



Universidad Nacional de Educación a Distancia
Máster en Lenguajes y Sistemas Informáticos

**Aproximación léxico-semántica para el
análisis de sentimiento en tweets: Pruebas y
evaluación en la colección TASS**

Autora: Esther Villar Rodríguez

Directora: Ana M^a García Serrano

Febrero 2013

Agradecimientos

Sin la beca que me concedieron el año pasado no estaría escribiendo esta memoria. Por lo tanto, la primera mención en este trabajo debe ser para mi responsable dentro de TecNALIA, Marta Gonzalez, que depositó en ya dos ocasiones su confianza en mí. Gracias por preferir repetir en vez de mandarme a la cola del paro.

Mis cuantiosas horas de incertidumbre y angustia me han desvelado también que los compañeros no sólo son desconocidos que ocupan mesas a mi alrededor.

Por último, a mi directora Ana García Serrano por el tiempo dedicado no sólo a mi formación sino al apoyo moral. Tus “ánimo Esther” han sido los mejores mensajes, la mejor lección.

CONTENIDO

INTRODUCCIÓN.....	7
PARTE 1: ESTADO DEL ARTE O PANORAMA TECNOLÓGICO.....	10
1. Comunicación en las redes sociales.....	12
1.1 Discusión.....	14
2. Trabajos relacionados.....	16
2.1. Preprocesamiento.....	16
2.2 Aproximaciones existentes en la bibliografía.....	17
2.4. Discusión.....	23
3. EVALUACIÓN.....	25
3.1 Recuperación de información.....	25
3.2 Medidas de evaluación.....	26
3.3 TASS: corpus de prueba.....	32
4. Planteamiento.....	34
PARTE 2: TRABAJO REALIZADO Y EXPERIMENTOS.....	37
5. Funcionalidades básicas.....	40
5.1 Preprocesamiento 1: Tratamiento de características del universo Twitter... 40	
5.2 Preprocesamiento 2: Extracción de polaridad de emoticonos.....	41
5.3 Segmentación básica.....	42
5.4 Preprocesamiento 3: Estandarización.....	42
5.5 Cálculo de mayúsculas.....	43
5.6 Polaridad de interjecciones.....	44
5.7 Segmentación compleja.....	45
5.8 Detección de entidades.....	56
5.9 Corrección ortográfica.....	57
5.10 Asignación de polaridad previa.....	58
5.11 Tratamiento de modificadores.....	60
5.12 Inversión de polaridad – Negación.....	64
5.14 Procesamiento del estilo directo e indirecto.....	68
5.13 Obtención de la polaridad final.....	69

5.14 Creación de características	71
6. Experimentación y pruebas	72
7. Conclusiones y futuras líneas de trabajo	78
8. BIBLIOGRAFÍA.....	83
PARTE III: ANEXOS.....	85
ANEXO 1: Estructura de datos.....	85
ANEXO 2: Cuestiones técnicas sobre la experimentación	88
ANEXO 3: Ejemplos sobre el nuevo corpus anotado	91

FIGURAS

Figura 1: Conceptos principales para la definición de medidas.....	28
Figura 2: Medidas de valor simple.....	32
Figura 3: Arquitectura general.....	38
Figura 4: Preprocesamiento 1.....	40
Figura 5: Preprocesamiento 2.....	41
Figura 6: Extracción de polaridad de interjecciones.....	44
Figura 7: Árbol sintáctico para el caso 1.....	48
Figura 8: Árbol sintáctico para el caso 2.....	49
Figura 9: Árbol sintáctico para el caso 1.....	51
Figura 10: Árbol sintáctico para el caso 2.....	52
Figura 11: Árbol sintáctico para el caso 3.....	54
Figura 12: Detección de entidades.....	56
Figura 13: Asignación de polaridad previa.....	58
Figura 14: Tratamiento de modificadores.....	60
Figura 15: Ámbito de los modificadores.....	62
Figura 16: Agrupación de modificadores.....	63
Figura 17: Inversión de la polaridad.....	64
Figura 18: Procesamiento del estilo directo e indirecto.....	68
Figura 19: Acto comunicativo.....	79
Figura 20: Diagrama de clases de elementos sintácticos.....	86
Figura 21: Pool de procesos.....	89
Figura 22: Proceso.....	90

TABLAS

Tabla 1. Definición de conceptos para el estudio de la sensibilidad.	30
Tabla 2. Ejemplo de entrada en el diccionario de emoiconos.	41
Tabla 3. Ejemplo de entrada en el diccionario de intersecciones.	45
Tabla 4. Ejemplo de entrada en SentiWordnet	59
Tabla 5. Límites de las polaridades.	59
Tabla 6. Cálculo de los modificadores.....	63
Tabla 7: Resultados del trabajo para la medida de precisión.	73
Tabla 8: Precisión para 20.000 tweets con la 4ª configuración y detección de entidades.	73
Tabla 9: Precisión para 20.000 tweets con la 4ª configuración sin detección de entidades.	73

INTRODUCCIÓN

Mi trabajo de fin de master se ha orientado en la dirección del *Sentiment Analysis*, una disciplina que esta última década ha acaparado la atención de muchos investigadores interesados en la posibilidad de hacer comprender el lenguaje humano a las computadoras y por ende las opiniones de los usuarios, donde se halla el último objetivo.

El verdadero propulsor de este movimiento analítico fue el boom de la Web 2.0 y las redes sociales que generaron nuevas plataformas de opinión donde los usuarios expresaban libremente sus sensaciones y debatían los diversos temas candentes del momento aflorando foros ipso facto. No tardaron mucho grandes firmas, analistas e inclusive el mundo de la política en ver el potencial subyacente bajo Facebook, Twitter, blogs,... y poner una gran cantidad de recursos a disposición de la investigación.

El objetivo se centraba en, de forma automática, comprender y clasificar las opiniones de los usuarios sobre los temas de interés para la investigación. Dicha clasificación permitiría seleccionar una polaridad para cada mensaje. Obviamente, estos resultados se convierten en semillas para fantásticas campañas de marketing, como *feedback* instantáneo para empresas que desean conocer de primera mano la valoración de sus productos/servicios o incluso como método de predicción de resultados de elecciones políticas.

En estos, mis primeros comienzos en la investigación, he pretendido comprender más a fondo el funcionamiento de estas técnicas e indagar en los resultados obtenidos por las investigaciones más fructíferas de la última década. Y qué mejor para adentrarse en este mundo que intentar crear un entorno para mi investigación mezcla de aquello que más interesante he encontrado en mis lecturas. Para dicho entorno, he elegido Twitter como elemento central.

En las siguientes líneas concretaré un poco más con el fin de esquematizar el contexto y mostrar una idea general de este entorno en que se plantea la investigación.

A la hora de abordar un problema de análisis de textos cortos, como en este caso, el primer paso es averiguar qué buscamos y dónde podemos encontrar pistas para lograr el objetivo. Estas pistas se tratan de características que si bien no son perceptibles a primera vista, pueden ser minadas y obtenidas tras un procesamiento previo. Independientemente del enfoque en que basará el sistema, bien un enfoque lingüístico o uno orientado a Minería de Datos, el conjunto de características suele converger en el mismo punto aunque el tratamiento difiere considerablemente.

En los sistemas de Minería, delegamos la creación de un modelo de clasificación al algoritmo elegido (en *Opinion Mining* suele emplearse con mucha frecuencia Máquinas de Vector Soporte) a partir de unas primeras características seleccionadas del corpus. Este modelo trata de extraer patrones observables en el corpus y los transforma de forma que el resultado recoja el conocimiento extraído de esa información estructurada.

