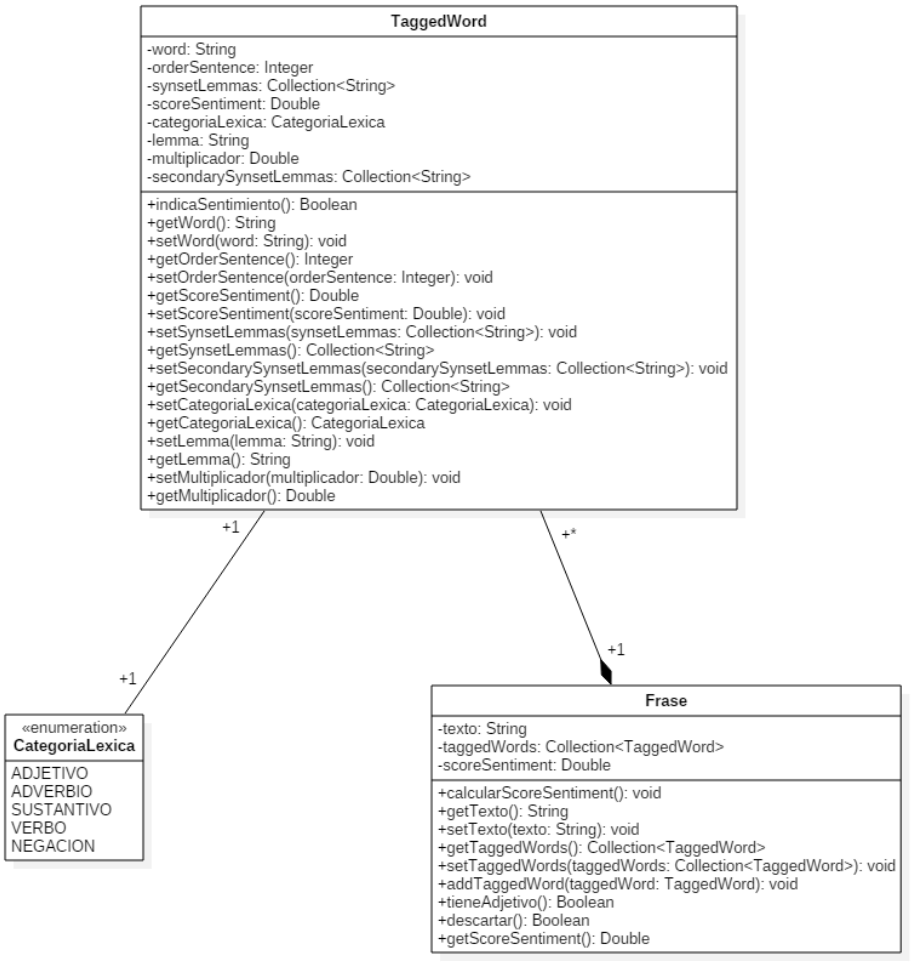

Figura 8: Diagrama de clases de dominio

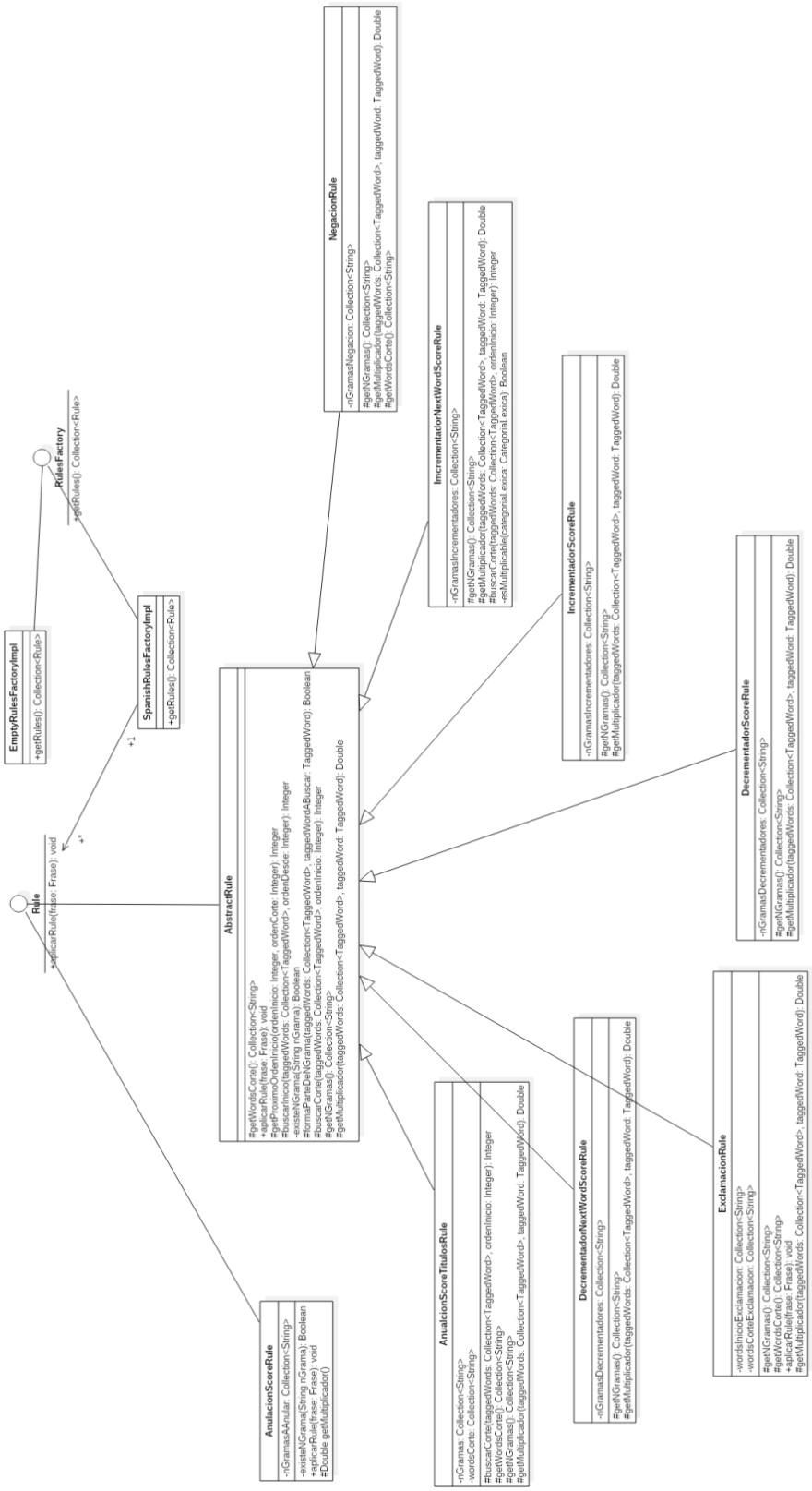

Figura 9: Diagrama de clases de reglas implementadas

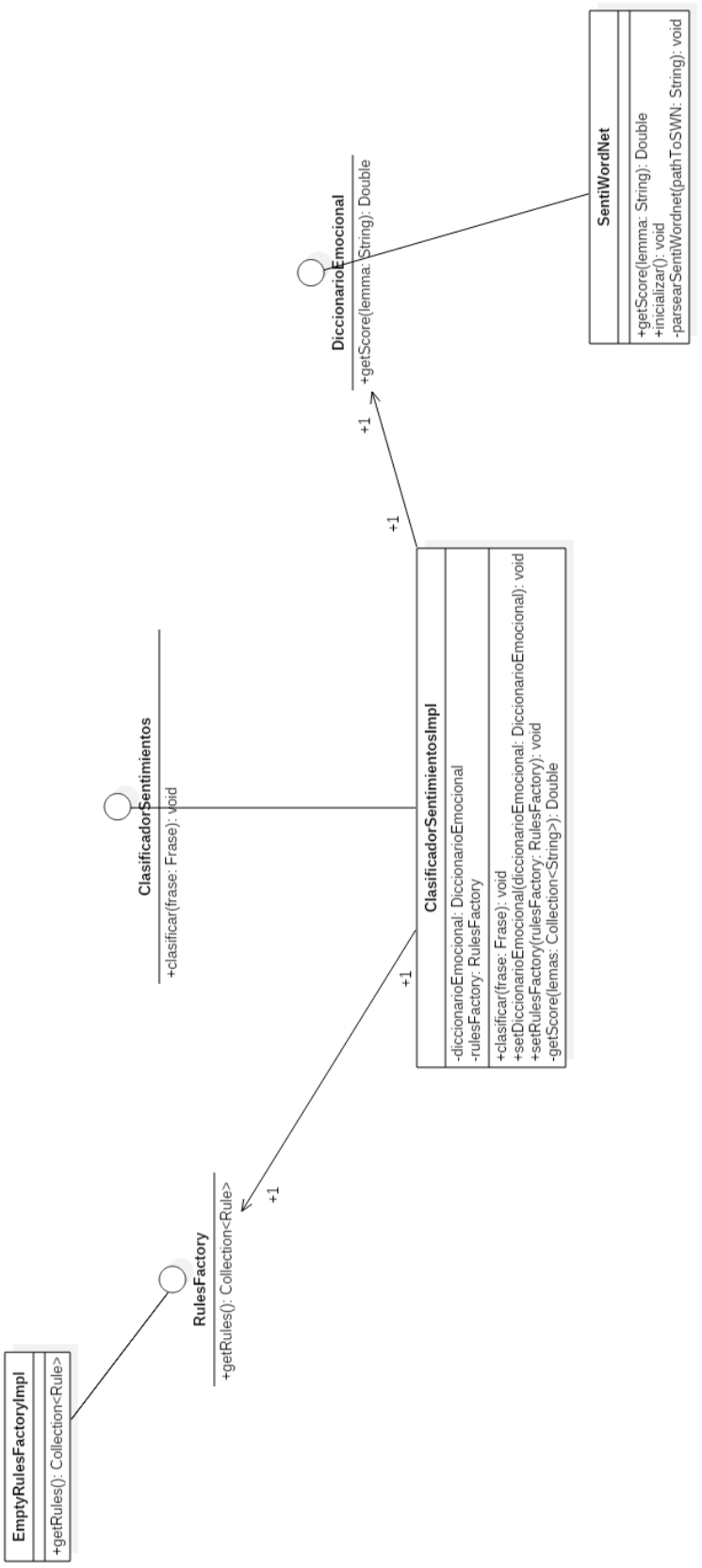

Figura 10: Diagrama de clases del módulo clasificador

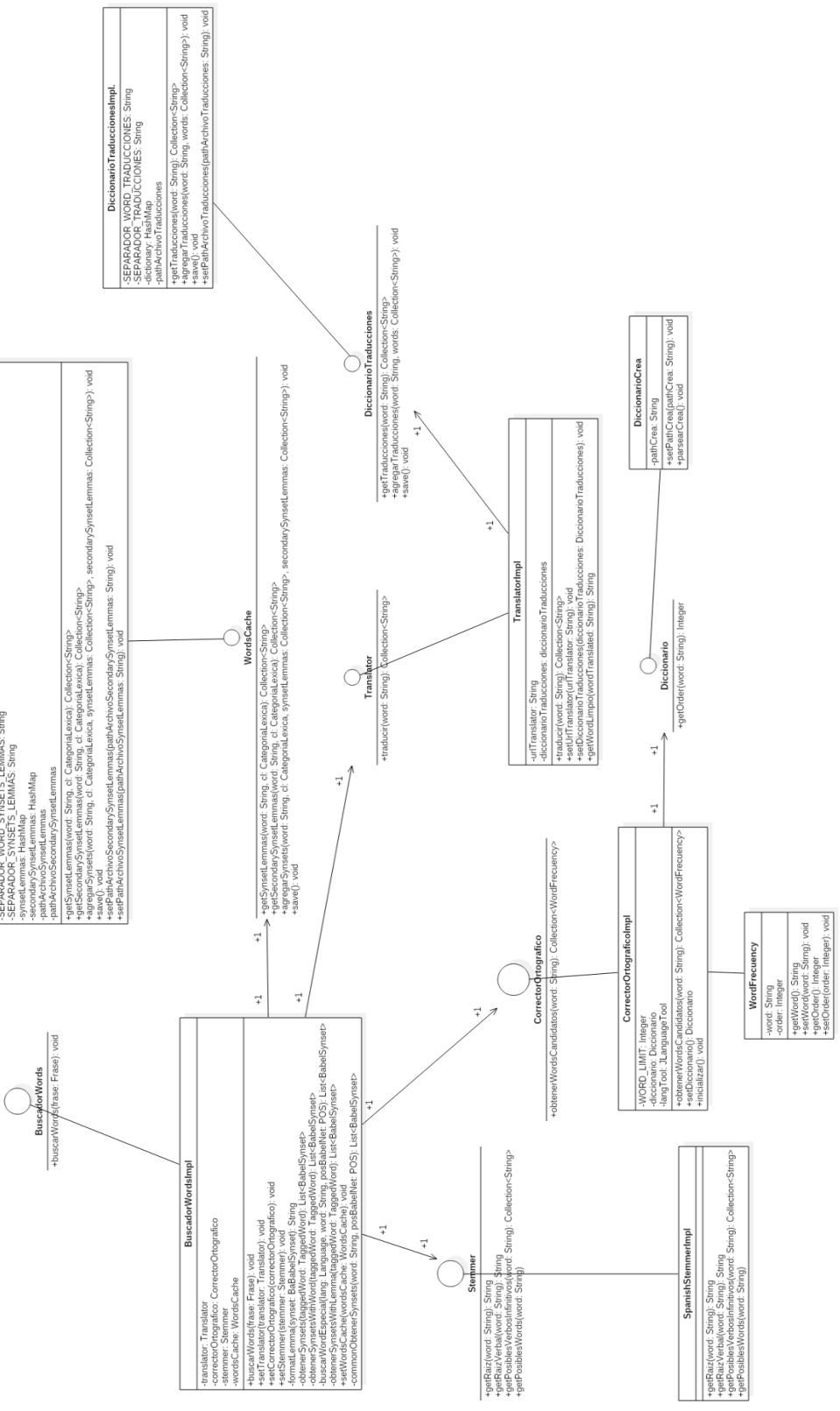

Figura 11: Diagrama de clases del buscador de lemas y sinónimos

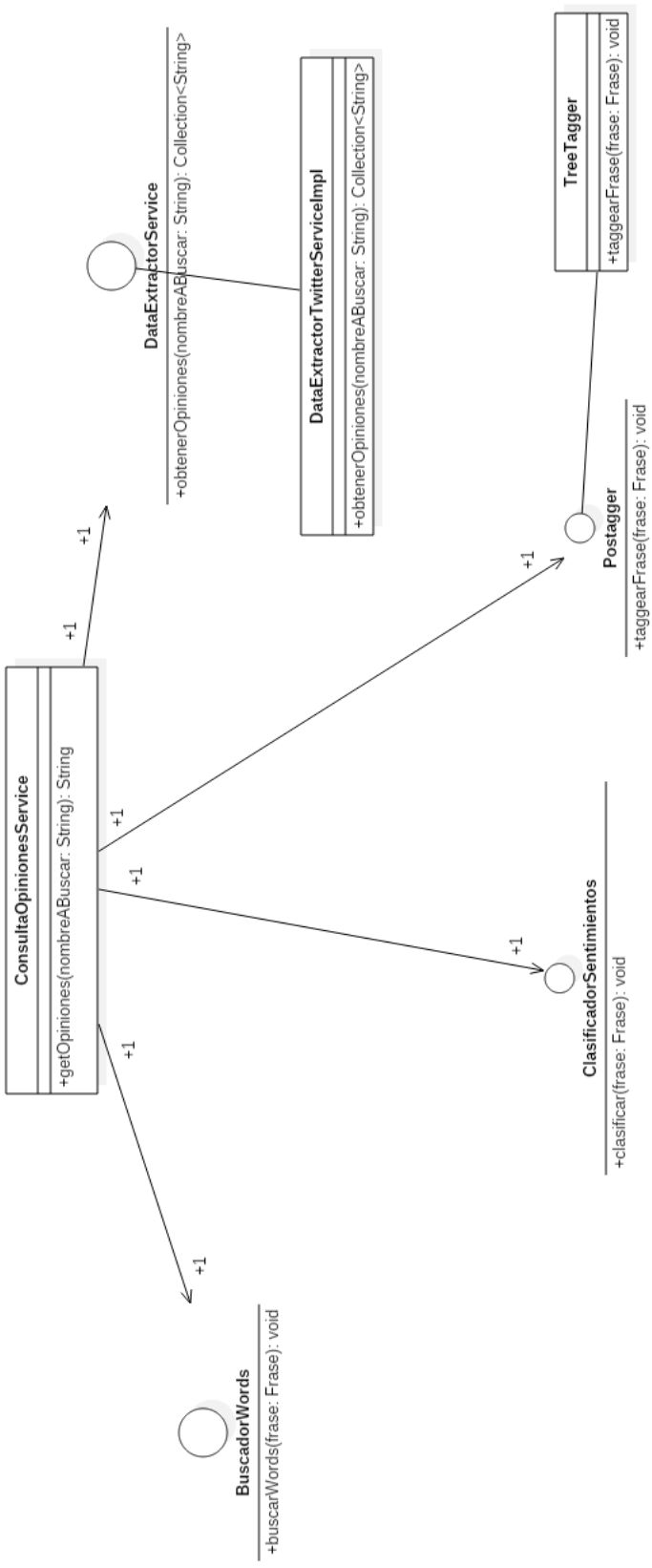

Figura 12: Diagrama de clases del caso de uso

9.2.4 Diagrama de secuencia

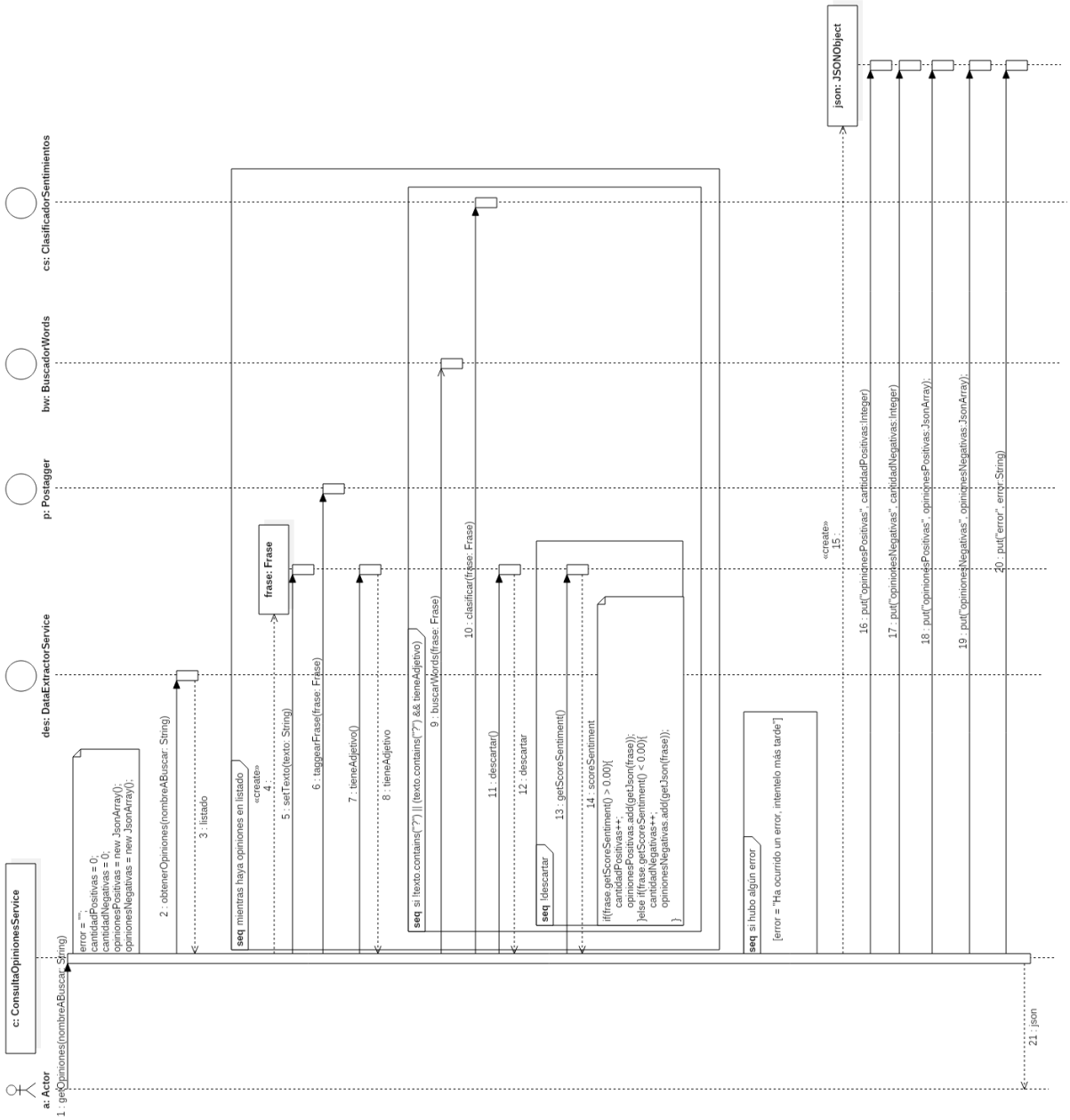

Figura 13: Diagrama de secuencia del caso de uso

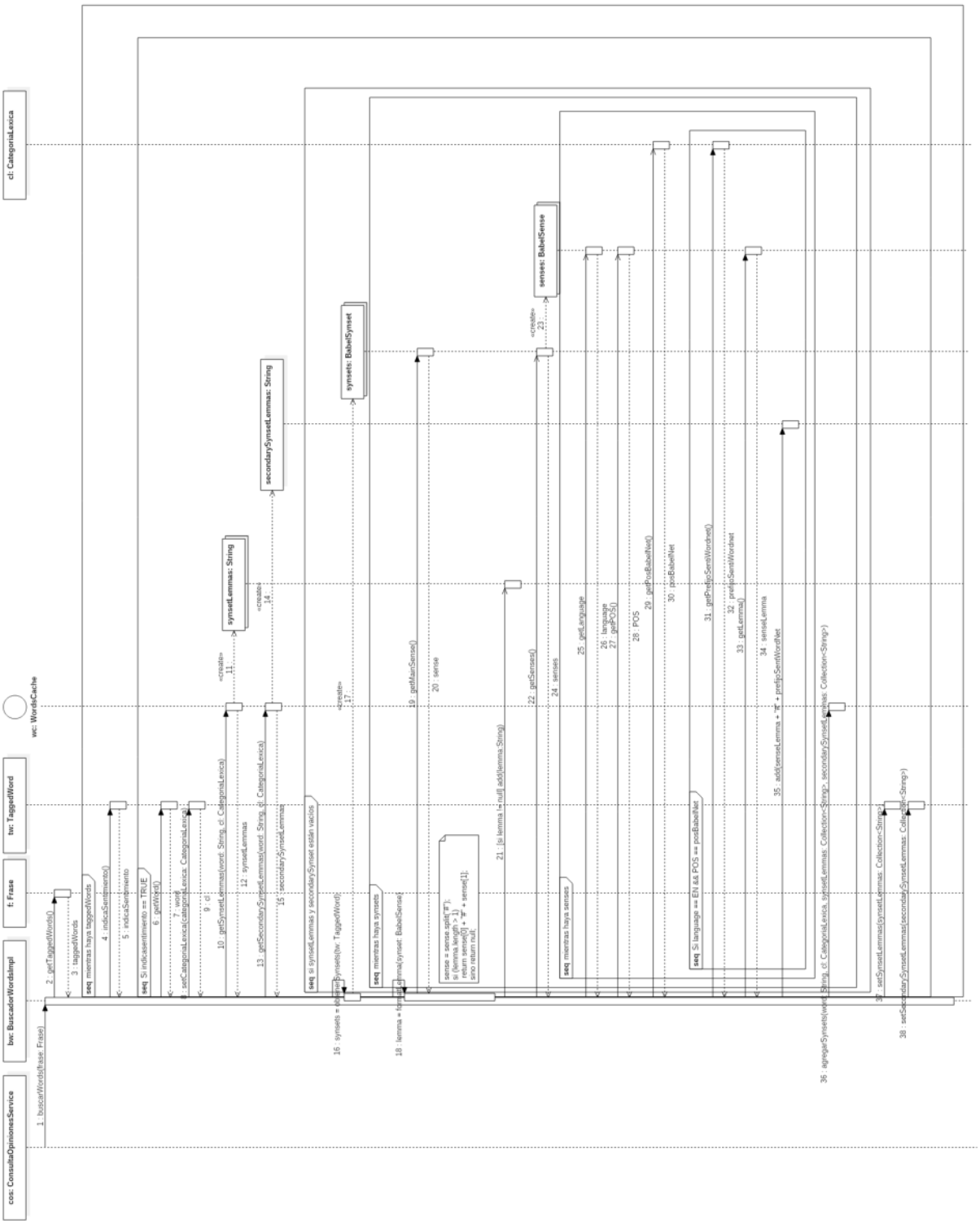

Figura 14: Diagrama de secuencia del método que busca los lemas y sinónimos de las palabras

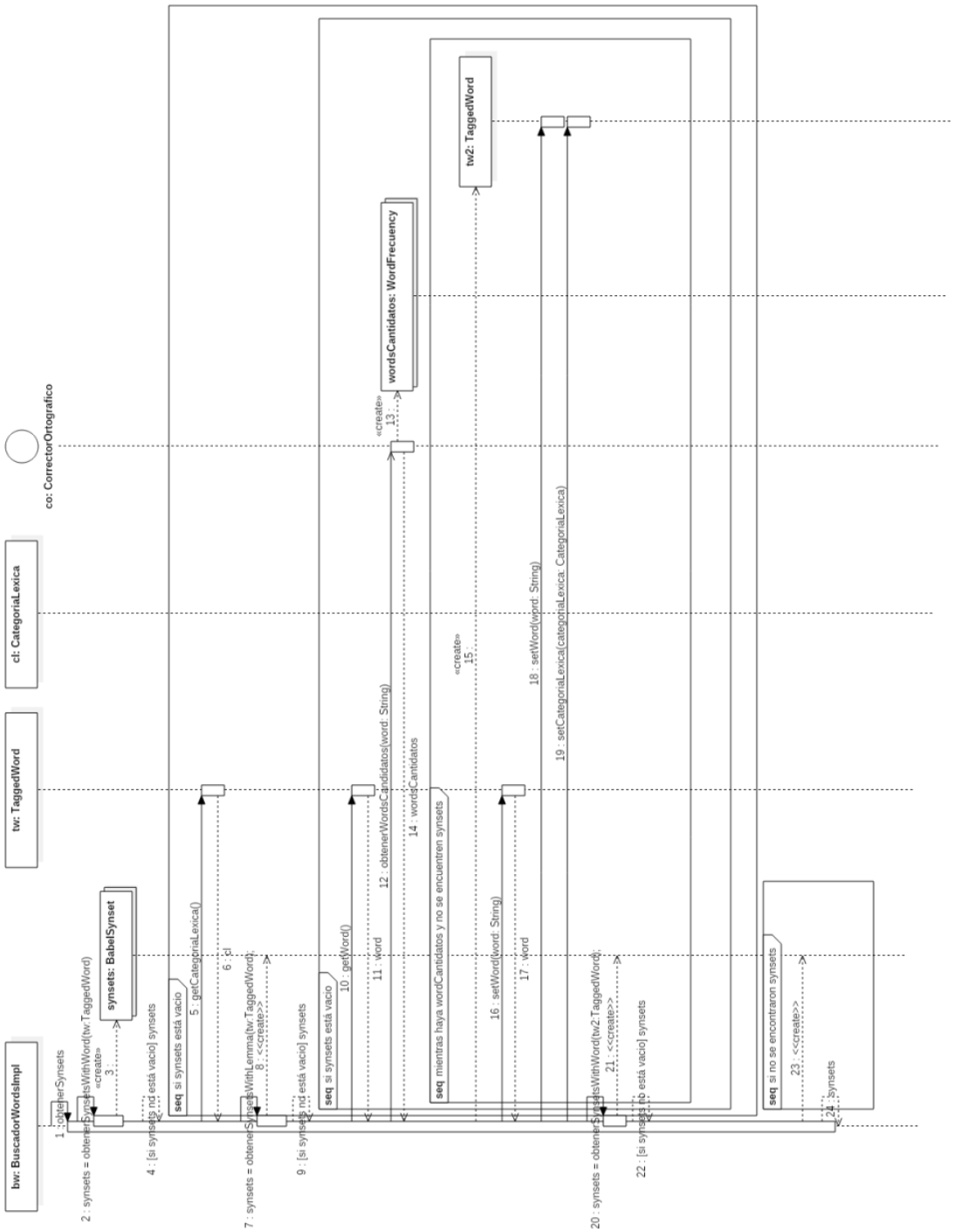

Figura 15 Diagrama de secuencia del método obtenerSynsets

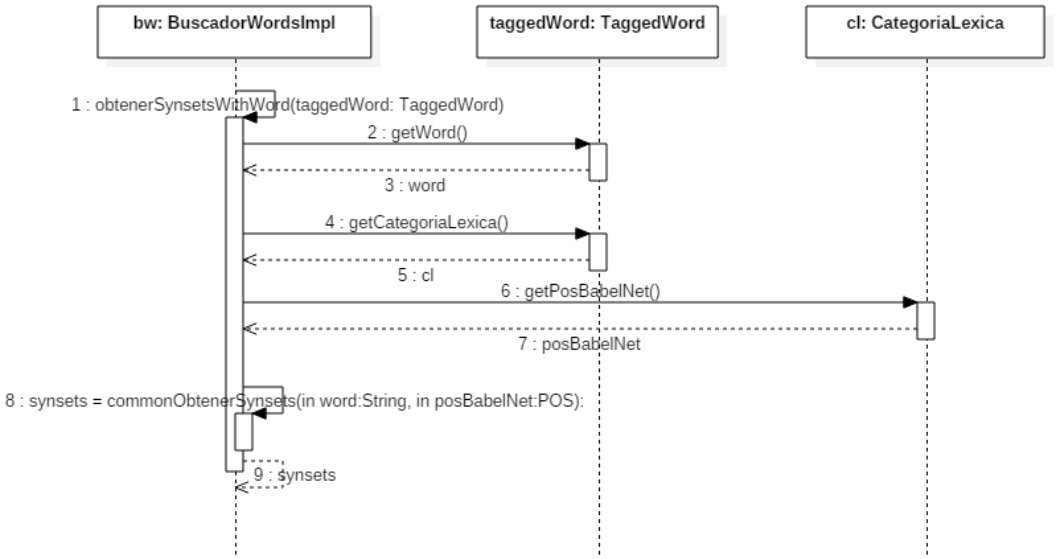

Figura 16 Método obtenerSynsetsWithWord

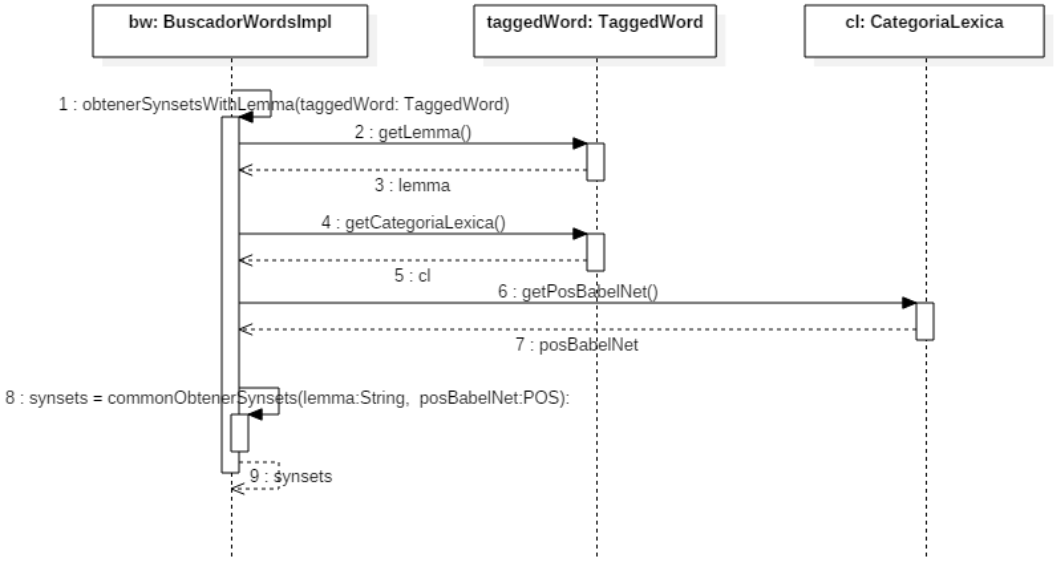

Figura 17 Diagrama de secuencia del método obtenerSynsetsWithWord

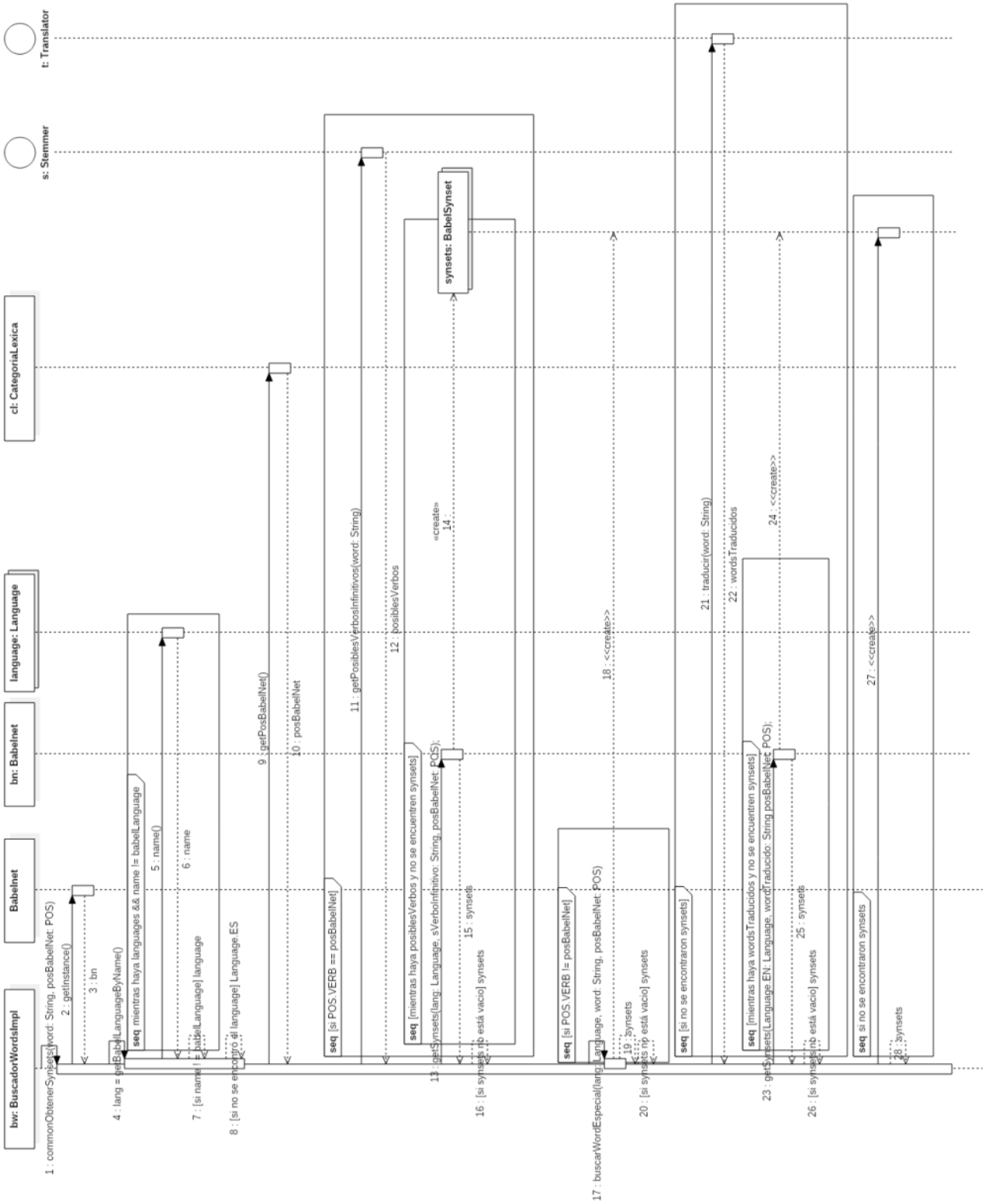

Figura 18 Diagrama de secuencia del método commonObtenerSynsets

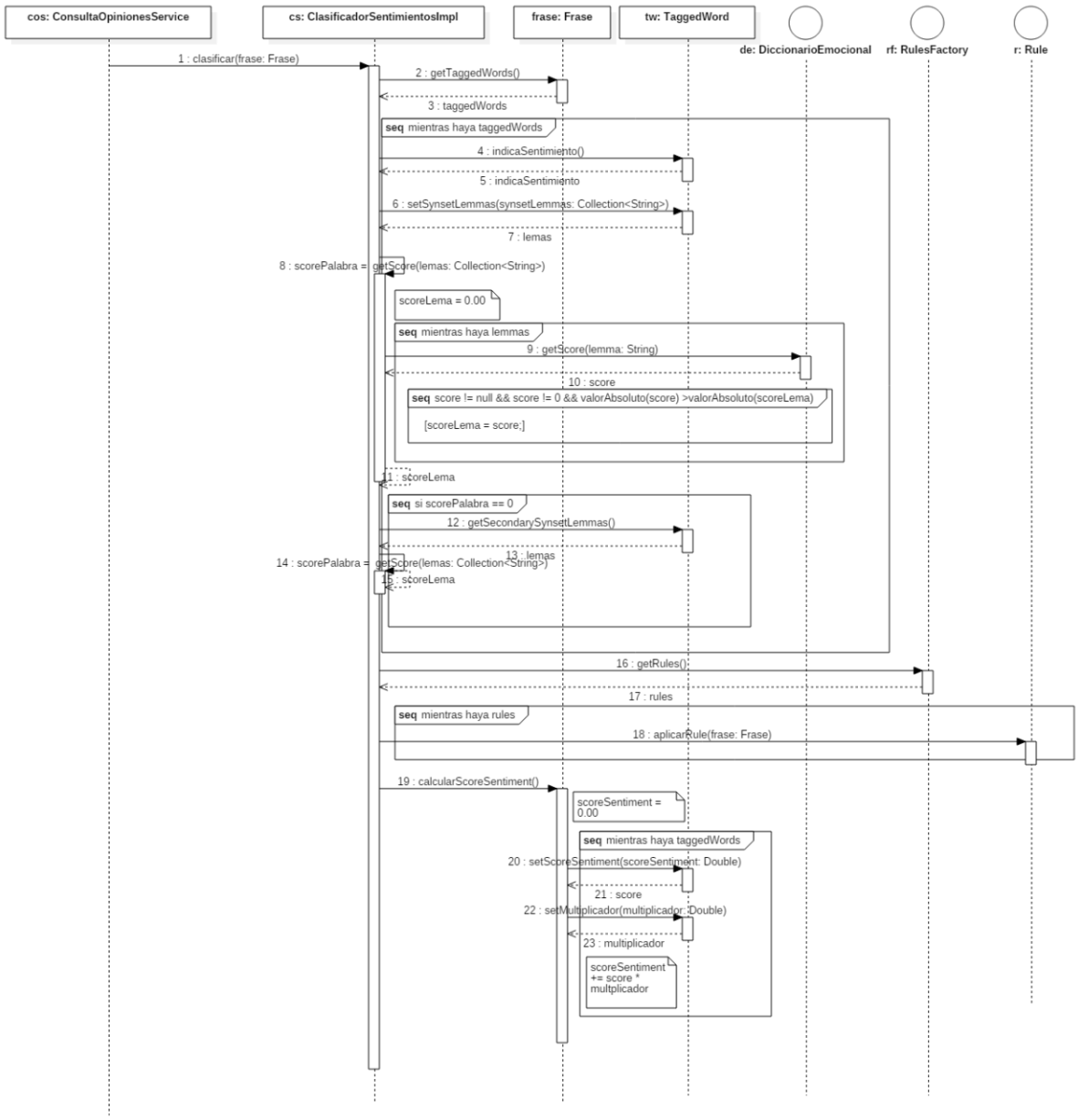

Figura 19: Diagrama secuencia del clasificador

9.3 CONCLUSIÓN

Como se explicó a lo largo del capítulo, la implementación se hizo de forma modular y abierta a poder extenderse a otros idiomas. Por otra parte, es sencillo reemplazar módulos, para por ejemplo, utilizar otro clasificador o postagger. El resultado fueron dos aplicaciones, una aplicación Web y otra mobile.

10. DESARROLLO - CAPÍTULO 8

10.1 INTRODUCCIÓN

En el presente capítulo se detallará la forma de trabajo utilizada para realizar el desarrollo de la aplicación.

10.2 DESARROLLO

A pesar que los requerimientos de la aplicación son claros, dada la complejidad del procesamiento del lenguaje natural, las diversas herramientas que hay en la actualidad y la incertidumbre acerca de la eficiencia del algoritmo pensado, entonces, se ha decidido por un proceso evolutivo, con el objetivo de lograr un prototipo.

Esta técnica de prototipado puede resumirse en el siguiente gráfico:

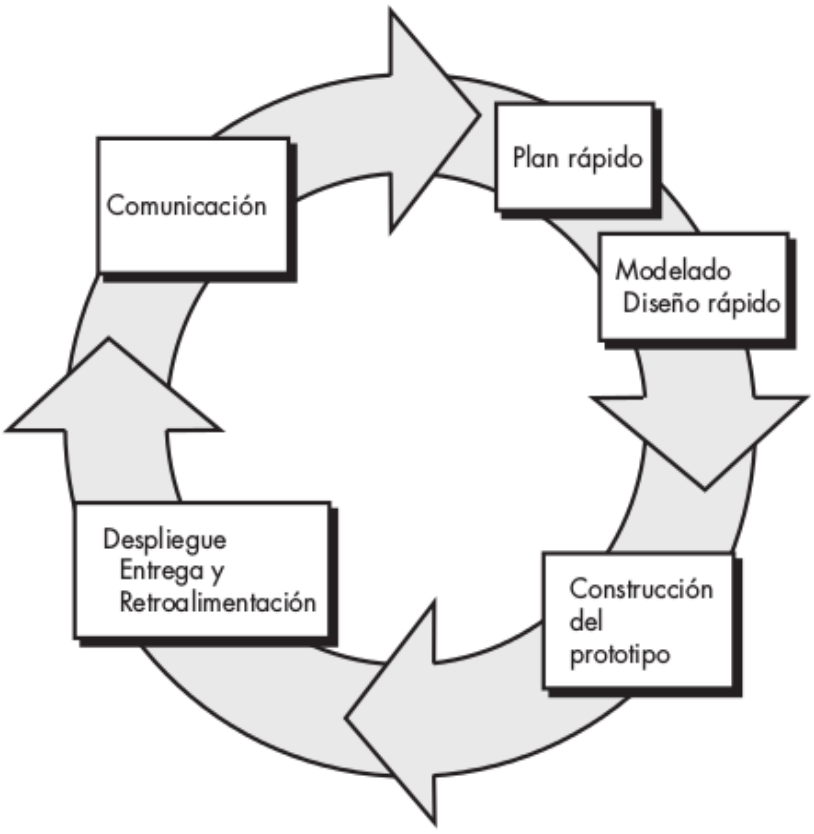

Figura 20: Técnica de Prototipado

Fuente: Pressman, 2010. Página 37 y 38

A continuación se detallará lo que hizo en cada etapa:

1. **Comunicación**: En esta etapa se definieron los requerimientos generales, y se estableció que el porcentaje mínimo de opiniones polarizadas correctamente debe ser superior al 70%
2. **Plan rápido**: En esta etapa se realizó una investigación inicial para determinar un modelo inicial para determinar la polaridad y las herramientas existentes que se podían utilizar para implementar el modelo inicial. Se definió primero realizar el algoritmo que polarice las opiniones y luego, una vez terminado, realizar la aplicación mobile.
3. **Diseño rápido**: Se definieron el modelo a implementar y las herramientas a utilizar, tanto para la primera versión como las posteriores. En la última etapa se hizo el diseño de la aplicación mobile.
4. **Construcción del prototipo**: En esta etapa se desarrolla lo definido en la etapa anterior.
5. **Entrega y Retroalimentación**: El cliente fue el mismo desarrollador del prototipo, donde lo que fue evaluado en cada versión del prototipo es:
    - Porcentaje de opiniones polarizadas correctamente
    - Razones por las que se obtenía una polarización incorrecta.
    - Performance del algoritmo

Más allá de que la aplicación final obtenga opiniones solo de Twitter se consideró conveniente que haya variedad de fuentes de las opiniones utilizadas en el conjunto de pruebas. El conjunto de pruebas estaba formado por más de 100 textos, en las que se encontraban:

- Opiniones obtenidas de Facebook
- Respuestas de un cuestionario de 5 preguntas realizado por el autor del presente trabajo
- Criticas cinematográficas

Para la evaluación de las opiniones mal polarizadas, fue necesario tener en cuenta que algunas opiniones iban a polarizarse mal por los puntajes que están en SentiWordNet. Una vez comprobado lo mencionado los refinamientos fueron los siguientes:

- Se detectó la necesidad de traducir ciertas palabras para poder buscarlas en BabelNet ya que al buscarlas en español no se obtenían resultados, es por ello que surgió el modulo denominado traductor
- Surgieron algunas reglas para el idioma español.
- Inicialmente no se utilizó TreeTagger para el postagging, se utilizó otro postagger. Al detectarse que habían opiniones que se polarizaban mal por un postaggeo erróneo, se investigó y se reemplazo por TreeTagger, obteniéndose así mejores resultados.
- Las primeras versiones de la implementación se demoraban en promedio 8 minutos ese tiempo no es aceptable para una aplicación para celulares; de este problema surgió el guardado de las traducciones y de las búsquedas a BabelNet.

10.3 CONCLUSIÓN

La forma de trabajo elegida permitió ir probando diferentes API's y formas de implementar los elementos del modelo, logrando así mejoras constantes a lo largo de todos los ciclos evolutivos del prototipo.

Se considera que ha sido una elección correcta.

11. DESARROLLO - CAPÍTULO 9

11.1 INTRODUCCIÓN

En este capítulo se expondrá la forma en que fue evaluada la implementación de modelo y los resultados obtenidos utilizando diferentes conjuntos de textos.

11.2 DESARROLLO

El modelo implementado se ha sido evaluado de dos formas:

- Utilizando un conjunto de textos extraídos manualmente de Facebook, criticas cinematográficas y una encuesta realizada por el autor del trabajo (Anexo 1)
- Utilizando textos extraídos de forma automática de Twitter (Anexo 2).

Dado que las personas se expresan de forma diferente de acuerdo al lugar dónde están emitiendo la opinión se optó por conformar un conjunto de textos de diferentes fuentes (primer conjunto mencionado) para ser utilizado durante la etapa de implementación y evolución del modelo. Utilizando el conjunto mencionado, se obtuvieron los resultados resumidos en la siguiente tabla:

Tabla VII: Datos de la evaluación del conjunto de textos seleccionados manualmente

Cantidad total textos	163
Cantidad textos polarizados correctamente	120
Porcentaje de textos polarizados correctamente	73,62%

Dado que la aplicación final obtiene de forma automática textos de Twitter (tweets) se decidió hacer un conjunto de textos integrado por dichos textos para evaluar el modelo implementado. La principal problemática de la obtención de forma automática de textos fue que había textos que no eran opiniones, para descartar la mayor cantidad de textos posibles se decidió descartar aquellos textos que:

- Tengan una url (indicado con un http), para evitar las publicidades. Demás está decir que los textos que tienen link muy difícilmente sean opiniones.
- Tengan el signo “?” y no tengan al menos un adjetivo, para descartar las consultas que se hacen vía Twitter.

- No tengan al menos una palabra cuyo puntaje obtenido en SentiWordNet sea mayor igual a 0.1 (valor absoluto).

El resultado obtenido de todo este procesamiento, es un conjunto de textos considerados opiniones por el autor, un conjunto de textos denominados indefinidos, que está compuesto textos que no son opiniones respecto la marca o empresa consultada y por un conjunto de textos descartados (se pueden ver los textos descartados en el Anexo II). El resultado de la evaluación del conjunto de textos considerados opiniones fue el siguiente:

Tabla VIII: Datos de la evaluación del conjunto de textos obtenidos automáticamente considerados opiniones

Cantidad total textos	256
Cantidad textos polarizados correctamente	181
Porcentaje de textos polarizados correctamente	70,7%