En los enfoques lingüísticos el procesamiento suele realizarse en línea, por fases, donde la mayoría de los módulos trabajan a distintos niveles la lingüística y la semántica para conseguir una clasificación basada en la comunicación y en el lenguaje como tal. Es un enfoque dirigido a la comprensión del lenguaje, en la línea de la lingüística computacional. Y esta será la dirección que seguirá mi trabajo.

Estudiando las aproximaciones existentes, se planteó la realización de un conjunto de módulos que combinados convenientemente, abordaran la tarea de identificar la opinión, el acuerdo o las entidades sobre las que se opina. Todo ello es una tarea de actualidad pero muy compleja, en la que intervienen aspectos lingüísticos (análisis de diferentes niveles, diccionarios...) y extralingüísticos y propios del contexto de los tweets (jerga, simbología de sentimientos, etc).

Especial interés se ha dedicado al estudio de la gramática española sobre fenómenos que aportan sentimiento a las expresiones, como son:

- El tratamiento de modificadores.
- El tratamiento de inversión de polaridad /negación.
- La detección de objetividad en estilo indirecto y enunciados de citas.

Este interés ha llevado a modificar la actual implementación de la librería Freeing.

Además había que probar la combinación de los módulos desarrollados para la tarea y evaluar los resultados. Para ello, es necesario procesar y analizar corpus que dispongan de juicios de relevancia. Se seleccionó TASS, que al ser un corpus de gran magnitud aumentó la complejidad de las pruebas con respecto a la eficiencia.

Para las pruebas (sobre 70.000 tweets) se definieron cuatro configuraciones para probar la influencia del tratamiento de: las entidades nombradas, los modificadores, la inversión o negación y el estilo directo o indirecto.

Los resultados se presentan en el apartado correspondiente de esta memoria, aunque tan interesante como el planteamiento, diseño y pruebas sobre el corpus TASS, haya sido el descubrimiento de los aspectos menos resueltos en esta área de

investigación y que me pueden permitir el planteamiento y la realización de una nueva investigación.

PARTE 1: ESTADO DEL ARTE O PANORAMA TECNOLÓGICO

En el mundo de la Web 2.0, Twitter es una de las redes sociales más utilizadas, en la que, además, la lengua española está en segundo lugar entre los idiomas empleados, seguido evidentemente del inglés (<http://www.20minutos.es/noticia/1700292/0/espanol/twitter/instituto-cervantes/>), por lo que su procesado es de incuestionable actualidad. En este apartado del estado del arte se comenzarán describiendo los aspectos más destacados de esta herramienta.

Twitter ha revolucionado el mundo del *Sentiment Analysis* dadas sus peculiaridades. El tamaño del texto provoca complicaciones extras a la hora de analizar la opinión pero también puede ser a la vez su punto fuerte dado que los usuarios se ven más forzados a ser doblemente explícitos y no “andarse por las ramas” (lo que en ocasiones causa estragos en los modelos de clasificación).

Lo que inmediatamente observa un investigador al revisar un corpus de tweets es que es prácticamente imposible manejar los mensajes sin una normalización previa [16]. Es por ello que el segundo apartado del estado del arte se dedica a este aspecto.

La normalización es un proceso común a cualquier tipo de enfoque, sea un enfoque por fases basado en lingüística o un enfoque orientado a la Minería.

Los primeros, también denominados enfoques léxicos, permiten un tratamiento más complejo a manos del investigador, sin olvidar que el resultado de esta computación puede ser también la entrada o comienzo a un sistema de Minería. Uno de estos tratamientos que se ha convertido en asunto de interés y que ha repercutido en modelos más fiables es la creación de modelos semánticos [17] que definen el dominio y modelan singularidades específicas del mismo a través de reglas y diccionarios. Este tipo de módulos trabajan más a fondo con la complejidad del lenguaje natural y abordan el concepto de contexto que en la vida real tanto repercute en el significado de las cosas. Ambas aproximaciones se presentarán con más detalle en el siguiente apartado.

Sea cual sea el modelo o el enfoque, el fruto final es un clasificador de polaridad que debe ser capaz de captar la esencia de los mensajes sea cual fuere su representación, su naturaleza o su contexto.

Para el cálculo del *agreement* o acuerdo, se han empleado medidas matemáticas como la desviación típica [17] que manifiesta el grado de variabilidad en las opiniones y es como consecuencia de ello, un síntoma evidente de diferencia de criterios.

Aún después de todo el procesamiento y una vez concluida una tarea concreta, los investigadores suelen toparse con la cruda realidad de que la precisión no es siempre la que se espera después de tan laborioso trabajo. *Sentiment Analysis* es aún un área en el que queda mucho por explorar. La propuesta concreta abordada y los trabajos realizados en este trabajo de iniciación a la investigación se presentan en la segunda parte de esta memoria.

1. Comunicación en las redes sociales

De acuerdo a una encuesta realizada por The Kelsey Group en 2007 [19], los consumidores están dispuestos a pagar entre un 20% y un 99% más si se trata de un producto calificado con 5 estrellas en vez de con 4. Este dato constituye la prueba evidente de que la opinión de los usuarios es el arma más potente en las estrategias de marketing de las empresas. Y este hecho es más acusado cuando se trata de hoteles o restaurantes. Dicho descubrimiento fue la mecha que propulsó la investigación de las opiniones creando una nueva área de investigación denominada *Sentiment Analysis* u *Opinion Mining*.

Entre las redes sociales, Twitter es de las primeras para varios indicadores: número de usuarios, comunidades, interés para las compañías etc. Twitter es un servicio de microblogging que usan ya más de doscientos millones de usuarios. Su gran apuesta radica en la agilidad de los mensajes cortos y en el dinamismo que se crea en la comunicación evitando que los usuarios deban invertir mucho tiempo en conectarse con el resto de la esfera social. Esta ligereza en la creación de los mensajes hace que se publiquen miles de ellos constantemente, que surjan comunidades en base a similitudes y que se cree un universo interconectado en constante movimiento donde las opiniones fluyen, se comparten y se contagian a la velocidad de la luz.

Un tweet es el mensaje que enviamos a través de Twitter y que consta de 140 caracteres máximo (aunque la mayoría de ellos rondan entre 14 y 78 caracteres). El Tweet puede contener letras, números, signos o enlaces. El estilo del lenguaje empleado es en su mayoría coloquial y suele incluir multitud de faltas de ortografía o expresiones idiomáticas. Como consecuencia de este tipo de lenguaje familiar y distendido, los mensajes no se rigen por un correcto estilo y en ocasiones, en un afán del usuario por plasmar sentimiento o enfatización se emplean indebidamente signos de puntuación, mayúsculas o repetición de letras.

Una vez recorrido de forma general el universo Twitter, vamos a concretar un poco más con el fin de esquematizar el contexto. El primer paso es averiguar qué buscamos y dónde podemos encontrar pistas para lograr el objetivo.

Las características se pueden agrupar dentro de la siguiente clasificación:

1. **Características pertenecientes al universo Twitter:** son aquellas que existen como tal sólo en este contexto ya que especifican propiedades concretas del lenguaje empleado en los tweets.
 - *Targets* o *mentions*. Muy distinguibles dentro de un tweet por tener una representación muy peculiar que consta de la @ seguida de un

usuario o grupo (Ejemplo: @Raul). Se usa para mencionar o contestar a otros usuarios.