Por más que el conjunto de textos considerados indefinidos no sean opiniones propiamente dichas de la marca o empresa que se ha consultado es interesante notar que muchas expresan algún tipo de sentimiento, positivo o negativo, por lo tanto se pueden utilizar para evaluar el modelo implementado que determina la polaridad del texto. Se calculó la polaridad del conjunto de textos indefinidos y se obtuvo el siguiente resultado:

Tabla IX: Datos de evaluación del conjunto de textos obtenidos automáticamente considerados indefinidos

Cantidad total textos	98
Cantidad textos polarizados correctamente	67
Porcentaje de textos polarizados correctamente	68,37%

Juntando los resultados de los dos conjuntos de textos obtenidos de Twitter, se llega al siguiente resultado:

Tabla X: Datos de evaluación del conjunto de textos obtenidos automáticamente

Cantidad total textos	354
Cantidad textos polarizados correctamente	248
Porcentaje de textos polarizados correctamente	70,06%

Para cerrar esta sección cabe mencionar que la principal razón por la se calculan mal polaridades son los puntajes de las palabras en SentiWordNet. Esto no quiere decir que los puntajes sean incorrectos ya que una misma palabra puede ser utilizada en múltiples situaciones, dándole un significado positivo o negativo; otro factor a tener en cuenta es que los puntajes que están en SentiWordNet son para el idioma inglés y este no cuenta con tantas variaciones como el español.

11.3 CONCLUSIÓN

Se han mostrado los resultados de evaluar el modelo con diferentes conjuntos de datos, obteniendo un porcentaje acierto importante para que el modelo e implementación puedan ser considerados para futuros trabajos.

12. CONCLUSIONES

12.1 INTRODUCCIÓN

Una vez explicado todo lo que hay actualmente y haber expuesto un método cuantitativo para determinar la polaridad de las opiniones junto con su implementación y los resultados obtenidos, a modo de cierre del trabajo se darán las conclusiones a las que se llegó y el trabajo futuro que se puede hacer sobre el presente.

12.2 DESCRIPCIÓN

Dado que actualmente no es posible acertar un 100% la polaridad de todos los textos, al inicio del trabajo se puso como objetivo tener un porcentaje de acierto en el cálculo de la polaridad de los textos superior al 70%. Los resultados alcanzados evidencian que el objetivo ha sido cumplido. Más allá de lo comentado, cabe mencionar que al evaluar el conjunto de tweets indefinidos, dado que no todas expresan algún sentimiento, se esperaba obtener un porcentaje menor de acierto en comparación a los otros conjuntos de textos. El principal aporte de este trabajo entonces, es un modelo cuantitativo extensible a diferentes idiomas que puede adaptarse a cada uno de ellos mediante las reglas que se implementen. A modo resumen, el modelo tiene las siguientes ventajas y desventajas:

Ventajas:

- Extensible a diversos idiomas
- Flexibilidad para tener en cuenta las diferencias entre los idiomas (mediante reglas). Por ejemplo las negaciones son diferentes para el español y para el inglés.
- No es necesario elaborar un conjunto grande de reglas, sólo es necesario elaborar aquellas que resuelvan problemas que se puedan en gran medida generalizar, cómo por ejemplo las negaciones y los intensificadores (sean los que incrementan o decrementan)
- El determinar la polaridad de un texto de forma cuantitativa permite abordar un problema complejo como es el entendimiento del habla junto con sus subjetividades de una forma más medible y estructurada, permitiendo

detectar problemas en la implementación y mejorar la efectividad del algoritmo.

Desventajas

- Fuerte dependencia a los repositorios utilizados y la efectividad del algoritmo depende la calidad del contenido de estos repositorios.
- Necesidad de elaboración de reglas para cada idioma para mejorar la efectividad del algoritmo.
- Es difícil detectar opiniones neutras (puntaje calculado igual a 0), por lo tanto es necesario un buen filtro de las opiniones a procesar.

12.3 FUTURAS LÍNEAS DE INVESTIGACIÓN

- Es posible mejorar la implementación actual creando más reglas. Por ejemplo, dado que el presente es un trabajo académico, se decidió no incluir una regla para improperios, esta regla permitiría detectar opiniones negativas por la presencia de éstas palabras. En el presente trabajo se hizo foco en las problemáticas más comunes: negaciones e intensificaciones.
- Detección del sarcasmo y modalidad, es lo que ha quedado fuera del alcance, es posible crear una o mas reglas que detecten el sarcasmo y modalidad.
- Cuando se extraen textos de una determinada marca o empresa y el texto a analizar hace referencia a los mismos junto con otras marcas o empresas el algoritmo actual determina la polaridad del conjunto, sin detectar las entidades afectadas por cada palabra, por ejemplo, si se está buscando opiniones de “Personal” y se está analizando el texto “Movistar y Claro son un desastre, su servicio es pésimo y tienen muy mala atención al cliente, no así cómo Personal” el algoritmo actual va a determinar que la opinión es negativa ya que prevalecen palabras con orientación negativa, cuando en realidad la opinión es positiva ya que se está buscando textos de Personal y no de Movistar o Claro. Esta problemática podría ser resuelta agregando más reglas.
- Dado que el foco del trabajo está puesto en el cálculo de la polaridad, la aplicación resultante extrae textos de Twitter y le aplica filtros simples, es

posible mejorar el filtro para un mejor descarte de textos que no sean opiniones.

13. REFERENCIAS BIBLIOGRÁFICAS

Alphabetical list of part-of-speech tags used in the Penn Treebank Project [en línea]. © 2003 [consulta: 10 febrero 2016]. Disponible en: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Baum, Leonard E. y Petrie, Ted, Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics. 1966. Vol. 37, no. 6, p. 1554-1563. DOI 10.1214/aoms/1177699147. Institute of Mathematical Statistics

BabelNet™ | The largest multilingual encyclopedic dictionary and semantic network, 2016. Babelnet.org [En linea], [consulta: 15 enero 2016]. Disponible en: http://babelnet.org/guide

Baccianella Stefano, Esuli Andrea y Sebastiani Fabrizio. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010), Valletta, MT, 2010, pp. 2200-2204

Baker, et. al, Modality and Negation in SIMT Use of Modality and Negation in Semantically-Informed Syntactic MT. *Computational Linguistics*. 2012. Vol. 38, no. 2, p. 411-438. DOI 10.1162/coli_a_00099. MIT Press – Journals

Balahur, Alexandra *et al*. Emotinet: A knowledge base for emotion detection in text built on the appraisal theories. In Natural Language Processing and Information Systems. Alicante: España. Springer Berlin Heidelberg, 2011. pp 27–39. DOI 10.1007/978-3-642-22327-3_4

Barbieri, F.; Ronzano, F. & Saggion, H. (2015), 'Is this Tweet Satirical? A Computational Approach for Satire Detection in Spanish.',

Procesamiento del Lenguaje Natural 55 , pp. 135-142 . ISSN: 1135-5948

Bayes, Mr. y Price, Mr., An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. Philosophical Transactions of the Royal Society of London. 1763. Vol. 53, no. 0, p. 370-418. DOI 10.1098/rstl.1763.0053. The Royal Society

Broke, Julian , *A Semantic Approach to Text Sentiment Analysis*. Simon Fraser University, 2009.