- *Hashtags*. Permiten agrupar mensajes sobre un mismo tema y su uso se estandariza mediante la # y el *topic* adherido. (Ejemplo: #Elecciones).

2. **Características pertenecientes al lenguaje de las redes sociales.** Son aquellas características que engloban usos especiales del léxico cuando se emplea en lenguaje coloquial.

- Emoticonos. Muy conocidos por su uso extendido en las redes sociales, SMS's y otras plataformas donde el lenguaje coloquial impera.
- URL's. En Twitter se dan una gran cantidad de debates sobre los temas más candentes en cada momento. Estas opiniones suelen ser apoyadas o difundidas gracias al uso de las URL's.
- Uso de Argot. En diferentes contextos sociales, los hablantes emplean una variedad del habla específica que les diferencia del resto y que suele ser utilizada en un periodo corto de tiempo. En Twitter existe un uso extendido de esta jerga social y es importante anotarla e incluso comprenderla dado que si se emplean diccionarios o recursos lingüísticos es posible que dichos términos o expresiones no sean hallados. Un típico caso es el de las abreviaturas, por ejemplo: "xo".
- Texto en mayúsculas. Suele emplearse para dotar de mayor énfasis a la sentencia.
- Interjecciones o expresiones típicas. Expresiones positivas y negativas ampliamente utilizadas en internet ("mecachis", "hurra", "snif",...).
- Repeticiones. Es un comportamiento muy típico en las redes sociales o en las plataformas de la Web 2.0. Dada la pérdida de expresividad y frescura y la estaticidad del lenguaje escrito, se usan formas inventadas para mostrar la espontaneidad del lenguaje hablado o de intensificar significados. La repetición y el uso inapropiado de mayúsculas son quizás las herramientas más evidentes.

3. **Características lingüísticas.** Obviamente, allí donde se usa el lenguaje existen aunque de forma subyacente y como producto de un exhaustivo análisis, características lingüísticas. Son aquellas que revelan conocimiento sobre la lengua empleada en sus diferentes niveles. Implican análisis morfológicos, gramaticales, sintácticos, ...

- Composición de los textos. Característica más vinculada a la sintaxis y a la comprensión de cómo estructuras sintácticas más pequeñas se aglutinan para componer estructuras mayores mediante la subordinación o coordinación.
- Categoría gramatical de las palabras. Aunque puede parecer a simple vista una característica con poco peso dentro del sistema, es todo lo contrario. En el mundo de la computación, recoger la semántica es un camino muy adverso. Las máquinas no comprenden de significados, sólo de símbolos. ¿Cómo distinguir entonces símbolos con significados diferentes? La categoría gramatical puede ser un primer paso para discernir entre palabras homógrafas. Más adelante veremos que existen otras pistas aún más explícitas.
- Morfología de las palabras. Permite entender cómo se forman las palabras mediante las raíces gramaticales o lemas y los afijos, y posteriormente ayuda a distinguir términos por lexemas en vez de por palabras.
- Uso de la negación. Detectar el ámbito de la negación y tratarla es uno de los pasos más complejos en el análisis de los textos pero extremadamente fundamental para detectar la polaridad, ya que puede revertirla completamente.

Como se puede observar, esta clasificación no es estanca y, aunque algunas de las características podrían verse clasificadas en más de un tipo, es un método rápido para comprobar en qué cosas nos podemos fijar para crear un modelo fiable de clasificación de polaridad según diferentes niveles de procesamiento o naturaleza.

1.1 Discusión

La comunicación en las redes sociales ha crecido vertiginosamente en la última década. La explosión de la Web 2.0 genera plataformas y hervideros para que las experiencias personales se compartan y se estimule el *feedback* constante de juicios. Bien es cierto que desde hace algunas décadas, cuando los medios de comunicación comenzaron a servir de escenario para las grandes firmas, la sociedad tiene la sensación de ser “arrastrada” o “engañada” por campañas de publicidad que distan mucho de ofrecer al usuario información veraz. Esta terrible desconfianza lleva cada vez más a los consumidores a buscar nuevas vías para cerciorarse de las capacidades de los productos. Es en ese punto exacto, cuando el potencial de las redes sociales y los blogs parece reafirmarse y dejar patente su labor de trasmisión de la realidad de

los ciudadanos *de a pie*. Y como una cadena más que se retroalimenta, volvemos a las empresas u organizaciones y a su afán de generar cuantiosos resultados. La posibilidad de comprender qué se está comentando acerca de sus productos/servicios es el arma más valiosa para darle una vuelta más a la cadena y emplear dicha semilla para reactivar las campañas de marketing. Para unos y para otros, la comunicación es la única vía para la sustracción de conocimiento.

Centrándonos en el análisis de sentimiento, nos encontramos ante varios retos de no fácil resolución. Algunos pertenecen a la complejidad del lenguaje natural, otros derivan de la necesidad de definir qué es subjetivo y qué no lo es y otros son específicos de la plataforma donde se aplican. En la mayoría de ocasiones, todos confluyen en la tarea presentando un conglomerado que paso a paso va reduciendo la precisión de los sistemas.

2. Trabajos relacionados

Es fácil intuir que cuando una temática resulta tan atractiva para varios sectores, la investigación coge un ritmo acelerado y en relativamente poco tiempo aparecen gran cantidad de trabajos al respecto.

Lo cierto es que *Sentiment Analysis* es aún un *recién nacido*. La mayoría de los esfuerzos se centran en un pequeño número de propuestas que parecen ir en la dirección acertada, pero aún queda un horizonte muy amplio por descubrir. Al menos tan extenso como posibilidades abarca el lenguaje natural. Y es que esta disciplina es tan compleja que parece que convertirá permanentemente a la minería de opiniones en un objetivo demasiado ambicioso en términos de precisión.

Analicemos uno a uno los campos donde los estudios han construido mejores estrategias, aunque algunas de éstas, como se podrá comprobar por la agrupación de referencias, son explotadas por casi todos los trabajos. La mayoría de ellos comparten enfoques e influyen decisivamente en los trabajos posteriores, creando una línea de investigación sin demasiados salientes o ramificaciones.

2.1. Preprocesamiento

Twitter ha revolucionado el mundo del *Sentiment Analysis* dadas sus peculiaridades. El tamaño del texto provoca complicaciones extras a la hora de analizar la opinión pero también puede ser a la vez su punto fuerte dado que los usuarios se ven más forzados a ser doblemente explícitos y no andarse por las ramas (lo que en ocasiones causa estragos en los modelos de clasificación).

Para abordar cualquier tarea sobre un corpus de tweets, es necesario aplicar a los mensajes una normalización [16]. Este preprocesamiento suele incluir la sustitución de hashtags, urls y targets, el remplazo de caracteres repetidos, la traducción de las abreviaturas e inclusive, los más escrupulosos, suelen proceder a una corrección ortográfica [1].

Otro de los elementos que requiere especial atención son los emoticonos. En el artículo [16] toman esta característica típica del *micro-blogging* como *feature* binaria para el clasificador y reportan buenos resultados en el modelo. En los enfoques léxicos [1, 7] los utilizan como indicios tempranos de polaridad y conforman una característica más a tener en cuenta.

La mayor traba, una vez que el corpus es medianamente manipulable, reside en encontrar el conjunto de características o pistas fiables que puedan ser claves a la hora de determinar la polaridad, el sentimiento el tema o cualquier otro aspecto. Y no es una cuestión de número, sino de su adecuación a las tareas y su correcta extracción.

2.2 Aproximaciones existentes en la bibliografía

Existe un punto en el que es fundamental establecer el enfoque del trabajo. Cualquiera de ellos se basará en la identificación de patrones relevantes a través del uso de sistemas informáticos pero existen notables diferencias que nos debe hacer decantarnos por unos u otros.

Los métodos léxicos requieren de grandes cantidades de conocimiento, ya sea lingüístico como semántico, que es muy difícil y costoso de obtener. Por el contrario, la ventaja obtenida es la capacidad de adaptar el sistema con más facilidad sin la necesidad de reconstruirlo prácticamente desde cero. Esta manipulación directa permite un control total sobre los resultados y la gestión del proceso en general.

Los enfoques basados en Minería requieren un preprocesamiento de los datos muy exhaustivo pero el resto del proceso es mucho más ágil. El problema radica en que los modelos creados suelen ser sumamente complejos y totalmente atados a la estadística y los números, con lo que en cuanto se descubre un cambio en el entorno, todo el modelo puede venirse abajo y se precisa una reconstrucción casi completa. Además, los resultados son difícilmente entendibles y por tanto la tarea de evaluación y aprendizaje puede ser muy laboriosa.

2.2.1 Aproximación léxica

Los enfoques por fases basados en lingüística [15, 17] permiten un tratamiento más complejo a manos del investigador sin olvidar que el resultado de esta computación puede ser también la entrada o comienzo a un sistema de Minería. Normalmente, incluso los trabajos que han optado por Máquinas de Vector Soporte no han podido obviar que Sentiment Analysis es un proceso que requiere de tareas de PLN (Procesamiento de Lenguaje Natural) y han añadido algunas de ellas como la tokenización u obtención de la categoría gramatical como primeros pasos para la definición de las características usadas en el modelo [16, 17].

La *tokenización*, que puede ir acompañada por la extracción de la raíz gramatical, es indiscutiblemente ineludible ya que los términos (bien recogidos en cualquier tamaño de *n-gramas*) son uno de los máximos transportadores de significado. Este hecho se constata cuando queda patente el aumento de precisión obtenido al añadir diccionarios de polaridad. Ejemplos de estos diccionarios son SentiWordnet [1, 18], DAL (Dictionary of Affect in Language) [7], MPQA [16],... Estos términos junto con su polaridad pasan a formar parte del vector de características pero el enfoque basado en términos independientes suele ser insuficiente.

Otros estudios aportan nuevas vías como es la extracción de patrones que permiten identificar expresiones subjetivas más allá de los *n-gramas*. No se apoyan, por tanto, en la secuencialidad de las palabras y permiten albergar construcciones más complejas [12]. En [3] analizan patrones típicos del dominio de la política y proponen un proceso de generalización de dichas características para entrenar el clasificador.

Por otro lado, el *POS Tagging* o anotación de la etiqueta gramatical ha demostrado ser una característica indispensable para ambos enfoques. Algunos estudios revelan que ciertas categorías tienen más responsabilidad que otras a la hora de definir la polaridad [8]. Los textos subjetivos contienen más pronombres personales (en especial la primera persona para hablar de sí mismos y la segunda para invitar a la audiencia), hacen mayor uso de verbos modales (ya que expresan emoción), emplean más verbos en tiempo pasado (ya que suelen indicar negatividad y demuestran disconformidad o negatividad hacia algo del pasado), utilizan más expresiones idiomáticas, y existe una aparición más notoria de adjetivos superlativos y adverbios (advirtiendo positividad).

Pero existen más características que son compartidas tanto por los enfoques de minería como lingüísticos. Estas características son las que podríamos agrupar dentro del grupo de modificadores de la polaridad. ¿A qué se llama modificador de la polaridad? La polaridad no puede verse como una evaluación aislada basada simplemente en un conjunto de términos definidos como características. La polaridad es un concepto más complejo que puede, y de hecho, debe sufrir modificaciones según procesamos de forma lingüística un mensaje. Así, no obtenemos la misma polaridad al hallarnos delante de una negación, o cuando aparecen síntomas de lenguaje coloquial haciendo uso de las pocas formas de enfatización que se permiten en el lenguaje escrito. Es ésta la parte del procesamiento que más ha cobrado atención por los investigadores y donde se han descubierto nuevas pistas con tanto o más carga semántica que las propias palabras. En este subgrupo entra la repetición de letras, el uso de mayúsculas, las interjecciones, los emoticonos o el empleo de la modalidad irreal (verbos en subjuntivo) [15].

Con los enfoques basados en Minería, algunos procesamientos lingüísticos básicos en este tipo de análisis, como lo es la detección de la negación, han sido superficialmente tratados y es ahí donde los enfoques basados en lingüística pretenden demostrar su potencial [15, 17, 24]. Estas estrategias imprecisas suelen emplear el uso de n-gramas compuestos donde adhieren una partícula negativa para tratar la compleja cuestión de la negación [8, 29]. Además de la negación, aparece otro asunto débilmente estudiado como es la segmentación [1, 17] y que puede ser crucial a la hora de localizar el ámbito de la propia negación o distinguir diferentes posturas u opiniones dentro de un mismo tweet.

La detección de entidades es un proceso íntimamente vinculado a la segmentación ya que cuando se trata de buscar la polaridad para una entidad concreta es necesario delimitar su estadio de actuación dentro del documento. Este hecho se agudiza cuando nos encontramos frente a un sistema de análisis de reputación [18] donde todo el análisis gira en torno a la figura de la entidad. El enfoque seguido por el artículo de Chao Yang, Sanmitra Bhattacharya y Padmini Srinivasan (Yang C, Bhattacharya S, Srinivasan P: [Lexical and Machine Learning approaches toward Online Reputation Management](#). Proceedings of RepLab, Conference and Labs of the Evaluation Forum (CLEF) 2012; September 17-20 2012; Rome, Italy) trata las entidades no sólo como palabras a resaltar y rescatar del texto, sino como un conjunto de términos o keywords que permitirán desambiguar dicha entidad sea cuales sean las palabras empleadas en el comentario.

Pero no siempre las características aisladas acogidas dentro de un tweet son suficientes para precisar la polaridad. Es entonces cuando irremediablemente hace falta acudir a otras estrategias para especificarla.

En [6], los autores proponen aprovecharse de las conexiones que los usuarios crean en las redes fundamentándose en el concepto de *homofilia*, que no es más que la percepción de que las personas tienden a relacionarse con personas de creencias similares. La construcción de un perfil alrededor del autor puede ayudar a determinar la polaridad cuando las otras variables permanecen en estado latente o son demasiado imperceptibles.

2.2.2 Aproximaciones basadas en corpus

La mayor traba una vez que el corpus es medianamente manipulable, reside en encontrar el conjunto de características o pistas fiables que puedan ser claves a la hora de determinar la polaridad, el sentimiento el tema o cualquier otro aspecto. Y

no es una cuestión de número, sino de adecuación a las peculiaridades ya citadas de la plataforma Twitter.

En este punto es importante determinar el enfoque que seguirá el sistema propuesto, ya que si optamos por un sistema basado puramente en Minería de Datos [4, 10, 16], la elección de características es la tarea primordial y casi más urgente dentro de la creación de un posible sistema y éste queda bastante restringido a *features* que son perfectamente visibles y manejables por el algoritmo elegido. De hecho, es tan crítico detectar las características correctas que existen diversos estudios únicamente centrados en dicha tarea [2, 9, 13].