Carter, Ronald and McCarthy, Michael, 2006, *Cambridge grammar of English*. Cambridge [England] : Cambridge University Press. ISBN: 978-0521588461

Choi, Yoonjung y Wiebe, Janyce. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. EMNLP 2014:. ISBN 1181-1191

Church, Kenneth Ward, 1988, A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the second conference on Applied natural language processing* -. 1988. P. 136 - 143. DOI 10.3115/974235.974260. Association for Computational Linguistics (ACL)

Ding, Xiaowen, Liu, Bing y Yu, Philip S., A holistic lexicon-based approach to opinion mining. Proceedings of the international conference on Web search and web data mining - WSDM '08. 2008. P. 231 – 240. DOI 10.1145/1341531.1341561. Association for Computing Machinery (ACM)

Halteren, Hans. Syntactic Wordclass Tagging. Dordrecht: Springer Netherlands, 1999. pp. 217-246. ISBN 978-94-015-9273-4;

Hu, Minqing and Liu, Bing. Mining and summarizing customer reviews. Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04. 2004. pp 168-177. DOI 10.1145/1014052.1014073. Association for Computing Machinery (ACM)

Ibañez, Jesus, Serrano, Oscar, y García, David. Emotinet: A framework for the development of social awareness systems. In Awareness Systems. Springer Berlin Heidelberg, 2009. pp. 291–311. ISBN: 978-1-84882-476-8

Jaffe, Eric, 2016, Why You're Bad At Understanding Irony. *Co.desing* [En linea]. 2016. [Consulta: 5 de Febrero 2016]. Disponible en: http://www.fastcodesign.com/3030622/evidence/why-youre-bad-at-understanding-irony

Jurafsky, D. & Martin, J. H., Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2nd ed, New Jersey: Prentice Hall, 2009. p 137 - 167 . ISBN 978-0131873216

Kamps, Jaap et. al. .Using wordnet to measure semantic orientation of adjectives. In Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, 2004, volumen IV, pp 1115-1118.. [consulta: 29 marzo 2016]. Disponible en http://humanities.uva.nl/~kamps/publications/2004/kamp:usin04.pdf

Karlsson, Fred. Constraint grammar. Berlin: Mouton de Gruyter, 1995. pp. 165-284. ISBN: 3110141795

Kempe, André. A probabilistic tagger and an analysis of tagging errors. Technical report, Institut für Machinelle Sprachverarbeitung, Universität Stuttgart, Germany, 1993

Kennedy, Alistair and Inkpen, Diana, 2006, SENTIMENT CLASSIFICATION of MOVIE REVIEWS USING CONTEXTUAL VALENCE SHIFTERS. *Computational Intell*. 2006. Vol. 22, no. 2, p.110-125. DOI 10.1111/j.1467-8640.2006.00277.x. Wiley-Blackwell

Kerstin Denecke. Using sentiwordnet for multilingual sentiment analysis. In Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on, pages 507–512. IEEE, 2008

Kim, Soo-Min y Hovy, Eduard, 2004, Determining the sentiment of opinions. Proceedings of the 20th international conference on Computational Linguistics - COLING '04. Association for Computational Linguistics (ACL). 2004. p 1367. DOI 10.3115/1220355.1220555.

Lakoff, George. Hedges: A study in meaning criteria and the logic of fuzzy concepts. J Philos Logic. 1973. Vol. 2, no. 4, p. 458–508. DOI 10.1007/bf00262952. Springer Science + Business Media

Lematización. Lematización (Universidad de Costa Rica) [En linea]. [Consulta: 30 de enero 2016. Disponible en http://maanvn.wix.com/lematizacion#!tecnicas/c19ov

Lovins, Julie Beth. Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics, Vol No.11, Issue No.1, 1968 pp 22-31.

Miller et al., 1990, Introduction to WordNet: An On-line Lexical Database *. *International Journal of Lexicography*. 1990. Vol. 3, no. 4, p. 235-244. DOI 10.1093/ijl/3.4.235. Oxford University Press (OUP)

Nishio, Minoru, Iwabuchi, Etsutaro y Mizutani, Shizno. The Japanese language dictionary. 5a ed. Tokyo: Iwanami, 1994.

Panessi, W., Bordignon, F. R.: “Procesamiento de variantes morfológicas en búsquedas de textos en castellano”, Revista Interamericana de Bibliotecología, Vol. 24, No. 1, 2001, pp. 69 – 88.

Pang, Bo, Lee, Lillian y Vaithyanathan, Shivakumar. Thumbs up?. Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02. 2002. pp 79-86. DOI 10.3115/1118693.1118704. Association for Computational Linguistics (ACL)

Paniagua Bernardo, J., López García, A. y Gallardo-Paúls, 1er B. 2005. Conocimiento y lenguaje. Valencia: Universidad de Valencia, 2005. p 415 ISBN: 978-84-370-6113-9.

Polanyi, Livia and Zaenen, Annie, 2006, Contextual Valence Shifters. *The Information Retrieval Series*. 2006. P. 1-10. DOI 10.1007/1-4020-4102-0_1. Springer Science + Business Media

Porter, M.F., 1980, An algorithm for suffix stripping. Program: electronic library and information systems. 1980. Vol. 14, no. 3, p. 130-137. DOI 10.1108/eb046814. Emerald

Potts, Christopher. On the negativity of negation. In Nan Li and David Lutz, eds. *Proceedings of Semantics and Linguistic Theory 20*. Ithaca, NY: CLC Publications: 2011. pp. 636-659. ISBN 2163-5951

PRESSMAN, Roger S. Ingeniería del Software. Un enfoque práctico. 7ma ed. México: McGraw-Hill, 2010. 805 p. Paginas 37 y 38 ISBN 978-607-15-0314-5

Quinlan, J. Ross, 1983, Learning Efficient Classification Procedures and Their Application to Chess End Games. *Machine Learning*. 1983. P. 463-482. DOI 10.1007/978-3-662-12405-5_15. Springer Science + Business Media

Quirk, Randolph et. al 1985, *A Comprehensive grammar of the English language*. London : Longman.ISBN: 9780582517349

Ratnaparkhi, Adwait, 2000, Syntactic Wordclass Tagging Hans van Halteren (editor) Dordrecht : Kluwer Academic Publishers, 1999, pp 103-131 ISBN 0-7923-5896-1. Computational Linguistics.. MIT Press – Journals

Saurí, Roser y Pustejovsky, James, 2009, FactBank: a corpus annotated with event factuality. *Lang Resources & Evaluation*. 2009. Vol. 43, no. 3, p. 227-268. DOI 10.1007/s10579-009-9089-9. Springer Science + Business Media

Schmid Helmut, Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester: UK. 1994. P. 44 - 49. DOI 10.1.1.28.1139.

Scholkopf, Bernhard y Smola, Alex. Learning with Kernels. MIT Press, Cambridge, MA, 2002. p. 2. [consulta: 15 marzo 2016]. Disponible en http://alex.smola.org/papers/2002/SchSmo02b.pdf

Sentiment Symposium Tutorial: Lexicons. *Sentiment.christopherpotts.net* [En linea], [consulta: 5 abril 2016]. Disponible en: http://sentiment.christopherpotts.net/lexicons.html

Sidorov, Grigori et. all, Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets. Advances in Artificial Intelligence. Springer Science + Business Media, 2013. DOI10.1007/978-3-642-37807-2_1.

Spanish stemming algorithm. Snowball.tartarus.org [En linea], [consulta: 20 enero 2016]. Disponible en: http://snowball.tartarus.org/algorithms/spanish/stemmer.html

Stome. Philip J. et al, General Inquirer: A Computer Approach to Content Analysis. 1st ed. Oxford, England: MIT Press, 1966. ISBN 978-0262690119

Takamura, Hiroya, Inui, Takashi y Okumura, Manabu, 2005, Extracting semantic orientations of words using spin model. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05. Association for Computational Linguistics (ACL). 2005. pp 133–140 DOI 10.3115/1219840.1219857.

Tausczik, Y. R. and Pennebaker, J. W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology. 2009. Vol. 29, no. 1, p. 24-54. DOI 10.1177/0261927x09351676. SAGE Publications

Turney, Peter D., Thumbs up or thumbs down?. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. Association for Computational Linguistics (ACL), 2002. . DOI 10.3115/1073083.1073153.

Turney, Peter D. y Littman, Michael L. Measuring praise and criticism. ACM Transactions on Information Systems. 2003. Vol. 21, no. 4, pp. 315-346. Association for Computing Machinery (ACM) , 2003. DOI 10.1145/944012.944013.

Vapnik, Vladimir. N., y Chervonenkis, Alexey Ya. *Teoriya raspoznavaniya obrazov: Statisticheskie problemy obucheniya*. (en ruso) [Theory of pattern recognition: Statistical problems of learning]. Moscú: Rusia, 1974

Voutilainen, A. (1995). Morphological disambiguation. En Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A.(Eds.), Constraint Grammar: A Language Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin. pp. 165- 284. ISBN: 3110141795

Voutilainen, A. Handcrafted rules. In van Halteren, H. (Ed.), Syntactic Wordclass Tagging, pp. 217-246. Kluwer, Dordrecht, 1999

Wilson, Theresa, Wiebe, Janyce y Hoffmann, Paul. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language procesing. Association for Computational Linguistics, 2005. pp 347–354. DOI 10.3115/1220575.1220619

Wilson, Theresa, Wiebe, Janyce y Hoffmann, Paul. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. Computational Linguistics. 2009. Vol.35, no. 3, p.399-433. DOI 10.1162/coli.08-012-r1-06-90. MIT Press – Journals

WNSTATS(7WN) manual page [En linea], [consulta: 5 abril 2016]. Disponible en: http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

14. ANEXOS

Los anexos del presente trabajo contienen los conjuntos de textos utilizados para probar la efectividad del algoritmo desarrollado junto con el valor de polaridad asignado por el mismo separados los que se clasificaron bien de los que se clasificaron mal. En el Anexo 1 se encuentran los textos seleccionados de forma manual, mientras que en el Anexo 2 se encuentran los textos extraidos de forma automática utilizando la API de Twitter (para este caso también están los tweets descartados)

14.1 ANEXO 1 – CONJUNTO DE TEXTOS SELECCIONADOS MANUALMENTE

Este anexo contiene los textos extraídos manualmente junto con el puntaje calculado por la implementación. Por un lado estarán los que se clasificó bien y por otro los que se clasificó mal.

Clasificados bien:

1.8914-¡El mejor juego del verano de 2014! Sin descargas, juega directamente en tu navegador
-9.7247-Meses y cada vez peor el servicio de @MovistarArg hace días prometen solucionar algo que no pasa !!! Existe el 3g
-2.8997-Es increible como decayó esta empresa @MovistarArg aun con mas usuarios el servicio es pesimo! Eso si, la factura te llega SI O SI.
2.0857-A mi me parece un excelente móvil. La batería dura aproximadamente un día con mucho uso.
0.5545-Un móvil estupendo en todos los sentidos. La única pega y preocupación,es que a veces se sobrecalienta. Espero que se solucione
0.6941-Hermosa foto
-0.055-Speedy, siempre tenia problemas y se cae el servicio, ademas q te obliga a tener una linea telefonica.
0.8648-Puntualmente, en los últimos 2 años tuve fibertel 3mb, y siempre me funcionaron a la perfeccion, es mas soy adicto a descargas de juegos y peliculas en alta definicion (1080 y 720p).

-0.9863-Arnet fue lo peor que me pasó en la vida, el soporte asqueroso, el servicio pésimo
1.209-Tuve Fibertel y para mi es el mejor proveedor de Internet que hay en Argentina comparado con los otros dos.
-0.6347-Que mal se vio Fernando Mayans.
1.0952-Decir lo que queráis de este tio, yo le estaré eternamente agradecido.
5.1174-La verdad que el sabor de Cacao que tiene está mucho mejor conseguido, es como más redondo y recuerda (un poco de lejos) al sabor de una taza de chocolate.
0.1932-Está para chuparse los dedos
1.4067-Da energia, es de Nestlé, producto de gran confianza
0.6098-Está rico
-0.1105-Pésimas propiedades nutritivas
1.2634-Sudamericano: Argentina le ganó bien a Uruguay y está en la final: La Selección
-2.322-UNA BASURA PERSONAL, soy de Salta y es imposible llamar, recibir mensajes.....un desastre !
-0.3506-Personal no serivis para mierda.. Me robas el credito tu internet es el mas lento q la misma m..... Ya me cambio de empresa... Yo les recomiendo a todos q se cambien de empresa
-3.0325-Hoy a la manana le cargue 50 pesos al celular, lo voy a usar por primera vez a la tarde para llamar a una amiga que tuvo familia y me dice que no tengo saldo suficiente para realizar el llamado. Que paso? Tan malo puede ser el servicio de esta empresa? Ni siquiera se si me van a dar la atencion que reclamo porque les mande una queja y hasta mi dni me pidieron, pero dudo mucho que alguien se comunique conmigo o me mande el resumen de gastos de mi celular como pedi. Una verguenza lo de Personal y como roban plata.
0.4814-Los q dicen q las hamburguesas del mc son una porqueria es xq no probaron un angus! :D a mi me encanta el tasty... ñam ñam!
-0.7235-Lo único feo del Mc son los cafés
-0.5773-NINGUNA..... !!! pésima la atención las dos ultima vez que fui me dieron fría y una cargada de hamburguesa cada vez mas chica...horribleeee
0.6415-Buenísimo McDonald sos lo mas
1.3244-Peanut Butter Cookies: crujientes, deliciosas, perfectas para una tarde con amigos.
0.73-Una de mis bebidas preferidas

-2.454-El viaje tiene horario incierto de salida, en los aviones los baños sucios, en mas de una vez la atencion es mala, desaparecieron en un viaje de 15 horas. Es de entender pues sin exajerar el 60% de los pasajeros no hacen caso y son irrespetuosos, embroman continuamente con su equipaje, cuando piden que se queden sus asientos justo le viene ganas de ir al baño o sacar algo del boslo, son sucios y dejan cosas tiradas en el baño, asquerosos al comer e irrespetuosos cuando piden algo, todo es un circulo hay que tener estomago una cosas es atencion, servir pero otra ser denigrado.
-2.0081-He dejado de volar en Aerolíneas porque el servicio que "sufrí" fué vergonzoso.
3.6561-Ahora prefiero a LAN, que es una empresa responsable y con muy buen servicio nacional e internacional.
-0.6045-Dan lástima. Hay que cerrarla y liberar el mercado subvencionando los vuelos deficitarios (que no haya regiones incomunicadas)
-2.9582-Pésimo, mal administrada por gente incompetente, con personal en función política, es una sangría inútil. Teniendo en cuenta el Inventario que se hizo en su momento las piezas de respuesto que e afanaron los muchachos, bastan para construir varias escuelas en Tartagal.
1.6646-Rápido, buen sabor, sencillo de preparar, varios tamaños disponibles.
0.9124-Sabor inconfundible, sencillo y rápido de prepara.
-0.1869-Una vez abierto el envase pierde calidad.
-0.5041-Gracias @MovistarArg por andar mal casi todo el tiempo.
-2.8997-Es increible como decayo esta empresa @MovistarArg aun con mas usuarios el servicio es pesimo! Eso si, la factura te llega SI O SI.
-0.3399-No crei llegar a esto @MovistarArg .... un desastre en varios aspectos. ..desde servicio de atención al cliente hasta prestaciones generales.
-1.0214-Un día duro el servicio aceptable de @MovistarArg. Ya anda muy mal, otra vez.
-0.4911-@MovistarArg no funciona el servicio de internet. No funcionó durante todo el día. URGENTE
-0.7237-@MovistarArg Son la basura mas grande del pais, no me mandan factura no me mandan sms por vencimiento, son lo peor de lo peor.
-0.0515-Que mal andas @MovistarArg , merecemos mejor señal y servicio...
1.8914-¡El mejor juego del verano de 2014! Sin descargas, juega directamente en tu navegador
2.182-No resultaba tan difícil ser un éxito, pero resultaba difícil mantener la apuesta. Sin embargo, la dupla conformada por los imbéciles de Channing Tatum y Jonah Hill lo han logrado por segunda vez: una gran comedia, que no tiene nada que envidiar a 'Mil

Maneras de Morir en el Oeste' y que será otra patada a los cínicos, además de ser una gran victoria que quiebra el fetiche de los 'intocables' años 80.
-1.0574-Un jueves flojo fue el 24 de abril para los cines. Apenas 54.000 entradas reporta Ultracine. Es una cifra baja para un jueves con un estreno importante como es el de El sorprendente hombre araña 2.
2.0857-A mi me parece un excelente móvil. La batería dura aproximadamente un día con mucho uso.
0.5545-Un móvil estupendo en todos los sentidos. La única pega y preocupación,es que a veces se sobrecalienta. Espero que se solucione
0.4388-Me encanta es super fluido una cámara impresionante lo único es la radio que no tiene por lo demás es estupendo
0.6941-Hermosa foto
-0.055-Speedy, siempre tenia problemas y se cae el servicio, ademas q te obliga a tener una linea telefonica.
-0.9863-Arnet fue lo peor que me pasó en la vida, el soporte asqueroso, el servicio pésimo
1.209-Tuve Fibertel y para mi es el mejor proveedor de Internet que hay en Argentina comparado con los otros dos.
1.1223-Muy amables en la atención , y siempre respondieron mis consultas a la brevedad
0.7537-Se demora unos dias el envio por temas burocraticos, igualmente llego en tiempo y forma desde la comunicacion de que se realizo el envio. Muchas gracias.
2.7334-Lo he recibido en perfectas condiciones y en tiempo dicho.gracias!
4.6625-Fui muy bien atendida, me explicaron sobre el producto,todo en excelente condiciones,gracias,saludos
-0.4716-Todo el día sin luz. #edesur me reintegra el dinero perdido x no poder trabajar?
-1.0803-Sin luz en Monte Grande, zona de Av. Fair. #Edesur ponete las pilas!
0.4836-Oficialmente soy fan de Benedict Cumberbatch porque su actuación es genial, y porque no gano al publico por "cara bonita"#aplausos #Sherlock
0.5774-Hoy, hace cuatro años que se estrenó el primer capítulo de #Sherlock. Una de las mejores series que vi nunca. pic.twitter.com/SkA1AuQSMj
-4.5419-El terrible remake de la original de 1977 y peor adaptación de la novela de H.G.Wells, es una mezcla tan desagradable de géneros y clichés que ni siquiera la presencia de Val Kilmer, Marlon Brando o el pequeño Nelson de la Rosa, que sí, aparece en la película, podrían solucionar.