En el caso de los métodos de minería, se pueden producir sobreajustes en los modelos debido a una mala selección de características. Siendo más concretos, con bastante asiduidad esto es debido al gran número de *features* en comparación con el tamaño del corpus. Algunos trabajos como [24] abogan por agrandar el corpus existente de tweets con nuevos mensajes extraídos de la API de Twitter para conseguir solventar este desequilibrio. De hecho, suele ocurrir que no todas las polaridades están igualmente bien representadas en el conjunto de *training*, lo que suele provocar serios problemas de clasificación y que el modelo no sea capaz de establecer correctamente las fronteras para aquellas polaridades con menor número de ejemplos. Una de las formas más directas para solucionarlo es emplear los emoticonos como arma para extraer mensajes de la polaridad menos reproducida dando por sentado que el emoticono es fiel representante de la polaridad en el mensaje [8]. Por ejemplo, [16] aprovecha un corpus llamado EMOT que contiene tweets rescatados siguiendo esta fórmula pero habiendo eliminado los emoticonos del mensaje posteriormente. Trabajos como [24], en cambio, los emplean para recolectar nuevos mensajes subjetivos y aprender y entrenar los clasificadores con corpus más potentes.

Para fijar el conjunto de características se han seguido diversas aproximaciones: el cálculo de la Información Mutua, Ganancia de información, Método del chi cuadrado [10] o conteo de *features* para cada clase, aunque éste último parece no ser una buena elección ya que algunos estudios han demostrado que el fenómeno de *hápax legomena* (palabras raras que aparecen poco frecuentemente) da muy buenos resultados en la clasificación de polaridad [8, 13, 28]. En [13] también se añade otro mecanismo llamado Similitud Distribucional basado en, como su nombre indica, hallar la similitud entre cada par de palabras (basándose en su uso pragmático o semántica contenida) y crear clústeres en base a esta medida de similitud. Finalmente, se tendría en cuenta la precisión (medida de evaluación típica empleada en los sistemas de Information Retrieval) no de una palabra aislada, sino de esa palabra en conjunto con las de su clúster (entendiendo el clúster como aquel

conjunto de palabras “similares”). O se toman todas las palabras del clúster o se eliminan todas ellas.

Por último, realizaré una mención especial para el tratamiento de la negación. La mayoría de las investigaciones acaban haciendo alusión en su apartado de conclusiones y resultados a este fenómeno y el motivo no suele ser el optimismo. Un vago tratamiento suele conducir a reducciones en la precisión, por ello cada vez más, los trabajos con enfoques lingüísticos prestan más atención a esta cuestión e intentan aprovechar las mejoras conseguidas por los estudiosos en la materia (que suelen optar por un enfoque de aprendizaje automático [27]). Según Roser Morante y Eduardo Blanco en [25] hallar el ámbito y el foco de la negación es una tarea muy compleja que debe emplear en su resolución completos corpus anotados. Estos corpus utilizan la técnica de anotación de roles semánticos que es la más acertada en este entorno y entre ellos se encuentran colecciones como Penn Treebank 1 o BioScope [26].

2.2.3 Tratamiento semántico pragmático avanzado

Cuando la estadística y el PLN no parecen suficientes para obtener una bondad significativa en el modelo, es necesario recurrir a niveles superiores y este hecho más se acentúa cuando tratamos con el lenguaje y sus peculiaridades.

Ya hemos indicado previamente que el contexto es un factor determinante en el significado. Esta afirmación es la que la pragmática defiende. Un texto sin entorno es un texto *cojo*. Por consiguiente, deberíamos estudiar el lenguaje en un discurso situado donde el contexto pueda interferir en las interpretaciones de forma crucial.

“la palabra es mitad de quien la pronuncia y mitad de quien la escucha”. Michel de Montaigne.

En el caso de artículo [3], los autores pretenden obtener un modelo capaz de predecir el comportamiento de los votantes en las urnas. Para dicho propósito, se centran en el estudio del dominio que les afecta, la política, y analizan patrones típicos con el fin de proponer un proceso de generalización de las características para entrenar el clasificador. Estas características o patrones son versiones avanzadas de expresiones o *n-gramas*; expresiones específicas del dominio tales como “El partido X ganará...” o “El partido Y no tiene...”). Obviamente, esta estrategia, como todas aquellas que traten de retratar un contexto, dará mejores resultados cuanto más reducido y acotado sea el dominio.

¹ www.cis.upenn.edu/~treebank/

Pero definir el contexto es un punto clave en ciertos procesos como es el caso de la extracción de entidades vinculadas a opiniones. En el pasado, algunos enfoques se basaban en diccionarios o en los hallazgos de palabras que comenzaban en mayúscula, pero estos intentos han resultado ser insuficientes. Pongámonos en el caso de una crítica hacia la famosa empresa Apple en un entorno donde la buena ortografía brilla por su ausencia y el uso de mayúsculas es menos informativo que los periódicos. ¿Qué pasaría si la palabra estuviera escrita en minúsculas? Si el asunto ya parece complicado, aún podemos darle una vuelta más de tuerca y añadir una nueva dificultad: el texto está escrito en inglés. ¿Cómo podemos discernir entre la empresa y la fruta? Algunos estudios recientes abogan por recopilar un contexto de las entidades a la vez que se recoge el contexto del documento. Dependiendo de las coincidencias podremos defender la detección de la entidad con más o menos acierto [20, 21].

El contexto también juega un papel esencial en el análisis de sentimientos. No siempre a una expresión se le puede asignar una cierta polaridad. Esta afirmación se sustenta en base a dos premisas: el dominio que se trata y la perspectiva desde donde se realizan las opiniones. En cuanto a la primera ya hemos comentado que una palabra, aunque pueda contener una polaridad previa, debería ser considerada desde el contexto en que se enmarca. Por ejemplo, ciertos términos inicialmente podrían considerarse positivos, pero cuando son aplicados a una expresión en un determinado dominio, la cosa cambia. El trabajo de Julio Villena Román, Sara Lana-serrano y Cristina Moreno García [17] lo argumentan con el siguiente caso de uso:

```
#INCREASE# increase|increment|grow|growth|gain|rise|go_up|climb  
#DECREASE# decrease|decrement|reduce|loss|do_down|descent  
  
reputation/#INCREASE# P  
reputation/#DECREASE# N
```

Este filtro especifica la polaridad de la expresión dados ambos términos obviando las polaridades previas que pudieran tener ambas palabras por separado. Esta regla forma parte de un modelo semántico generado para dicho contexto específico. Estos modelos contienen además diccionarios que permiten que los análisis sean más exhaustivos dada su adecuación al dominio. Por último, es necesario citar el posible uso de taxonomías cuando el objetivo final no se trata de definir simplemente la polaridad, sino que va más allá buscando clasificaciones de usuarios en cuanto a sus juicios sobre ciertos asuntos [22]. Para sustentar dicha complejidad estructural, se hace forzoso la creación de una taxonomía que puede incluso servir como método de inferencia para la consecución de la clasificación final.

En cuanto a la perspectiva, el artículo [21] analiza dicha característica cuando se trata de hechos objetivos en los análisis de reputación. A pesar de ser casi contrario a la

“legislación” y pautas del *Sentiment Analysis* que, como su nombre bien indica, debe analizar sentimientos y por lo tanto lenguaje subjetivo, los hechos factuales pueden repercutir en la reputación de una empresa. Veamos un ejemplo ilustrativo donde podremos comprobar que seguramente dependiendo de si te remites al cliente o a la empresa, la valoración será bien distinta:

“La empresa X aprovecha el tirón de la demanda para aumentar los precios.”

2.4. Discusión

Una vez que hemos estudiado los trabajos más relevantes en el área, pasamos a finalizar con algunas conclusiones.

Sea cual sea el enfoque perseguido, ciertos procesos son inevitables. La mayoría de ellos tienen en común dos tipos de procesamiento: los procesamientos lingüísticos básicos atados al lenguaje natural y aquellos vinculados a la herramienta Twitter y que gestionan las características relativas a las redes sociales y en particular a dicha herramienta.

Los primeros son inherentes cuando tratamos con el lenguaje dado que las características no son tan perceptibles a simple vista y deber ser obtenidas tras un procesamiento previo. Por ejemplo, al igual que en muchos sistemas de IR, no es factible o provechoso lidiar con términos en cualquiera de sus variantes ya que las coincidencias deberían basarse en raíces gramaticales y en el significado que transporta la palabra. Indistintamente tratándose de un plural o de un singular, la semántica contenida es la misma y el modelo no debería realizar distinción. En el caso de los enfoques basados en Minería, no tener en consideración esta apreciación podría acarrear la construcción de un clasificador con demasiadas *features* que podría producir un sobreajuste en el modelo.

Respecto a los segundos, nos encontramos ante un planteamiento o situación similar que en ambos enfoques trae numerosos problemas si no realizamos un preprocesamiento previo y convertimos unos textos improcesables desde el punto de vista del PLN en unos textos estandarizados que siguen las directrices de las reglas gramaticales de la lengua castellana.

Una vez que los mensajes pueden ser comprendidos por cualquier oyente, parecen estar preparados para su posterior procesamiento.

En las aproximaciones asentadas sobre SVM, el mayor esfuerzo recae en la selección de las características óptimas en número y conceptualmente significantes para la creación del clasificador. En los enfoques léxicos, en cambio, existe un arduo trabajo

previo de estudio de la lengua en los diferentes niveles que el sistema se dispondrá a analizar. Estos niveles, como hemos visto, pueden ir desde los simples procesamientos gramaticales a los complejos exámenes semánticos y pragmáticos.

Mientras que los sistemas de Minería intentan descubrir nuevas características que añadir a sus modelos y exploran en algunos casos teorías a veces poco fundamentadas, las estrategias lingüísticas argumentan la necesidad de un análisis más profundo del discurso como *quid* para obtener resultados fructíferos. Lo cierto es que hasta los más claros defensores de los clasificadores basados en Minería de Datos apoyan este razonamiento y acaban incluyendo en sus trabajos características heredadas de los enfoques léxicos.

El enfrentamiento se haya en decidir si el lenguaje, como otras disciplinas, yacen sobre la matemática y pueden resolverse gracias a esta o si es necesario un conocimiento humano de difícil traducción numérica para *lidiar* con la inexactitud y la complejidad de la comunicación.

3. EVALUACIÓN

Le evaluación es el proceso mediante el cual generamos una valoración en base a los criterios establecidos, interpretamos los resultados, los comparamos, si es factible, y sobre todo realizamos un análisis de los resultados obtenidos con el fin de descubrir qué aporte o contribución se ha realizado y qué posibles mejoras se podrían desplegar en un futuro.

Individualmente, como proyecto independiente, podría verse como la última parte del proceso, pero si tratamos de visualizar la investigación en su conjunto, observaremos un ciclo interminable donde la retroalimentación es constante y es el mecanismo por el cual logramos avanzar.

“Todo sale bien al final y, si no sale bien, es que no es el final”. Anónimo.

3.1 Recuperación de información

El crecimiento del uso de Internet ha puesto a disposición de los usuarios grandes cantidades de información, sin embargo, encontrar información relevante resulta en algunos casos una labor complicada. De hecho, este concepto de relevancia puede verse reflejado cuando Grossman y Frieder anotan que recuperar información es:

“encontrar documentos relevantes, no encontrar simples correspondencias a unos patrones de bits” [22].

No se trata únicamente de localizar datos, sino más bien de generar información pertinente que con normalidad requiere de procesos dedicados al concepto de *selectividad*, a la aplicación de sucesivas discriminaciones u operaciones hasta recoger la esencia de lo que se busca.

Los sistemas de recuperación de información (IR) tratan de extraer información estructurada de contenido desestructurado siguiendo dicho concepto de *selectividad*. Así, es posible realizar preguntas concretas sobre documentos y obtener respuestas precisas. A grandes rasgos, es una manera de codificar la información o indexarla de forma que se pueda interactuar con ella.

Dichos sistemas se aplican en diferentes áreas y, dependiendo de la finalidad, su creación y procesamiento serán bien distintos. De hecho, estos sistemas requieren como primer paso definir las necesidades de información. Estas necesidades, junto con la definición del formato y soporte donde se encuentra almacenada la

información, repercutirán de forma directa en el diseño y capacidades del sistema. Los motores de búsqueda web indexan el contenido y a través de ecuaciones complejas y tratamientos de refinamiento de consultas acceden a él gracias a las *keywords* almacenadas. Otros sistemas de IR buscan respuestas concretas a preguntas realizadas en lenguaje natural. Esta modalidad se denomina búsqueda de respuestas o Question-Answering y debe localizar de entre todos los documentos, no sólo qué documento puede estar relacionado con la pregunta, sino qué trozo en particular puede contestar con más precisión.

La disciplina de *Sentiment Analysis*, como ya he comentado con anterioridad, trata de extraer del contenido generado por el usuario una conclusión sobre su opinión acerca de un tema. Esta conclusión se traduce en una polaridad. Por lo tanto, es otro tipo de sistema de Recuperación de Información. Pero aquí, el objetivo no es que nos remitan documentos ni extractos de texto, sino una respuesta que no se halla como tal en el contenido, que debe ser procesada y finalmente decidida gracias al conjunto de tareas que comprenden el sistema.

3.2 Medidas de evaluación

Como todo campo científico, la Minería de Opiniones o el Análisis de Sentimientos requiere de unas valoraciones “objetivas” que aporten confianza sobre la calidad de los resultados.

Realmente, un sistema de RI podría ser evaluado desde diversas perspectivas dependiendo desde qué parte del proceso se quiere observar. Sería factible que un usuario deseara conocer qué tiempos de respuesta tiene dicho sistema, o que un técnico se viera más atraído por conocer qué enfoque y qué velocidad permite el almacenamiento.

Pero normalmente, aunque la eficiencia de los sistemas de IR preocupa a las comunidades investigadoras, éstas suelen centrarse más en el carácter inteligente y en la capacidad de responder con exactitud. Es por ello por lo que nosotros, al igual que en las evaluaciones que se llevan a cabo en las conferencias TREC, CLEF, TASS, y otras, orientaremos las medidas hacia la relevancia.

Para evaluar el sistema en base a este tipo de medidas es necesario identificar inequívocamente las necesidades de información a las que la consulta hace mención. El problema de estos sistemas es que a diferencia de la recuperación de datos donde el criterio de evaluación puede establecerse en base a la exactitud de lo recuperado, el éxito de la búsqueda en IR es el grado con el que la respuesta satisface dichas

necesidades, es decir, de la percepción sobre lo útil que resulta la respuesta. En definitiva, conlleva una subjetividad inherente. En el campo en que ahora nos movemos, esta satisfacción es incluso si cabe más subjetiva ya que la polaridad puede diferir entre distintos usuarios/investigadores. De hecho, al manejar un lenguaje más coloquial y dado que el contexto siempre interfiere en la interpretación, se producirán muchos casos en los que el lector reproduzca ironía (incluso siendo inexistente) en su lectura o que incluso su estado de ánimo afecte a su juicio.