0.9014-Este es el mejor celular del mundo con mejor linterna la verdad q todavia lo tengo hace 9 años que lo uso sin ningun problema obviamente tengo el samsung galaxy trend pero mi nokia va conmigo a todas parte lo amo nose q haria sin el. Es lo mejor! Saludos desde Argentina.
2.3491-Tiene buena señal, buen audio y es muy durable. El teclado es de un gran diseño. Ideal para el que necesita estar comunicado. El tiempo lo convertirá en un clásico.
3.8944-jaja el mejor celular, tiene todo el aguante. se supone qe un celular sirve para llamar y mandar mensaje y este lo hace!. El qe qiere escuchar musica qe se compre un mp3!. Encima no te lo roban ni en pedo
10.2902-Este celular no te da complicaciones para nada, ya que carece de pelotudeses como el mp3 o Internet, siendo facilísimo de usar y especial para viejos o gente que no se lleva bien con la tecnología. Pero yo que uso mucho esas pelotudeses en otros celulares, detesto este celular, pero tengo que reconocer que la linterna es de mucha utilidad y la deberían traer más celulares. A mi que se y me gusta tocar el piano, me gusta el compositor
-11.2653-A principio de diciembre de 2013 un temporal fuerte corto mi cable de telefonía que "ustedes suministran" obviamente RECLAME ESTE ECHO, y lo ice mas de una ves... como pasaron los meses Y NO REPARÓN LA LINEA, no volví a pagar ¡PORQUE NO ES JUSTO PAGAR SIN NO ME BRINDAN EL SERVICIO! luego de unos meses me acerque a Telefónica ubicada en Arieta 3660, alli me dijeron que para restaurar el servicio PRIMERO PAGUE LOS MESES ADEUDADOS (meses que no use porque no me arreglaron la linea) En fin pague, pero para recuperar la linea tuve que anclarme al teléfono y llenarlo de reclamos. PERO DE NADA SIRVIÓ. Los días pasaron, y para mi suerte vi pasar un vehículo de la empresa por la puerta de mi casa y lo pare, le explique y le pedí que por favor lo arreglara, y así fue como volvieron a poner el cable que corto la tormenta. PERO SOLO RECUPERE INTERNET porque el teléfono no tenia tono. reclame hasta que me canse y no volví a pagar PORQUE NO ES JUSTO PAGAR POR UN SERVICIO QUE NO TENGO..! Ahora el pasado 15 de Julio me llego una "notificación de pago" la cual exigía el pago en 48hs o inician acciones judicial. por este echo fui a pagar el monto exigido en la notificación, la cual cuenta con 3 códigos de factura. Bueno ahora fui a pagar y me dicen que debo MAS DE 2.000 PESOS. Por este echo, pague solo 2 facturas. yo quiero una SOLUCIÓN CLARA. NO QUIERO PAGAR POR UN SERVICIO QUE NO TUVE, Y NO TENGO
9.6247-La verdad que más haya de estar disconforme con sus plazos de resolución debo informar y agradecer que fue soluciónado el problema en el día de hoy!muchas graciasy

seguiré disfrutando de su servicio ahora que lo tengo nuevamente!
-1.12-Las instalaciones están deterioradas y empeoran su estado desde hace más de 10 años
-1.7541-18 horas sin luz. No hacen 40°, nadie usa el aire acondicionado, no hubo inundación, ni rayos y tampoco una lluvia torrencial. Seguro que esta vez fue por un terremoto en Haití y/o la erupción de un volcán en la loma del traste. LCDTMEdesur.
-4.6527-Gracias a Dios no formo parte de la gilada que bardea todo lo que tenga que ver con Inglaterra porque a un grupo de ellos se le ocurrió invadir Malvinas, una argentinada boluda más, repetir "el que no salta es un inglés", etc.. Encima muchos son los 1ros en ser fanáticos de cosas de ese país, fútbol, bandas, etc.. Además de que como país nos dejan chiquititos, cuando van a allá argentinos se los recibe de 10, etc.., boludeces de esta sociedad vándala y boluda tan triste que tenemos
-0.6745-Nuestro mundo va cada vez de mal en peor
-0.5186-La guerra contra las drogas ha fracasado
-4.2857-Se viaja mal, tambien tiene que ver con la cantidad de inversion que se hace en transporte publico. En mi opinion se viaja mejor que hace unos años, pero sigue siendo deficiente.
0.9217-Mas alla de las cosas negativas que tiene la universidad, como profesores poco capacitados o desganados y todas las cosas relacionadas con tramites burocráticos. Siempre respondieron de buena manera, la verdad no puedo quejarme de mi facultad. Todas las universidades tienen sus pro y sus contra.
3.7748-Pese a no compartir la forma de hacer de hacer politica y algunas de las medidas que tomo. Creo que hizo grandes cambios que si perduran en el tiempo con los siguientes mandatarios, en un futuro se van a ver como positivos. En un balance general creo que diria que mi opinion es positiva hacia ella, pese a las diferencias ideologicas o mejor dicho relacionadas con los medios y procesos para llegar al resultado que espera.
-0.5066-Mediocres y chantas.
3.5756-Calidad de educación elevada. Infraestructura (en Ingenería al menos) buena y mejorando. Podría ser mejor.
-2.0568-Propaganda política muy molesta.
-0.2231-El servicio es un desastre sobretodo cuando se generan altas demandas como en el verano
-0.6175-Aunque en muchos sectores se ven mejoras todavia no son brindan un buen servicio
-1.3209-el estado de algunas instituciones educativas es desastrosa y si le sumamos

algunos dinosaurios que todavia hay dando clases cambiaria a negativa.
1.0484- Para acotarlo, creo que la institución funciona bastanta bien en fiuba para el escaso presupuesto disponible. La voluntad y calidad de los docentes, en su mayoría, es la esperada para la carrera.
7.2411-Es polémica! Pero creo que en promedio estoy satisfecha con su gestión, tiene cosas buenas y malas como todo, pero creo que se ha avanzado positivamente.
-0.6818-Inconvenientes en reiteradas ocasiones.
-0.0947-Siempre repleto de gente.
-0.3723- El servicio es pesimo
-0.6896-Se viaja muy mal
0.975-Excelente universidad la uba
-0.3164-No estoy de acuerdo con sus politicas
0.2801-Tengo Personal, y no suelo tener problemas.
-3.7661-Se viaja pesimo en horas pico. El transporte publico esta siempre muy sucio y no tienen buen mantenimiento. Sin mencionar que siguen subiendo los boletos y no hay muchos camb
-0.4526-No me gusta
-1.9506-La señal no es buena, se corta todos los días. El servicio de internet móvil funciona poco y de manera intermitente, además que su velocidad es cada vez más lenta.
-3.3428-Ella, junto con su gobierno y partido político que está al mando del país desde hace 11 años, tuvieron tiempo para revertir situaciones principalmente económicas que siguen perjudicando a la mayor parte del pueblo Argentino. Hubo mejoras en varios aspectos sociales pero los considero "parches" para aparentar una evolución general en toda la Nación. Considero que cualquier político que esté al frente tanto de una Municipalidad, como Localidad, Pueblo, Ciudad, Provincia o País es responsable de la calidad de vida de la gente que lo habita, por lo tanto mientras haya gente que no tenga trabajo, educación y ni goce buena salud, ellos son los culpables por permitirlo. Especialmente en una Nación con la Argentina en la cual los alimentos, el territorio y los recursos para satisfacer las necesidades de las personas sobran pero son mal administrados no sólo en la actualidad, sino desde hace décadas.
-0.2206-Ante cualquier inconveniente, suelen poner su típico disco automático y no atienden. Creo que tienen el nivel saturado de demandas cada vez que sucede algún apagón o corte de luz.
4.1012-Creo que si bien muchas veces viajamos apretados, y esto genera malestar.

Dentro de todo lo negativo, podría ser peor, es por eso que lo tildo como positivo. Lo que generó el Gob. de la ciudad estuvo muy bueno, el metrobus sí es un bueno medio de transporte.
-0.4441-Considerando que debería ser un servicio que anda siempre y no falla, tiene bastantes problemas en el verano con los cortes de luz. ni hablar de los diluvios que hubo y en consecuencia los días sin luz de la gente
-1.3261-Todas las empresas de telefonía son malísimas. La factura me aumenta sin explicación alguna.
5.7639- Es muy variado este punto. No creo que pueda poner en la misma línea a todas las facultades, pero en general la educación en Argentina sigue siendo, por suerte, muy buena.
-0.5286-No solucionan los reclamos en tiempo y forma.
-0.3576-Muchos cortes de luz, incluso en epocas de frio
0.992-UTN: Estudié en ella y tiene un nivel excelente.
0.6184-No tuve inconvenientes con Edenor hasta la fecha y cuando se corto la luz un 31/12 colocaron un generador a las 2 hs. del comienzo de corte.
-2.1025-Movistar negativo tuve que dar de baja una linea de celular porque no les entraba a tiempo el debito en la tarjeta de crédito y me cortaban la linea,sin tener nada que ver yo como usuario. Ya que era un tema de la compañía del teléfono con el banco de la tarjeta
0.1624-Se destaca la excelencia de la enseñanza a pesar de los conflictos docentes
0.1217-No he tenido grandes problemas con Edenor.
1.2583-Utilizar la SUBE para viajar ha sido un gran avance.
1.3095-Muy contento con la educacíón de la UTN.
-0.4904-stoy viviendo en un edificio eléctrico, y hace meses que estamos esperando que nos coloquen la trifasica. siempre tienen un problema por todo.
-0.2998-Nunca tengo buena señal en ningún lado.
-0.5751-No hay la cantidad de unidades que se necesita en las horas pico. Mala frecuencia.
1.2953-Me parece que cualquier estudio universitario es bueno, sea en calquier universidad.
0.7338-es muy util, la gente no colabora en su cuidado.
0.0935-de estos establecimientos salen los profesionales que el pais necesita.
-0.6052-Siempre hay algún problema con el 3G

-1.2302-En hora pico se viaja muy mal.
-0.6118-Durante corte de luz no se podia acceder a un operador humano
-1.7051-Pesima respuesta a reclamos por servicio deficiente y facturaciones incorrectas
-0.2236-No hace mucho mas que agregar, servicios pesimos, infraestructura en pesimo estado
-0.8619-Solo observar los indicies de pobreza/economia/etc hacen responder a esta pregunta por si misma.

Clasificados mal:

1.1303-Las llamadas se cortan hay zona sin señal y solo el 35%. De las antenas permiten #3G @personalAR @movistararg @ClaroArgentina
1.0435-hoy en día tengo Arnet, y para que te des cuenta, alguien levanta el tubo y se corta la red por ratos (como si estuviéramos en la época del Dialup) mis reclamos no se a donde van a parar.
1.7687-El Nesquik solo puedo tenerlo en mi retina puesto que era mi preferido de pequeño. Pero tiene un sabor como si estuviese muy diluido, poco potente. Y el tacto que comentaba antes es un poco más incomodo.
-0.7808-Cómodas y fácil de comer
0.0976-Viajo por todo el mundo y definitivamente Aerolíneas Argentinas es la peor aerolínea.
0.4514- es mas caro que otros nescafes que puedas tomar de otras marcas.
1.1303-Las llamadas se cortan hay zona sin señal y solo el 35%. De las antenas permiten #3G @personalAR @movistararg @ClaroArgentina
0.4466-El servicio de @MovistarArg sigue siendo tan vomitivo como siempre. Eso sí, para aumentarte la tarifa, ahí no les tiembla el pulso. Pésimo.
0.5762-SIGO RE CALIENTE CON @MovistarArg, no me olvidé que los odio eh.
0.3627-@MovistarArg y x el pésimo servicio q prestan, Haganme los descuentos x TODAS los reclamos del mes pasado pic.twitter.com/Q79mGWilqj
0.428-Por favor, no hay nadie que conteste mi reclamo. @MovistarArg
0.907-'Ride Along' fue un éxito de taquilla en los Estados Unidos, posicionándose cómodamente incluso durante la segunda y tercera semana de estreno (donde suelen desmoronarse todas las producciones luego de la semana inicial); pero la crítica la devastó en su mayoría. Y esta vez, todo indica que no se han equivocado.
-0.1088-Fibertel no me puedo quejar funciona bárbaro
-1.2533-Muy conforme, recibido todo ok !
1.311-Varios días con problemas de luz,hoy llaman de #Edesur para decir q' hasta el 30 a las 00 hs no solucionan el inconveniente #TeOdioEdesur
0.7198-La tecnología ha avanzado y no cuenta con otras prestaciones como mp3, cámara, web, etc. El tiempo pasa...
0.1139-Todo el resto, salvo la batería son contras. Es monofónico, no tiene color, no tiene internet ni mms, y muchas, muchas más

0.2644-Yo estoy sin linea hace mas de un mes y ahora me dicen q tengo q esperar hasta el dos de agosto para q me lo solucionen....
0.1932-mi inteernet desde la semana pasada a empeorado ha subido demasiado el ping por q???
1.448-A ver gente de telefónica, estoy adherido al sistema de factura sin papel y pago cuando me llega el mail indicandome cuanto debo abonar, uno de los problemas que surgen, es cuando deseo ver la factura completa e ingreso a la página de telefónica y no me permite ver mis datos por que no aparezco como usuario registrado, deseo registrarme y no me permite por que me dice que el usuario ya está registrado. Me vuelven loco, así que no te puedo dar los datos de la factura por que no la puedo ver por culpa de ustedes
-0.3571-Sin lugar a duda la tecnología en nuestros días ha alcanzado un alto nivel de desarrollo
-0.4467-La tecnología ha acercado fronteras, abierto posibilidades y unido culturas
0.6344-La empresa gana dinero por servicios que nunca brindó
0.3195-Siempre tardan en atenderte. Varias veces me cobraron cosas de mas de servicios que nunca pedí. No te avisan cuando te cambian de plan.
0.219-Pésimo aunque mejorando.
0.7769-Solo les interesa vender el servicio no que funcione
0.427-En mi casa anterior la luz se cortaba muy seguido. Hice varios reclamos e inicié un pedido de cambio de fase pero nunca tuve respuesta.
0.401-Nunca es clara la prestación/es que te dan con el plan dado los packs que te agregan. La última vez tenía números gratuitos que en el consumo me cobraron. Llamé y me dijeron que eso no estaba vigente aunque yo en la web en el detalle de mi usuario lo tenía. Pedí que me vuelvan a enviar la factura por correo y oh sorpresa es un beneficii inckuido. Una chantada total.
0.3446- Se encuentra desbordado para la cantidad de usuarios que lo consumen. El estado de alguno de los vehículos/vagones en algunos casos no es seguro.
0.0123-Nunca responden ante los reclamos
-0.1926-nunca tuve problemas, pese a los cortes ...gracias a Dios! nunca fueron de mucho tiempo.
0.7348-Cortes del suministro frecuentemente durante todo el año, no sólo en épocas de altas temperaturas en las que el consumo es excesivo.

1.2362-Deberían haber más colectivos para que mejore la frecuencia, lo mismo con los subtes. En este segundo caso es importante que se extienda considerablemente la red por su escaso alcance en la ciudad.
1.6162-Curso en la FADU, perteneciente a la UBA. Ultimamente el nivel académico bajó con respecto a años anteriores. Además en la parte administrativa le falta mucho trabajo para acondicionarse a los tiempos actuales. Falta mantenimiento edilicio, mejor oferta horaria y sedes mejor ubicadas en la ciudad.
0.7049-Es neutra porque no se puede decir que funcionan como debería ser. En muchas partes de la capital no tengo señal, sería necesaria la instalanción de más antenas para ampliar la señal.
1.4067-Su rol es díficil a nivel humano, tiene que combatir muchas exigencias de todos lados y negociar lo mejor posible muchas cosas. Igualmente ese es el rol, por definición ella optó por ese mando y gracias a la voluntad de unos cuantos fue electa. No creo que su desempeño haya sido el mejor, se maneja con un sinismo incríble y hasta limita con la psicosis su visión de la realidad, pareciera vivir en otro país en vez de Argentina cada vez que sale en cadena nacional.
-0.8051-Viajo en varios medios de transporte público y se vieja de forma incomoda ya que tienen poca frecuencia, tanto colectivos, trenes o subtes. Si la frecuencia se cumpliría todo sería muy distinto.
-0.2235-Nunca tuve probemas con Claro.
0.2727-Es la representante máxima de un gobierno de mefiosos y corruptos.
1.0272-No veo la hora de que termine su mandato. No solo me parece una mala persona, sino que realmente no creo que quiera ayudar a las clases pobres, como tanto ella dice.
0.2064-las companias de celulares estan saturadas y la calidad del servicio baja.
-0.1601-En todas es posible acercarse a diferentes disciplinas de estudio
0.6791-Siempre buscan el beneficio de la corporacion, donde en otros paises regalan los equipos, en argentina experimentamos estafas

14.2 ANEXO 2 – CONJUNTOS DE TEXTOS OBTENIDOS DE FORMA AUTOMATICA

Este anexo contiene los textos extraídos de forma automática junto con el puntaje calculado por la implementación, discriminando los que pertenecen al conjunto de textos considerados opiniones y los que son considerados indefinidos. Además se encuentran los textos descartados.