La disciplina de *Sentiment Analysis*, trata de extraer de contenido generado por el usuario, una conclusión sobre su opinión acerca de un tema. Esta conclusión se traduce en una polaridad. Pero aquí, el objetivo no es que nos remitan documentos ni extractos de texto, sino una respuesta que no se halla como tal en el contenido, que debe ser procesada y finalmente extraída gracias al conjunto de tareas que comprenden el sistema.

En lo que sigue se presentarán las medidas de evaluación más utilizadas y las más acordes a la tarea de *Sentiment Analysis*.

3.2.1 Medidas de precisión y recall

Los resultados obtenidos por un sistema de IR deberán ser enfrentados a los juicios de relevancia durante el proceso de evaluación. Dado que el objetivo es determinar si la polaridad obtenida es la correcta o no para cada tweet, los juicios de relevancia contendrán el conjunto de mensajes que deberían ser clasificados como pertenecientes a cada una de las polaridades posibles.

Las medidas empleadas para calcular la bondad del modelo son las empleadas en los sistemas de IR (medidas de *precisión* y *recall* o *exhaustividad*). Estas medidas se pueden explicar con la siguiente imagen:

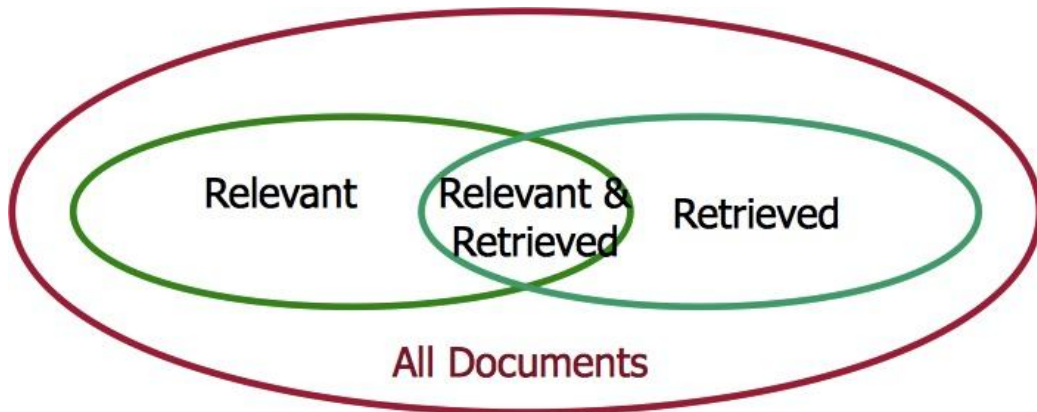


Figura 1: Conceptos principales para la definición de medidas.

En este contexto, se denomina *silencio* a los documentos relevantes no recuperados y *ruido* a aquellos recuperados no relevantes.

La *precisión* es el subconjunto de los documentos que son relevantes dentro del conjunto de todos los obtenidos ante una consulta o *query*. La *precisión* mide, pues, la calidad del conjunto de documentos recuperados. De forma genérica se representa por la ecuación:

$$\text{Precisión} = \frac{\text{Documentos relevantes recuperados}}{\text{Número de documentos recuperados}}$$

En este universo de *Sentiment Analysis*, la *query* se corresponde con la polaridad que quieres analizar y el concepto de documento relevante se equipara al de tweet correctamente clasificado para esa polaridad. Si extrapolamos la definición anterior al universo que ahora nos ocupa, obtenemos la siguiente ecuación:

$$\text{Precisión} = \frac{\text{Clasificaciones correctas para la polaridad "x"}}{\text{Total de tweets recuperados para dicha polaridad}}$$

El *recall* o *exhaustividad* en IR mide la eficacia de la recuperación y constituye la fracción de documentos relevantes obtenidos dentro del conjunto de todos los documentos que son realmente relevantes.

Exhaustividad = Documentos relevantes recuperados / Número de documentos relevantes

Para este ámbito, redefinimos la fórmula de esta manera:

Exhaustividad = Clasificaciones correctas para la polaridad "x" / Total de tweets con dicha polaridad

La *precisión* y la *exhaustividad* tienden a poseer una relación inversa donde cuanto más crece la primera más decrece la segunda.

También se pueden evaluar los resultados unificando las medidas anteriores con la media armónica llamada medida F.

Medida F = $((B^2 + 1) * P * E) / (B^2 * P + E)$

Donde B es un valor preestablecido, teniendo en cuenta que si B es igual a uno, estamos dando la misma importancia a la precisión que a la exhaustividad, si B es mayor que uno de damos más importancia a la exhaustividad y si es menor de 1 justo lo contrario.

Un contexto de evaluación es lo que ofrece TREC (Text Retrieval Conference) y su herramienta TREC-EVAL. TREC es uno de los foros de más prestigio dentro de la comunidad científica enfocada a sistemas de recuperación de información. Cada año reúne a multitud de investigadores que, gracias a los bancos de pruebas y la herramienta TREC-EVAL han evaluado sus correspondientes sistemas y desean comentar y comparar sus resultados con el fin de hacer que la comunidad aprenda y avance en conjunto.

TREC_EVAL es el nombre de la herramienta usada para la evaluación sobre grandes colecciones de datos (corpus) en los que ejecutar una serie de pruebas previamente estandarizadas y acordadas por el foro. La finalidad de estas estandarizaciones y conjuntos de pruebas es lidiar con las evaluaciones que de forma particular se practicaban sobre pequeños conjuntos de documentos o sobre documentos de mínima variedad temática y cuya extrapolación a grandes colecciones era dificultosa.

¿Y cómo se evalúan esas pruebas? Para este cometido existen los llamados juicios de relevancia que ya mencionamos anteriormente, que sirven para poder puntuar los resultados. Los juicios de relevancia son llevados a cabo por jueces o asesores (personas miembros de universidades, laboratorios y empresas) e indican la relevancia/no relevancia de cada documento de la colección una vez ejecutada una query. Atendiendo a la subjetividad y a la percepción personal, es obvio entender que los juicios emitidos por los jueces pueden no concordar y se hace necesaria la valoración de sus juicios en conjunto.

3.2.2 Otras medidas

Además de las mencionadas del ámbito de IR, existen otras medidas que mantienen similitudes aunque se aplican en otros campos.

De hecho, estadísticamente, la medida de *recall* se corresponde al concepto de *sensibilidad* que se aplica típicamente en los diagnósticos. Vamos a hacer un breve recorrido sobre esta teoría estadística en nuestro contexto pero antes definamos los conceptos:

		Resultado de la prueba	
		Documentos recuperados	Documentos no recuperados
Veracidad	Documentos relevantes	VP	FN
	Documentos no relevantes	FP	VN

Tabla 1. Definición de conceptos para el estudio de la sensibilidad.

Donde **VP**, serían los verdaderos positivos; **FN**, los falsos negativos; **VN**, serían los verdaderos negativos y **FP**, los falsos positivos.