Textos considerados opiniones clasificados bien:

-0.625-#fibertel te odio con todo mi corazón.
-0.1315-rt @marianogoro: desde el 1 de julio, fibertel aumentara 18% los servicios de internet, o sea se va a 400 mangos solo la conexion. ¿compart...
1.7386-@fibertel perfecto !
-3.3864-son todas iguales!!!: telecentro, fibertel y ahora speedy una cagada
0.4588-@flxpiacentini yo también tenía esa velocidad hasta el viernes ahora coloqué una de 25 mgs con fibertel
0.0397-gran servicio negro @fibertel
0.8035-hermoso tener fibertel
-0.9997-@emapiumetti @fibertel yo por 6 y lo hago varias veces al día, es una cagada!
-5.3491-@fibertel cada día peor el servicio! ni al lado del modem me llega la señal al celular! un espanto!
-1.6806-@cablefibertel @rcachanosky @rcachanosky fibertel , que tal si en vez de perder el tiempo contestando lo inviertes en solucionar el problema
-1.0567-sobre la angustia de interactuar bajo presiòn y de cómo no funciona hbo on demand en mi habitación pero no quiero llamar a fibertel.
-0.0114-@fibertel y de nuevo con los problemas para conectarse con @blizzardcs_es
-0.6398-@nico14eg me mata esto. siento que estoy hablando con un operador de fibertel o movistar
-0.0627-@luculpable el otro día pasé por tu cuadra y desde el suelo salía humo de losNegativa
-1.8807-@cablefibertel @fibertel tan desastres es el servicio q tengo q apagar wifi del celular para q funcione. uno paga para no usarlo. desastre.
-0.625-rt @tomasvankooten: #fibertel te odio con todo mi corazón.

-0.6295-que internet del orto te odio @fibertel
-0.3245-terminar de estar haciendo las cajas de la mudanza y que fibertel te patee de battle net es para suicidarse
1.1868-@victorvaldivias en ciudad bien, pasas el ferrilo y horrible. quédese en movistar no más. yo ningún problema en movistar
-1.0398-una suerte me voy a comprar el celu había un movistar cerrado por duelo el otro no tenia el lg g3
-0.1971-@movistarchile chicos de movistar tengo un problema, quiero comprar un chip en usa y me dicen que mi iphone está bloqueado, me pueden ayudar
0.4608-@eltreceoficial apagalo @barvelez o cambia el numero. si tens claro es gratis! calculo q en personal y movistar tmb. t tiro un dato, d onda
-0.5275-estúpido y no sensual movistar que no lee la sim.
-0.2121-@movistar arreglen la mamagueva señal mamaguevos.
-1.0231-me anda como el pico el wifi movistar culiao
-5.1027-pero k pasa tan lento esta el net, tan penca la señal de movistar niun brillo
-0.7572-rt @aarenasmunoz: avisenle a @movistar que el #turndownforwhat paso de moda hace como 3 meses ya #vamoschilectm 10 gb por mes
-0.8181-llamé a movistar para cambiarme a un plan más bajo y me dijo sisi, cuando corto me quede sin servicio
0.0714-@cyynvega_ un motorola amiga, no sabes lo que esta ese celu
-0.5751-@robordon7 el modelo del moto g3 es horrible, si es motorola me compro un moto x ooohhhh noseeeeeee la putamadre
0.0686-pasó de un motorola año 2006 a un smartphone 2016 y yo sigo sin creerlo
-0.6-para la próxima adquisición me compró un htc o un nexus. pero motorola ya no
-0.3382-rt @etchartpali: le encontre un nuevo sinónimo a poronga y a cagada: motorola
0.2371-por eso amo los motorola
0.8409-@marty_batiato hasta mi ex motorola xt914 era mejor que nexus 7 desde que actualizaron a 5.1
0.4098-los cargadores motorola son de un metro,literal. creo que me salvan la existencia

0.5265-@dahorseone io tengo el motorola r no me kejo
1.4175-@crakermc yo te recomendaria motorola, los sony de gama alta y los nexus,tambien los xiaomi suelen salir buenos en rendimiento
0.4598-somos del mundo de las radiocomunicaciones, por eso icom, kenwood y motorola son las marcas que traemos para ti
-0.6162-@sonyxperiamx como extraño mi xperia tuve el error de comprar un motorola y me arrepiento espero pronto tener un xperia de nuevo
1.8067-rt @c_palmere: ojalá mi corazón fuera como el nuevo motorola, para que nunca me lo rompan.
-0.3382-le encontre un nuevo sinónimo a poronga y a cagada: motorola
-0.6443-@motorola g2 es el peor celular del mundo, lento y totalmente detestable. nunca vuelvo a comprar alguno de sus productos. #motorolasucks
0.5042-buenas noticias para los que aún tienen motorola c115. se les metió el snapblack. xd
0.7915-@doblasftw cualquier motorola, se te puede caer mil veces y sigue bien, y la batería es re duradera
-0.3796-@yazawasbf @taetaesgf pero motorola es una caca y no me deja tener lockscreen y homescreen diferentes u.u
-0.5003-no termine de desbloquear el motorola que se me tildó por todas las notificaciones que no le llegaron desde septiembre
-2.9973-@luis_uf yo he usado últimamente motorola. los quiero pero sé que ya no trabajan con claro.
3.3702-muy bonita y todo la #camiseta blanca de #lacopaamerica centenario.. muy cuca !!pero #colombia es #amarilla @adidas #quevuelvalaamarilla
1.053-rt @caarovaallejo: la moda de ahora es ser xoni pero combinarlo con nike y adidas para que no quede tan cateto
0.625-amor eterno a las zapatillas nike y adidas
0.7415-a mí si me gustan estos diseños nuevos de las camisetas adidas en selecciones.
0.5733-uff que adidas mas bellas
0.4272-rt @valeibanez1: que cosa perfecta que es la ropa de adidas
0.7765-que hermoso son los nuevos palos adidas y ni hablar los botines .
0.9303-lindo que adidas tenga para colombia la gran idea de la nueva camiseta en blanco, azul y rojo. cómo todos los hijueputas equipos del grupo.

-0.674-rt @s4ntiagoc: mi sentimiento por la selección ha bajado... y aparte adidas nos pone a jugar con uniformes que no identifican de ninguna m...
0.3318-cada vez que veo a la @fcfseleccioncol me acuerdo de ese cm de adidas.
1.0666-el diseño de la camiseta adidas de colombia es lo que estará en todos lados esta temporada, variando los colores obviamente.
0.8182-2 pares más de adidas y soy feliz
-0.0173-@contrerashector jajaja no parece patrocinio de adidas sino de lechona tolimense el gordo manuel jajaja
0.5653-confirmo día a día que adidas es la mejor marca de ropa deportiva.0.1364
1.053-rt @caarovaallejo: la moda de ahora es ser xoni pero combinarlo con nike y adidas para que no quede tan cateto
0.625-amor eterno a las zapatillas nike y adidas
1.3032-amo la calsa nike porque es re calentita
0.5994-hermosas las nike air max que me voy a comprar
1.4964-me encanta todo lo que sea nike, todo es hermoso
0.1395-curry ha dicho que ve videos de lionel messi para motivarse. seguro anoche vio uno de lio... pero con argentina... #noaparece
-0.1578-el clipconverter descarga con menos velocidad que messi en la final del mundial.
0.4521-es como comparar a messi con alexis sanchez
1.0718-rt @miguegranados: lo hermosa que le queda la barbita a messi
1.6578-@francozidane @faitelson_espn se supone que argentina tiene los mejores jugadores y a messi. argentina debe ganar fácilmente.
0.103-rt @guidobucci12: ahora todos quieren a messi en cancha cuando supuestente en la final del mundial y la copa américa pasada no corrió ni hi...
0.1771-mauricio me hace acordar a messi y es un bebé asi que nunca podria caerme mal
0.6221-tengo miles de fotos de messi y neymar creo que es amor
0.7034-@kattyfrancis está werto loco los quiere a todos hasta a messi
2.1364-me tocan a messi o roman, y me vuelvo loco!!
0.8145-messi cada día está un poquito más bueno
0.2329-se me termina el polo y estoy entre un kaiak o un lacoste

-0.7461-racias al trauma de que hayan descontinuado la lacoste gris, me he comprado tres botellas de la polo double black. ahora: la pobreza.
0.3852-quiero aprender a doblar remeras como en la propaganda de lacoste.
0.4808-hoy me compré una playera lacoste y se quedó bien dormida. □□□
0.7933-@prisofpersia mi mamá pagó chirolas un vestido genial de lacoste porque no tenía cocodrilo salvo en la etiqueta.
0.9501-no puedo entre el tipo que hace pesas con una mano mientras contesta el cel con la otra y el tipetín que entrena con su polo de lacoste
0.0102-rt @textrovertidark: quiero aprender a doblar remeras como en la propaganda de lacoste.
0.5129-coño directtv por eso es que te amo, pasas a esta hora películas buenas de terror.
-0.0018-maldito directtv y su poderío apara conseguir derechos de transmisión
0.101-dios quiero directtv o que se quede el @eltankearias @ad_americatv @gonzaperucarajo vargas noooooo
0.1017-rapidito dieron de baja cablevision y compraron la antena de directtv, claro.. prioridades.
0.2765-@faticollazos tienes razón *cambiandome a directtv*
-0.9715-@alewp8 igual agarre la transmisión de directtv ni se cual es mejor jajaja
2.5569-ihoy por la noche vamos a disfrutar del show de @ramazzottieros en el directtv arena con todos los #superfans!
0.2909-@fernandezjosec en directtv pasan todo. nuestros canales no. atv sólo pasa un partido al día, usualmente a la 1:30
1.0292-estaría lindo tener directtv para poder ver la eurocopa
-1.6484-@javierpiatigor1 @poleroana #directtv $600 sin programación #fox. #personal mas internet $570. #luz $173 barato, café tostadas juguito $75
-0.2614-@hijadelpapa @ocasito ese está en canal 8. suelte el directtv.
26.2659-@lurdesmoreno1 feliz cumple!!!!! espero que la pases bien y que te regalen muchas pepsi
0.125-aguante la pepsi
-7.0E-4-rt @marce_lp: señores de @pepsicolombia, igual en barranquilla y en la costa, no se toma pepsi, así que háganle... #notomopepsi #juniortupa...
0.3398-@demasiadonada la sagrada pepsi compartida con lospi jaja los recuerdo
1.6517-aguante la pepsi batida sin gas y muy fria

-1.0344-@darkvader2015 @sin1presidente peor! prefiero tomar agua en lugar de pepsi o big cola
1.3577-seagusta la pepsi me gusta la cocn me gusda besarte en la boca
0.48-@luizasaantos_ prefiro pepsi
0.5956-rt @sin1presidente: @eljorgebsc las ganancias son muy bajas los supermercados prefieren vender pepsi que no tiene la misma salida pero más...
-5.5252-esa pepsi que está en la cocina me puede mucho pero no se si subo vivo o endemoñado
-1.0399-rt reporte_futbol: opinión sobre messi de anellogaby rt anellogaby: en vez de sangre le corre pepsi por las venas.... mucha publicidad poco...
0.0682-@castell_melu @isaalvarez_20 @la_alecastro @leydimuralles la primera fui yo con la pepsi
1.2765-un licor de menta con sprite con el amigo natha total es lunes recien
0.625-rt @camidening1: necesito estar saboreando un vino con sprite en este preciso momento
1.7138-a tener problemas, en ese momento de debilidad, salió el amigo sprite aprovechó la ocasión para salir d la friendzone e hizo bien su trabajo
0.7501-mi mama diciendo que va a comprar un pack de botellitas de sprite para ella y mi novio, siempre lo mima jajajaja
0.2908-que cosa rica la sprite y el ades
-0.6317-llegó ruben con una bolsa de gomitas y sprite , como hacer dieta con este hombre
0.3832-que piola fue cuando nos fuimos a comprar sprite con nacho y thiaku
0.0262-mi vida con los ratatoing es un comercial de sprite.
0.2637-que lindo encontar sprite en la heladera y saber que tenes un gancia en el mueble que te llama
0.7631-necesito una sprite d litro y mucho hielo
-5.9144-hijo no toma refrescos en general pero ayer nos veía con los vasitos y quería de lo mismo. le pusimos agua y dos gotas de sprite en uno.
-1.0257-odio que me guste tanto la sprite y no poder tomarla porque me hace mal.
0.6142-que ganas de una sprite o una pepsi biennnn fria
0.6757-siempre me agarra sed a esta hora toda la paja levantarme, necesito una manguera de sprite

0.5598-#daysgone #ft #gollum xd! o por lo menos era el mismo sprite #e32016 #sonye3
-1.2279-dos años de cárcel a los pendejos que, siendo de méxico, se van a echar un café a un starbucks... pinches mamones.havuck, presidente 2018
0.1507-agustina cobró y metimos mc, starbucks, dot, helado y neverlannnd
3.5832-@starbucks_cr hola, está demasiado rico. súper recomendado! :p
0.757-tramitame un burger king y un starbucks en san juan
0.2917-alguien para aprovechar el happy hour en starbucks. □
0.2663-la crema batida de starbucks es por lo que vivo
2.1891-rt @ccr_02: me urge un starbucks está vida sin lácteos es muy dura
-0.5013-rt @minfame: la verdad es que a mi sí me da pena llegar a un starbucks y pedir un café de sumatra en prensa francesa.
-0.8223-como la vez que, con lana y carl's jr. cerca, gasté lo equivalente a una hamburguesa "comiendo" en starbucks y morí de pedos toda la noche.
0.3684-rt @irvinlozano_: quisiera que hubiera starbucks y dq a domicilio□.
0.0988-@caleebmiranda cz:te amo a mas no poder,te extraño,necesito tardes de starbucks
0.1234-vamos a ver quien se prende para un starbucks mañana☕
1.7753-@_josh19 ahhh xd jajajaja es que starbucks saco dos nuevos sabores de frapucchino y uno de eso es de churro y sabe bien
0.6838-le dije a mi mama que me comprara un cafe y me compro la tienda completa de starbucks #queladillaserburgués
0.9531-me acabo de comprar unas gafas bien hipsters, ahora si puedo ir al starbucks con dignidad.
5.3289-el té de hierba buena de starbucks es lo más rico y relajante de éste planeta!
0.7783-amigos de @movistar tengo un plan smartphone y bam en @entel_ayuda si me ofrecen alguna promoción buena me cambio
-5.3516-#doctorfilemv el #noaa pronóstico tormenta solares, eso explica los problemas en @movistar @vtrchile @entel y señal hd nacional nula!
-0.5546-rt @rgcopernico: @runrunesweb @nelsonbocaranda la pregunta es: y que dice movistar de violación de contrato, porque entregó algo a quien no...
-0.625-movistar no sirve para una mierda

-2.217-#doctorfilemv desde q se reprodujo x primera vez en el programa el sonido infratierra la señal del canal se escucha horrible x movistar,raro
-1.2162-movistar me tiene la vida triste
-0.0947-puta que son malos los controles de los deco movistar
1.1956-que mojigatos me parecen algunos servicios en este pais, para mi lo mejor en bancos: provincial, lo mejor en celular: movistar
-0.4305-no quiero pagar más de 20 lucas y las opciones son bien callampa: 15mb en claro y 8mb en movistar y con límite de 500gb en la última
-0.1908-@randompiece @fmonroy @movistarve nadie me quita la idea de que hubo soborno a algún empleado de movistar. nelson es una celebridad..
-0.6147-jajajajajajaj bue , estoy esperando que sean las 12 así movistar de mierda me da los megas y me dejo de joder
0.5988-movistar bamos vien ...
1.6023-mejor no me quejo de movistar verdad
-0.7232-que ladilla con movistar que no actualiza la hora.
0.1615-no veo la hora q terminen de cablear los de fibertel así sacamos este intenet de mierda
-0.1887-@fibertel no tengo internet desde ayer
-0.3786-todos miran el partido y yo mirando que movistar aumenta un 14%
-1.2159-@sinbustos movistar cagando comerciales desde tiempos inmemoriables!
0.172-he usado mi celular todo el dia y aun tengo el 50% de bateria gracias motorola
-0.5677-@solviegass sii una bronca, casi incendio el local de adidas cuando me dijeron eso jajajajja
1.0495-rt @carlosnewells: tuve sueños mojados con la barba de messi
-0.7533-rt @tinchowest: no me jodan con starbucks, a mi dejame con el puestito callejero que tiene el mejor café a 10pe
0.1693-rt @jazgrieco: necesito en la plata : dean & dennys, starbucks, burger54 y un shopping por favor
-0.2709-los megas de movistar se van mas rapido que un suspiro es un robo
-0.2413-@santimaratea hace más vídeos por favor, me cago de risa en el laburo que está lleno de viejos soretes , bardealos a los putos de movistar.
-0.36490958270292084-para variar no hay internet... y van.... @cablefibertel @fibertel

-0.21652615563419014-rt @jerosahagun: me cago en la vida y en las familias de toda la gente que trabaja en fibertel
-4.377485887215847-@cablefibertel y @fibertel para variar anda lento internet!!! cuando van a solucionar eso??? hace 3 meses venimos con lo mismo.
-0.6335632849508939-rt @pachoriva: es un desastre @fibertel , no funciona *no importa cuando leas esto*
-0.8024147586144605-rt @lucislux7: fibertel otra vez cortando la señal y transmitiendo para el orto. #tresperiodistas
-0.5173933209647494-@juanvalentin86 @cablefibertel @fibertel mandales hasta hartarte, son de 4ta!!
0.29345769627593843-me voy a dormir nomas... necesito tener fibertel o arnet, una mugre todo lo demas!
-0.08819153211744563-no decido si @pedidosya es más mierda que @fibertel o viceversa.
-0.5276286767790545-rt @pedrito_vm: estoy alterado y no se porque... sera porque los de fibertel no me habilitan un buen inet? la re puta que los pario.
-0.7079494648421623-parece que los pelotudos de fibertel se divierten cortándome el wifi cada 2 minutos, porque es literal. váyanse a la mierda, manga de forros
-0.05596847762272504-@fibertel @cablefibertel otra vez no funciona el servicio de internet
0.9202983333120799-@nahu__casla tiene que putearlos mijo, aprenda de mi con fibertel
-0.5788567493112947-fibertel te agradezco por dejarme sin internet 1 semana
0.5987088360610043-@crisgreen95 ch cris si por ahí hay fibertel pos te lo recomiendo
-0.2229520540748442-rt @jonhy_call: @todonoticias @c5n en @fibertel se cagan tanto en los clientes, ke despues d decir ke no funciona me mandan una encuesta!!?
0.23515928782049517-el dia que tenga fibertel voy a estar re plaga en tw insta snap lo que sea..
-2.5334614460253615-lo bueno de fibertel es que funciona tan mal que te hace sentir que tu casa es enorme y no llega el wifi del living a tu cuarto.

Textos considerados opiniones clasificados mal:

0.1111-@fibertel estimados, otra vez tengo problemas con la conexion. se cae. media pila!! se paga x adelantado!!
0.3514-@emapiumetti. @fibertel te hara hacer unas pruebas, reiniciar el.modem y la visita tecnica para al mes volver a lo mismo @cablefibertel
1.9595-@aquimarco no quiero desilusionarte pero lo mismo me pasó con movistar
-0.3526-@fbrcrojas hola , trabajo para movistar y en promoción trío talla s desde $ 32.990 , instalación gratis , deco principal en dvr
-0.6096-chao, aguante el test drive de movistar.
-0.3904-mientras tu me ignoras, movistar me ofrece tiempo aire gratis
-0.2747-gracias a jesús el miércoles me dan el motorola. sino con esto me iba a morir
-0.242-@ezelagorio bueeeno si estoy con la verga esta y al motorola se lo di a la novia de mi hermano jajajajaj
0.4098-@moto_mex exorto a las personas que no adquieran equipos motorola si no quieren ser decepcionados
-13.2081-rt @fiamasalamanca3: quiero ir a la casa adidas y comprarme todas las zapatillas!!!!!
-0.0571-quiero unas adidas y me las voy a comprar
-0.1148-quiero unos nike:(
0.0078-tengo las votitas vanss que no las uso por que están sucias y las championes nike los use 4 veces y como se educaron nuca más los use
-0.0327-rt @camimartinezzok: entras a nike y te volves loca/o
-0.1148-rt @santinchu: quiero un rompeviento nike
-0.0975-quiero unas gorras nike que vi.
-0.1332-quiero una gorra nike
-0.262-rt @alprimertoque: piqué: "es rarísimo que tres estrellas como messi, neymar y luis suárez se lleven tan bien. eso es impagable" #apt
-0.125-preciso de oakley, lacoste, hurley, thugnine, lost...
-0.1148-quiero esos lacoste
-0.1257-@lacoste jejejejeje quiero todaaaaaaaa la coleccion
0.0542-loco estoy re triste, posta... #directtv saco de la guia los canales de musica de un dia para otro sin aviso

-0.1242-@profesorlucasp una pena. hubiera sido lindo ver este nuevo gp europa en vivo, buscaré alguna cuenta de directtv para poder verlo
-0.3929-pensé que era mi parabólica, me acordé que tengo directtv entonces fue error de mtv #malumahostmtvmiaw
-0.1074-rt @jaasoqui: ayer americatv, hoy atv... ya mejor voy ahorrando para contratar directtv para el mundial 2018
0.5676-rt @cantantedrogado: giles son los que sobran, pero mas giles son los que le ponen pepsi al ferne. no me lo cagues asi hermano
0.2654-te juro que no entiendo, cuando estaba en el secundario tomaba 3 litros de pepsi y 10kg de doritos y estaba perfecto
0.3382-rt @miguemartinezf: coca cola + coquan = con pepsi no pegaa #notomomaspepsi @pepsicolombia
-0.1565-quiero una sprite
-0.1543-@chavaluria @s_emaforo es tanto café del starbucks.
-0.3444-quiero un starbucks, quiero nikkori, quiero todo
-0.2652-conttiisoletta sii, tengo ganas de ir a oroño ósea starbucks y pan de queso
-0.1148-quiero un starbucks
-0.053-rt @karym_andrea: quiero un starbucks, alitas y boneless, chips fuego, hershey's, nuggets, una hamburguesa y tacos
-0.5-con ganas de ir a starbucks.
-0.5-rt @tufavorite: con ganas de ir a starbucks.
0.9511-tengo dos opciones para contratar internet en el depto: claro y movistar, puta que te echaré de menos vtr
0.67-rt @leslyramirez_: @fibertel mal servicio sin señal
0.0336-rt @jucanoma: me llaman a cualquier hora para ofrecerme promociones pero para cancelar me indican q el horarios es de l
0.6571-rt @pofirella: movistar dejame de joder y funciona bien un rato por lo menos
0.7098-esta verga no hace ni 2 meses que la tengo y ya la quiero dar contra la pared motorola que te pasoooo
1.0276-porque mierda motorola funciona con google play music ahora la caga me caducó y tengo que pagar, linda la wea

-0.6623-@tomaswagner19 coincido con @cristiannmillo . soy taliban de apple pero en su defecto motorola es magnifico
0.9443-@z4lseo bue piensa que yo tengo un motorola g 3ra generación, no me ha durado ni 8 meses y ya está re malo :'v
-0.875-rt @tatianaabetros: como la rompen las adidas superstar , explotan
0.2891-@tar_zan que adidas hizo una campaña que decía "columbia" en lugar de colombia con fotos del equipo colombiano
0.6508-adidas ni se tomó la molestia de diseñarle una pantaloneta a la camiseta blanca peye esa. espero lo noten
-1.4431-ese uniforme blanco de colombia no tiene ninguna gracia. no nos identifica de ninguna manera, le hace falta sabor a esa vaina @adidas
-2.4457-@agudelabarrera olvidate, con ese frio nos ponemos un camperon adidas, salimos afuera escuchando el pepo y nos cabe una!
-0.2312-@camilatamaraaa y lo que salen, deja. sabes el conjunto nike que me compro en vez de esas camperas chetitas
0.2955-@yeisson724 esperar que pronto jueguen messi para que la pechern xd
-0.3459-@nescalderon11 cuando messi era niño no lo quisieron en argentina y mira lo que es hoy.... a trabajar señor
-0.4312-@alejobostero @peronistsheep fue porque estaba corriendo hacia el arco rival a lo messi, señor.
-1.1312-rt @alextoremember_: vuestras gorras de lacoste mi abuelo las llevaba hace 5 años.
-0.8176-fan de la ropa lacoste, explotada mal por favor
-0.6682-rt @iaurab: compré una polo lacoste y se quedo dormida.
1.2573-#futboltotaldirectv estimados amigos de directtv la ultima vez que perú le ganó a brasil fue en el 75',no el 58' como informaba su reportero
0.9568-#futboltotaldirectv iiivenezuela la euro es por meridiano!!! directtv duplicó su tarifa y no la transmie @giraltpablo @alex_candal
0.0078-@javiermendezm :(aca el servicio de directtv es caro...
0.2059-avisenle a directtv que va 1
-0.2068-#euro2016portlaymtv sus comentarios son una mierda, déjenos ver esa vaina por directtv pajuos
-0.96-fernet con pepsi de donde saliste mogolico ahre
0.4711-cuan mamau tienes que ser pa decir "dame una pepsi"

0.4308-que dios perdone a los que prefieren la pepsi y no la coca cola
0.1044-@sin1presidente @darkvader2015 la pepsi es un asco. prefiero tropical o manzana
1.1544-los más lindos de los equipos son love y villar y son más malos que una pepsi caliente.
-0.0682-así como hay gente que mi puede vivir sin puchos, yo no puedi vivir sin pepsi
-0.4292-quiero con locura el tenedor enrolla spagethi de sprite #borntorfrsh
-0.6402-@matiasbusajm no es rica,tiene un sabor horrible, la sprite es riquisima
-0.3977-tengo miedo de tomar alguna decision que se vea afectada por mi estado de animo y terminarme arrepentiendo o trabajando en un starbucks
-1.5269-quiero conocer seattle e ir a tomar un latte en el local original de starbucks, con el cielo nublado tan característico de ser posible
-1.1445-rt @fedecorriente: un starbucks solucionaría la vida de muchos, café rápido y listo perro te vas
1.3493-saben que me dormí un rato y se fue el internet y obvio movistar no me deja ver todo el tl ayy lmao
0.1535-mi mamá tiene tal vida de mierda que se está peleando con movistar a las 12 de la noche en vez de no se fumar un faso ahr
0.4611-rt @doctoramolina: hoy cumplo 17 días desde que @movistarmx extravió mi línea y @mimovistarmx no soluciona. si pensaban cambiarse a movista...

Conjunto de textos considerados indefinidos clasificados bien:

1.8544-llegar a q movistar funcione mejor que @fibertel @cablefibertel @dndconsumidor uno paga fortuna p/nada. todos los días lo mismo.
-1.0235-rt @josesaa62: @cablefibertel @rcachanosky @rcachanosky fibertel , que tal si en vez de perder el tiempo contestando lo inviertes en soluci...
-0.9504-edesur me dejó sin luz todo el sábado. para no ser menos @telecentroayuda sin cable e internet sábado y domingo. vuelvo a fibertel mañana.
-0.8136-@claroelsalvador @romeolemusam claro al igual que tigo, digicel y movistar son lo peor que le pudo pasar a el salvador en cuento a telefonia
-0.5614-rt @diogara: @aquimarco vos proba e igualmente deja la puerta abierta xq movistar no tiene atención personalizada, ojalá te vaya bien

0.1646-ojalá vivir en el movistar arena
0.1188-@ariel_bn hola , si gusta cotizar en movistar me avisa , salu2
-0.2453-que alguien le diga a movistar que aparte de que el turn down for what ya pasó de moda no saben usarlo.
-0.439-@ayudamovistarcl no se ve el canal 13!! llamé a movistar y no supieron ayudar
0.7154-rt @jeanmary_: qué chévere. ya no tengo datos del plan movistar porque aunque me la paso conectada al wifi, al parecer, también me chupo la...
-0.3503-tengo que ir a movistar a buscar el chip, alto bajón aparte pierdo todos los contactos
-0.6734-@marcepalaciosmp te falta poner gracias por el oxigeno roamong movistar
0.2186-#movistar migue granados no entra en el locker jajaja. te banco gordo
1.2458-@guiastacticases @thegrefgyt en vez de un teaser era una propaganda del nuevo motorola moto g4 :v
-0.2137-@jotacervantes14 te contamos que el equipo motorola moto e dual se no se encuentra disponible en nuestros centros de experiencia tigo
-0.8142-no se si pedir un samsung a5, el j7 o un motorola mmmm q difícil
0.2815-@telcel es liberado de fábrica, es un motorola g3 turbo edition , doble ranura sim , android 6.0
0.7083-@golcaracol gol de colombia #iscolombianotcolumbia @adidasco @adidasus @adidas
0.428-xq será que paraguay siempre tiene último los nuevos modelos de adidas
0.262-#lomasextrañodeldía es que mi mama me haya comprado un balón de fútbol y adidas o.o
0.1932-curioso q ni #nike ni #adidas vistan a los árbitros #copaamerica adidas,,los guayos y nike el balón #voit la ropa
-0.4756-rt @maartipison: me hace mal ver ropa de nike
0.2418-@nike_hernandez @lacopamx gracias
2.12-@nike_hernandez @lacopamx @segundadivfmf esa idea es buena, sólo que a la @fmf solo le interesa el dinero
0.2419-los dioses no están muertos, kid, están paseando por tu barrio y por el mío, cambiando el mundo con sus nike and reeboks

0.2883-me acabo de enterar que en mi reto de nike+run, está el célebre @genarolozano, a quien siempre leo.
0.2462-rt @edgarbaez95: siempre gano con la zapatilla nike
-0.783-yo quería unas nike sb y me tuve que comprar unas dc, todo mal, todo mal. wee agradece que te comprar zapas sin que lo pidas tarada
0.9727-rt @iiskani: vi un compilado de messi y ahora voy a ir a jugar al fútbol e intentar todo lo que hace. por ahora me sale re bien la de vomit...
-0.1286-rt @rinaudoopina: el pulga rodríguez es el messi que en vez de tomar hormonas, tomó vino toda su infancia.
-1.0281-rt @cihs22: @nescalderon11 jaja ni que fueras messi pendejo, eres más malo que el "avión" ramírez de @chivas ino te queremos en el más gran...
-0.7-tu camisa podrá decir "lacoste", pero tu cara dice "quetzalcóatl".#esdenakos
-0.2225-@robertolanzas75 @inakiatm @ripollrafa tiene apolillado el lacoste la gorda esta
-0.8302-tu ropa dice lacoste tu cara dice quetzalcoatl
-0.6315-rt @senoracatolica: tu camisa podrá decir "lacoste", pero tu cara dice "quetzalcóatl". bendiciones.
-1.5986-rt @chembito: a mi no me gustan las retros env mejor me quedo con los lacoste, pumas, polo, huarache, aldo y con eso uno palte con cualquie...
-0.4986-tu novio el que cree que por usar tenis lacoste es millonario.
0.0826-botas lacoste. podes tenerla aprovechando el plan #ahora12 #jueves #viernes #sabado #visa #cabal #mastercard
-0.3155-la embajada de italia debería pagarle a iliana calabró para que no aparezca por ningún lado. como hacía lacoste con los wachiturros.
0.1814-la propaganda de los perfumes lacoste me da ganas de jugar grand theft auto: san andreas
-1.2022-#gameofthrones por duro no tengo directtv y cuando veo futbol por rcn, siempre me arrepiento
-0.5893-#copaamericaporgamatv pura publicidad. veamos el futbol por directtv.
-0.2197-es un robo que solo #directtv tenga la exclusiva con #euro2016 #fútbolempresa
-0.343-rt @chasquetti: acabo de ver en directtv como le hacían bulling a buysan, un periodista venezolano, un argentino y un mexicano. periodismo...

-1.0159-@meridianotv que triste lo q se convirtió este canal tiene derechos d la eurocopa no pasa ningún juego y no deja q lo veamos por directtv ,
-2.1143-rt @luis_sanze: @javierpiatigor1 @poleroana #directtv $600 sin programación #fox. #personal mas internet $570. #luz $173 barato, café tost...
2.6097-@sebasdecker hola sebastián! me llamo juan carlos albuja, te estoy viendo por #directtv soy de ecuador y estudio periodismo en arg. saludos!
0.8295-vacaciones=deportes #tourdesuisse #euro2016 #copa100 #nbafinals genial @bertenschilders #bigscreen #directtv #espn
0.4023-#copaamericaxdirectv saludos desde quito a fabián gallardo y todo el equipo de directtv
0.9164-@jmtellezm la vez pasada participé y si me las gané en pepsi
-0.9031-son pendejos los que beben sprite con benadrex y se sienten negros.
-0.0835-@25maaar yo tome 2 mientras ustedes se desesperaban para que le den sprite para bajarlo, ca go na
0.5152-rt @afatimaaalvarez: gabriel y charon se pagaron la sprite y los carlitos
0.1672-pororo, sprite y mirando pavada en la tele
0.0171-@falsestreet nosotras las hemos tenido con todos revueltos. y jugo de naranja o sprite jajaja
0.0257-encontré una sprite con limón que prepare ayer para la resaca y no la tome toda , me vino como anillo a el dedo
1.1675-@lion_vb97 merca, pepa, faso, pasta base, nafta, forros, sprite, rollo de papel higienico todo
0.9225-"estaba en mc y me acordé cuando estábamos en starbucks que nos mostraron el culo" jajajaj
-0.3016-capital esta llenísimo de lugares fotografiables. por el amor de dios, cortenla con el vasito de starbucks.
-0.8536-hablame de cuando nos dijeron "chicas pusieron un starbucks cerca de la 25 y las enfermas corriendo casi llorando" que echada de humo
-0.7522-ya paren el mame, nadie les va a comprar su pizza, o sus nachos, o un starbucks por una publicación de face. mmmmmecos.
0.213-en el starbucks del abasto, había un chico alto de amplia espalda, pelo negro con anteojos en la caja.después un coreano facha hacia el café
-0.1932-@patolazbal te haces el boludo y no viniste a fisico por ir a starbucks

-0.4318-rt @gimemoreno111: a san juan le hace falta un burger king, y un starbucks
3.576-nueva discreta de 25 años. tetona, estudia comunicación en la uanl. $1000 la hora. muy buena onda la conocí en el starbucks. monterrey, nl.
-0.3519-pésimo servicio de facturación @calufecafe en querétaro. lástima del buen café pésimo servicio. ámonos a starbucks chafa pero sonrien.
1.0015-betos y starbucks con mi novio, bien obesos

Conjunto de textos considerados indefinidos clasificados mal:

3.3118-@fibertel che 3 reclamos en una semana y me dicen q lo "solucionan" pero no me carga ni un audio de whastapp. no da más de lo mal q anda
1.7466-rt @s4db4rbie: yo bien emocionada xq pense q el internet se puso rapido y resulta que no estaba conectada al wifi que peooooooo ya movistar...
-0.4463-@maylen15reyna @matias_munozz hola, les ofrezco movistar , trío hogar desde 32.990 , instalación gratis , deco pppal en dvr .
-1.0141-@movistarmx si podré ir a las tribunas movistar el jueves, me mandaron mensaje
-0.4004-@demiansex hola , si quisiera cotizar para movistar me avisa , salu2
-1.0896-rt @sinbustos: alguien que me explique el comercial de movistar "vay al arco"
-0.9491-colgué en ir a movistar, mañana voy
-0.0367-@patricnor b tardes, las llantas en 2500 las 4, el cel ya se vendió, voy a publicar otro cel en 900 pesos un motorola droid, a sus ordenes
0.0322-vengo del futuro y #adidas experimenta tanto con la camiseta de la selección que la proxima sera violeta y la alterna fucsia.
-3.8151-rt @alejandroolaz: disculpame la cipayeada pero si en algún nike store me cruzo con la camiseta color negro de usa, ni lo dudo. la uso pa'...
-0.5991-ya aprendi a transformar mi camisa lacoste en un perfume
0.4236-@matiaschiriotti @melicavallo no es así en caso que haya un regalo que sea de corazón igual recomiendo #tommy #apple #dior #lacoste #diesel
1.061-@analiafranchin va a usar lacoste cuac!
0.3974-@eleonoranavatta me parece que el resto de los partidos que no van por tnu van por directtv
0.0656-@arielinostroza @manueldtp @telecanal yo tengo directtv y no los

transmiten pero si gustas ver lo puedes hacer lo en rojadirecta
0.6038-@matiasbazaes @fpetrocelli @fabig08 en venezuela no lo podemos ver por directtv solo por el chavista canal meridiano que desgracia.
0.1515-que criminal que esté la eurocopa sólo en directtv.
0.1123-@aguslavigne2 yo lo miraba hasta que a directtv se le ocurrió deshabilitarlo :'v
-7.2153-rt @barbieharp: este 25 de junio el concierto de belinda en el pepsi center!!! no puedo esperar más #cdmx
-0.1765-rt @belindapop: “@ra0lin: @belindapop harás un dvd del tour catarsis #askbelinda”// seguramente vamos a grabar todo el concierto del pepsi...
0.184-@javierlopezdiaz en este momento riña entre dos conductores en el cruce de la pepsi frente a clínica abc dirección entronque
0.8329-hoy me compre un chocolate y una sprite en el kiosco y me salió 30 pe con eso me compro una petaquita amigo kalmateeee
0.0799-@agustinbarbos10 jajaja yo estaba comprando una sprite en pantuflas re crota jajajjaa y vos salias del pasillo ahi te la vi
-0.0417-gancia con sprite con mamá
-0.2372-@falsestreet jajajajajajajaja y después vodka con naranja y tres esquinas con sprite! eso que va tome de todooooo jajaja
-0.3405-@el__venegas @ch14_ hahahaha tomando sprite y unas chips verdes
0.7067-gancia con sprite sin limón no es un mismo el limón le da ese gustito especial
0.1693-necesito en la plata : dean & dennys, starbucks, burger54 y un shopping por favor
-0.0605-@nico_ferrer1 m.a te amo futura cuña jajajajajajaja te extraño, pagate starbucks
-0.1012-holis voy a trabajar en starbucks
0.2815-nadie quiso ir conmigo a starbucks

Textos descartados por ser considerados consultas:

-@fibertel internet sigue andando para el orto y uds no responden, que es lo que pasa con el servicio!? llamo al 0810 y no atienden! no jodan

-@fibertel hace dos horas estoy sin servicio. cuando van a arreglarlo y brindar un servicio acorde a lo que pago?
-@fibertel me desconectaron el wi fi como es ? asi solucionan los problemas ?
-@fibertel ya les escribí por privado y no respnden. deben estar mucho muy ocupados con su maquina de escribir invisible. o no?
-cuando estas paveando, internet anda como los dioses.. pero si tenes que hacer algo de vida o muerte se vuelve una caca. @fibertel ke onda?
-alguien con movistar???
-alguien que me preste un celu movistar x unos dias?
-@attmxayuda pase mi numero de movistar con ustedes pero en el carrier aparece at&t 3g y luego iusacell lte por que aparece esto ?
-mi plan es movistar con equipo huawei, me afeito con guillete,le doy masterdog a mi mascota.
le soy fiel a alexis?
#partychilensisftlaroja
-rt @doctorcsp: ¿por qué nadie sanciona a movistar por esas cagás de comerciales?
-rt @eblaquier: #unaluzeneltunel mientras bla bla bla, movistar y personal subirán los precios hasta 14% a partir de julio. ¿no lo sabía? ok...
-movistar me tuvo 11 horas sin señal. sentí que faltaba un pulmón. como podíamos vivir sin celular? que mierda.
-@androidmx @karlos17tdv un motorola force, si lo tienes? en cuanto me lo vendes?
-@el_leeguiza tenes un motorola no?
-pregunta para entendidos : samsung grand prime o motorola g 4g? pros y contras ? murio mi celu y me ofrecen esos dos
-@movistarplus @yomvi ¿cuándo tendrá yomvi compatibilidad con el motorola nexus 6?
-de motorola que terminal recomiendan amigos? @elcamionerogeek @elcamellogeek saludos desde mexico!
-@ruiz4647 es motorola, qué rayos esperabas de esa mierda?
-@dropex_droiidyt y ke klazhe de motorola tiene?
-@motorola_mx estoy por comprar el moto x style, saben de algun código de promo? msi o alguna promoción adicional ?

-@motorolasoporte perdí mi cargador turbo de mi motorola g turbo. ¿donde lo venden?
-hablando en serio, @nike me parece que es la mejor en tenis, baloncesto, golf (?), pero en fútbol @adidas es más nivel lejos.
-que bonitos trabajos hace adidas con los jerseys de colombia. ¿porqué no nos pueden hacer algo así?.
-a quién hay que hacerle los mandados en adidas para que me den las camisetas de todos los equipos que visten?
-columbia? no en serio? #adidas
-rt @primadeguilsmit: a quién hay que hacerle los mandados en adidas para que me den las camisetas de todos los equipos que visten?
-hablando en serio, @nike me parece que es la mejor en tenis, baloncesto, golf (?), pero en fútbol @adidas es más nivel lejos.
-@valentiorrego_ a cómo tenes el nike?
-@christianleave nike or adidas?
-esuve toda la noche con una campera nike, unas ganas de salir a robar tenia (?)
-@ichibolo -¿y qué opinas de zapata?
+mmmm... me encantan, sobre todo si son nike o adidas
-@romano_434 @marca jajajaj messi con argentina? estas seguro que lo has visto jugar con esa camiseta una azul con lineas blancas o viceversa
-@menino_messi @fiorettiisa_ q mina arrombado?
-rt @deporlider: crees que messi debería jugar ante panamá o seria conveniente resguardarlo para bolivia y los play-off?
-@martinsouto messi no es más fundamental para su equipo que guerrero? □□
-rt @decontragolpe: en el partido contra panamá ¿crees que messi tiene que jugar de titular?
-#mipreguntaes ¿crecerá alguna vez messi?
-rt @gavioto_: -me dijeron que te fuiste a casa con la pivita esa borracha del polo del cocodrilo, ¿eh pillín?
--lacoste.
-vaya maricón está...
-y si me regalan uno de esos feos perfumes de lacoste □?
-@ripollrafa @esrespeto ¿porqué ese hombre siempre lleva un polo negro lacoste,

no tiene más ropa ese vividor de la subvención pública?
-juan se viste en lacoste y andy polo
-@juancaherg esa solo va por directtv?
-buenas tardes,servicio técnico de directtv no tengo señal desde hace una semana
-por favor desconecte la antena,¿de qué color es la antena?
-solo x directtv ???? no hay plan b?
-@futboltotal@directtv, vargas si sacaran con un gol de mano a colombia hablarias igual?
-@directvuy no esta en directtv sports?
-sabran en #directtv que añadieron reglas en para la #copaamerica??? super perdidos!! #brasilvsperu
-@domingatto esa es por directtv ??
-solo directtv transmite la eurocopa?
-alguien sabe como robar cable? no quiero pagar $50 por #directtv lmao
-los partidos de la #eurocopa solo los va a transmitir directtv ???
-qué onda #directtv y sus espectaculares reportajes !!? *-*
-y yo que quería ver la #euro2016 pero directtv compró los derechos, no? damn!!
-alo gente la euro copa la transmite sólo directtv ??? nadie más ??
-@fernangood ¿me llamaste? sorry estaba peleando con los de directtv.
- cita y responde
78. ¿coca cola o pepsi?
-rt @dailyvotingpoll: coca-cola or pepsi?
-rt @yaressicisneros: @patotorrescast ¿domingo de enchiladas y starbucks o que?
-los starbucks me tiene las bolas al plato, los yankees nos van a enseñar a tomar café? little latte dame un café con leche enfermo
-¿cuánta publicidad ven de starbucks en la calle? aparte de sus letrerotes en sus tiendas???
-@zonarojas tuiteás desde starbucks?
-@lethyss me like, me like, a starbucks? ella ama el café, andale si?

Textos descartados por no tener una palabra con puntaje mayor a 0.1:

-que servicio de mierda manga de forros!!! @fibertel
-para cuando vienen a bariloche @fibertel
-anda para el orto el wifiiii. fibertel te odio
-@fibertel que generan ustedes .
-sos una mierda fibertel
-fibertel y laconchadetumadre
-aaaaaaaaaaaaaaaaaa @fibertel los odioooooooooooooooooooooooooooooo
-@cablefibertel che gente no me estaria andando ni fibertel ni cablevision
-fibertel una mierda, cuando más lo necesito no anda
-jajajajajajajaja me siguió fibertel
-malditos tus módems @fibertel
-@fibertel solicito la baja del servicio.
-@fibertel mal servicio sin señal
-fibertel del orto #twittermusicfestival
-@c5n @epoliticac5n @robdnavarro como todos ls dom. la transmisión está cortada para los clientes de fibertel
-la re concha de tu madre fibertel hijo de re mil puta
-fibertel te re contra mil cago odiando
-quién inventará los comerciales de movistar #partychilensisftlaroja
-por culpa de movistar :).
-movistar me esta tomando el pelo !!!!
-rt @lukelechero: el comercial de movistar me da vergüenza sorry not sorry #partychilensisftlaroja
-que mal que andas movistar.
-hay 2 hitos que me dan una rabia incontrolable: los comerciales de @movistar y de @santanderchile. los del banco me enchuchan #losodioatodos
-rt @beelgrg: unas ganas de que movistar ande como la gente
-me anduvo para la mierda movistar, pero me comen el credito igual -.-
-que internet poronga movistar y la concha de tu madre
-rt @mrbzn: quién inventará los comerciales de movistar #partychilensisftlaroja
-@tomaswagner19 motorola es una masa. si no tendría iphone seguro lo compraría.
-@marttmicolich jaja me canso el motorola
-@anavonrebeur motorola g por mucho

-he comprado: 'motorola moto g (3ª generación) - smartphone libre android 6 (4g, pantalla 5", cámara 13... via @amazon @moto_col
-@stphns en lo personal se me hacen mas chidos los telefonos de motorola
-@ramirezmacielc nono, mi celular es motorola, el bb es otro jajaja
-si me regalaran un iphone lo vendería para comprarme un motorola es en serio
-q drogaaaaaaaa,q odioda motorola
-rt @mermeladx: si me regalaran un iphone lo vendería para comprarme un motorola es en serio
-@facucontreras11 el j7 es lo mas, y motorola el de tercera generación también esta buenisimmmo
-adidas me patrocina
-cejas de adidas ajjajajjajajaaj
-rt @zamorandres: gol de columbia! @adidas
-ajkjaajajajaja un traje adidas, que rata
-@antonellafinis venite a las pulgas y compra 5 camperas adidas por $ 100 jajaja
-rt @locombia__: los busos adidas>>>>>>>>>>>>>>
-hola @adidasco @adidas, it's culombia, not columbia.
-gol de columbia! @adidas
-adidas esta robada con las camisetas que saca
-en facebook apenas se enteraron que @adidas cree que es columbia, no colombia.
-"gol de columbia" - @adidas
-rt @lejuanito: columbia 1 / @adidas 0 #copaamerica
-columbia 1 / @adidas 0 #copaamerica
-adidas repitan después de mi: colombia, no columbia.
-los busos adidas>>>>>>>>>>>>>>
-@brianpushparaaj adidas
-poco a poco! acabo de correr 8.30 km con nike+. #nikeplus
-adidas se está tranformando en nike con las alteraciones de colores que hace
-acabo de correr 3.23 km con nike+. #nikeplus
-nike podia me patrocinar ...
-acabo de correr 4.60 km con nike+. #nikeplus
-acabo de correr 3.96 km a un ritmo de 6'54"/km con nike+. #nikeplus

-acabo de correr 4.67 km con nike+. #nikeplus
-acabo de correr 13.0 km con nike+. #nikeplus
-acabo de correr 4.01 km con nike+. #nikeplus
-acabo de correr 4.63 km con nike+. #nikeplus
-acabo de correr 6.01 km con nike+. #nikeplus
-acabo de correr 4.06 km con nike+. #nikeplus
-acabo de correr 3.37 km con nike+. #nikeplus
-@camilatamaraaa por otro lado nosotros somos pobres. yo vi un conjunto nike hso en el sport y todavía no me lo puedo comprar
-acabo de correr 6.10 km con nike+. #nikeplus
-acabo de correr 7.09 km con nike+. #nikeplus
-@christianleave nike or adidas
-acabo de correr 7,50 km con nike+. #nikeplus
-acabo de correr 3.84 km a un ritmo de 11'00"/km con nike+. #nikeplus
-acabo de correr 7.22 km con nike+. #nikeplus
-acabo de correr 6.37 km a un ritmo de 7'06"/km con nike+. #nikeplus
-acabo de correr 0.30 km a un ritmo de 11'14"/km con nike+. #nikeplus
-#run #monumentoalarevolucion #cdmx #mexico #mx #correr #nike+ #nikeplus acabo de correr 5.57 km con nike+. #nikeplus
-acabo de correr 5.31 km con nike+. #nikeplus
-@malenasesma trae iphone y zapatillas nike plis aa
-ahí la llevo... acabo de correr 6.01 km a un ritmo de 6'00"/km con nike+. #nikeplus
-@eze_giuliani eso por que no me puse los nike jajaja
-acabo de correr 5.00 km con nike+. #nikeplus
-acabo de correr 5.58 km a un ritmo de 5'50"/km con nike+. #nikeplus
-acabo de correr 5.15 km con nike+. #nikeplus
-acabo de correr 4.61 km con nike+. #nikeplus
-ta re juerte messi con barba
-rt @tacuara_cardozo: argentina sin messi, brasil sin neymar, uruguay sin suárez, bolivia sin irrazabal, paraguay sin tacuara. no estrellas,...
-mas traidor el falso messi ese ! #gh2016
-te doy hasta que messi deje de evadir impuestos
-@pipirc22 @marlonnando @elpipasanchez @hugohlamadrid dinho y el diego no

entran en la misma tabla. messi tampoco.
-@lokorojinegro @zorrotapatio @luissiemprefiel @gustavoguzmans @atlasfc si, ese messi es todo un muerto
-curry esta tocado con la misma vara que messi, lo marcan 3 a él solo.
-curry es el messi de la nba, lo defienden hasta los aficionados
-agustin cuantas tenés bajas
eeeeee messi
-rt @rinaudoopina: dani la muerte se recuperó en 20 días de 7 balazos y messi hace 15 que está llorando por una patada en las costillas
-rt @marianogottig: dani la muerte se recupero en 20 dias de 7 balazos y messi hace 15 que esta llorando por una patada en las costillas
-je lacoste en lacoste
-#lacoste super miercoles!!
-@inakiatm @ripollrafa el guarro hijo de puta lleva con el mismo lacoste veinte años
-lacoste & mendrez
-enamorado estoy de mi suéter lacoste
-sale un lacoste entra otro lacoste **
-@mari_nia_t @pollomaldonado la caja de zapatos de freddy valentín y las polos lacoste de ñañito #regalospalajunta
-usamos droga de lacoste
-nos anda todo de lacoste
-@tomiirolon (lacoste vos no andas en cualquiera).....
-usamos drogas de lacoste
-@ner12k vans , adidas, lacoste ☹□❤□
-@cecilee1000 la banane lacoste, la chicha
-@analiafranchin al prox si es varon le pone "lacoste"
-rt @bonsoirmessieur: @cecilee1000 la banane lacoste, la chicha
-cartagena se veria afectada por el aumento del nivel del mar asosiado al cambio climatico. "mathieu lacoste" día de los océanos. @cm_botero
-@saulhr es como los cocodrilotes de 15 centímetros en los polos lacoste. o el gap estampado en el culo de los pants.
-@gravdur enzo lacoste

-calzado #lacoste #super #miercoles !!!
-@kiemadrid @inakiatm @ripollrafa el sempiterno polo lacoste
-@moranzonidiego directtv y si no el canal de la ciudad
-alguien con directtv pls :(jajaja
-rt @peppertagalleta: @andynsane los que tienen directtv recién estarán gritando gol
-ya encontré link de directtv, lero lero.
-@mariamercedez11 fox. en directtv es el 202.
-estos de directtv se dan palo durísimo entre todos.
-@pelotazo tienes días libres y directtv
-cagada ser pobre y no tener directtv.
-#eng - #rus
fútbol fútbol fútbol □□□
(directtv o telecanal)
-el grito de gol del narrador de directtv !
-@andynsane los que tienen directtv recién estarán gritando gol
-@bollino te acaban de saludar en la transmision de directtv...
-ese momento en el que escuchas tu nombre en un partido de ecuador retransmitido por directtv sport y tu o.o, ah caramba.
-rt @ludmicarrera23: exijo un delivery de pepsi.
-@crazymadafaka16 un dia, a un bato lo mandaron a usa por coca (c0cayna) y trajo pepsi
lo llevaron manejar a un restaurant y lo chocó
-5 conto numa pepsi
-pepsi con tingui
-eles falando do pepsi , q lindoss
-exijo un delivery de pepsi.
-@ritasolangee @_japochavez @gasteeea rita copate y buscarme la pepsi abajo □
-necesito tomar pepsi
-rt @solo_junior: "llevo dos años sin tomarme una pepsi" cheché hernández.
-rt @eber_bard: que relajación es tomar sprite❤
-qué ganas de un calentadito de fríjoles con sprite.
-muero por una sprite y doritos

-que relajación es tomar sprite❤
-rt @arugirard: que ganas de vino con sprite diiiiooosss
-enrola que tengo una moña más verde que linterna verde y botellas de sprite
-codeina con sprite pal carajo el ron
-rt @emiliaezv: mendas ganas de un tintillo con sprite
-me vino la latita de la sprite a tope y me empape
-tomando una sprite chorra
-rt @bassiststark: @virshepard 7up no, por dioooooooooos xd sprite si eso...pero pls xd
-me trajeron gomitas y sprite y se las regale a ellos , dieta y lpm
-jurao jurao' que yo me atrevo a cambiar a kevin love por dos cheese dogs y una sprite de lata. pelo a pelo
-esta sprite sabe a alka-seltzer o como sea que se escriba
-@celesgarbin @halecsis 1 smirnoff entre 3 con una sola sprite de 2 litros entendela
-@kisiva19 vele la sprite pues jajajaja. mera 'esprai'
-donde estan las chicas del vip las que toman champagne con sprite
-hamburguesas y sprite
-@dahiitb jajajaajajaja sabelo ñe vamo a tomar ese tinto con sprite
-mendas ganas de un tintillo con sprite
-@_zpta jaja la próxima agrégale también sprite al jarabe!!!
-rt @tatifloww: enrola que tengo una moña más verde que linterna verde y botellas de sprite
-tu, yo, happy hour starbucks... no sé, piénsalo.
-@soyjoseray @analaloga 8:30 pero starbucks cierran más tarde.
-unas ganas de starbucks
-@juancruzfr o starbucks y acompañarlo con cine
-rt @davidcorrrea: primer día de dieta y ya me tomé un frappe de starbucks
-@geekbarista @starbucks @starbucksar tienda 100 #mastil
-qué ganas de starbucks
-ni crean que vine a starbucks porque está al 2x1 todo el día.
—yo siendo súper pobre :(
-primer día de dieta y ya me tomé un frappe de starbucks

-@axelcardin y el cafe starbucks
-rt @marianelav08: muuuuero por un starbucks, una crepa y spikes !!
-@camiiayelen08 ay la puta q lo parió, te conté que tengo un menú especial. igual me consuelo suponiendo que en starbucks hay leche deslacto
-@jazgrieco para! como que no hay starbucks!
-@juarezdai @marciopedreira4 yo te vi a vos en el starbucks d callao cerra las nalgas daiana

15. TABLA DE ILUSTRACIONES