La *sensibilidad* es la medida extraída de esta fórmula (donde podremos observar su similitud con la medida *recall*):

Sensibilidad = $VP / (VP + FN)$ → Este denominador es lo que nosotros en nuestro campo reconocemos como "Número de documentos relevantes".

Acompañando a esta medida, nos podemos también topar con la *especificidad*, que se codifica de la siguiente forma:

Especificidad = $VN / (VN + FP)$

Es por lo tanto, la fracción de verdaderos negativos hallados correctamente.

Se presenta a continuación otras medidas que también han sido objeto de debate, dado que la relevancia o no relevancia es una característica difícilmente cotejable objetivamente y que depende de la percepción del que la maneja. En el campo que nos concierne, muchas de estas medidas son difícilmente aplicables, pero de todas formas merece la pena citarlas brevemente para entender sobre todo que no hay demasiado consenso en el conjunto de evaluadores.

Por un lado, se ha pretendido definir juicios de relevancia llevados a cabo por el usuario. Es esta figura finalmente la que puede otorgar utilidad a la respuesta del sistema ante una pregunta que él mismo haya realizado. Constituye una especie de traslado desde el significado de relevancia al concepto de utilidad y a la definición de pertinencia. Se proponen otras medidas como la *cobertura* (documentos que el usuario contaba con recuperar), la *novedad* (aquellos que le han llegado por sorpresa) o la *exhaustividad relativa* (ratio que define cuántos documentos ha examinado entre los que está dispuesto a examinar). Por último, no es menos importante la medida de *esfuerzo de exhaustividad* que hace alusión al número de documentos que ha tenido que pasar el usuario hasta encontrar la información útil que esperaba hallar.

Por otro lado, se ha hecho hincapié en la intromisión de la medida de *tasa de fallo* que recoge el porcentaje de documentos recuperados no relevantes de entre el conjunto de documentos no relevantes (puede comprobarse la similitud con la *especificidad*). Esta medida es especialmente importante si se considera que la precisión se encuentra muy sujeta a posibles variaciones en el contenido de la base de datos/colección, y se observa que la *tasa de fallo* no adolece tanto de esta dependencia.

Desde la perspectiva de la eficiencia del sistema, existen también medidas orientadas a evaluar el acceso a la información y a las prestaciones tecnológicas que confiere. Entre ellas se encuentran los tiempos de acceso (medidas de rendimiento) o el número de instrucciones necesarias para computar una consulta.

Por último, mencionar que además de las comúnmente usadas, otros autores proponen emplear otras nuevas llamadas de *valor simple* porque combinan las anteriores de diversos modos (según la teoría del autor) pero consiguiendo todas ellas un único valor ordinal. Aquí agrego un cuadro de las más representativas:

Autor	Expresión
Borko	$I = E + P$
Meadow	$M = 1 - \frac{\sqrt{(1-P^2) + (1-R^2)}}{\sqrt{2}}$
Heine	$D_1 = 1 - \frac{1}{\frac{1}{P} + \frac{1}{R} - 1}$
Vickery	$V = 1 - \frac{1}{\frac{2}{P} + \frac{2}{R} - 3}$

Figura 2: Medidas de valor simple.

3.3 TASS: corpus de prueba

Para realizar los trabajos que se describen en esta memoria, el primer paso es encontrar un corpus que contenga mensajes de Twitter y otros recursos necesarios (juicios de relevancia, por ejemplo). Entre los existentes, el corpus TASS contiene más de 70.000 tweets escritos en español en la época comprendida entre Noviembre del 2011 y Marzo del 2012. Todos ellos han sido escritos por periodistas, políticos y famosos.

En este corpus, cada mensaje contiene un identificador, la fecha de creación y otro identificador correspondiente al usuario. Con estos códigos inequívocos es posible recuperar el contenido del mensaje e información de los usuarios.

Además de estos datos básicos, cada tweet es anotado con su polaridad global que comprende 5 niveles [Muy Positivo (**P+**), Positivo (**P**), Negativo (**N**), Muy negativo (**N+**), Neutral (**NEU**)] junto con uno adicional que especifica que el mensaje no contiene sentimiento alguno. En el caso de que existan entidades, las polaridades pueden afectar a dicha entidad en particular, en vez de a todo el mensaje, lo que significa que es factible que un tweet tenga más de una polaridad asociada.

En el mismo contexto de la polaridad, se añade el nivel de acuerdo/desacuerdo que hace referencia a si la polaridad global es constante a lo largo de todo el texto o si es fruto de una conjunción de diferentes polaridades encontradas y por tanto, existe desacuerdo en el escrito.

Por último indicar que cada tweet es anotado según su pertenencia a diferentes áreas temáticas o *topics*: política, fútbol, literatura o entretenimiento.

Un ejemplo completo es el siguiente:

```

<twit>
  <twitid>000000001</twitid>
  <user>usuario1</user>
  <content><![CDATA["UPyD contará casi seguro con grupo gracias al Foro Asturias.]]></content>
  <date>2011-12-02T00:21:01</date>
  <lang>es</lang>
  <sentiments>
    <polarity>
      <value>P</value>
      <type>AGREEMENT</type>
    </polarity>
    <polarity>
      <entity>UPyD</entity>
      <value>P</value>
      <type>AGREEMENT</type>
    </polarity>
    <polarity>
      <entity>Foro_Asturias</entity>
      <value>P</value>
      <type>AGREEMENT</type>
    </polarity>
  </sentiments>
  <topics>
    <topic>política</topic>
  </topics>
</twit>

```

El corpus se divide en dos conjuntos: el de entrenamiento y el de evaluación. El primero, a diferencia del segundo, contiene mensajes acompañados de su polaridad para ayudarnos a crear el modelo mientras que el de evaluación nos servirá para ponerlo a prueba y evaluar el sistema. Aquí es donde entran los juicios de relevancia y el cálculo de las medidas más extendidas ya comentadas.

4. Planteamiento

Después de haber realizado un recorrido suficientemente representativo en el estado del arte, se presenta a continuación la motivación que nos lleva a plantear el trabajo finalmente desarrollado.

La tarea de Opinión Mining está cobrando mucha relevancia en estos últimos años. Ser capaces de comprender e interpretar de forma automática lo que los usuarios pretenden expresar en sus comentarios proporciona a los analistas, propietarios de empresas o investigadores información crítica acerca de cómo ven las personas el mundo que les rodea o ciertos aspectos concretos acerca de un tema en particular.

Además, el uso de las redes sociales se está extendiendo considerablemente como mecanismo y plataforma para el debate y la crítica por parte de sus usuarios. El conjunto de ambos contextos ha generado el interés masivo por parte de los investigadores para mejorar las técnicas de Minería de Datos y adecuarlas al mundo de la opinión.

Pero no sólo ha acaparado atención en el ámbito de la Minería de Datos, sino también en el más cercano a la lingüística que intenta demostrar que el Procesamiento del Lenguaje Natural es una asignatura troncal para determinar los sentimientos subyacentes. En el enfoque lingüístico, el procesamiento se realiza normalmente en línea, por fases, donde la mayoría de los módulos trabajan a distintos niveles la lingüística y la semántica para conseguir una clasificación basada en la comunicación y en el lenguaje como tal. Es un enfoque dirigido a la comprensión del lenguaje, en la línea de la lingüística computacional. Así, analizan pistas estudiadas en los textos que se consideran claves para poder determinar qué tipo de opinión se encuentra tras las palabras.

Es en este marco se incluye este trabajo que, es una aproximación léxico-semántica para la extracción de la polaridad en los tweets. ¿Y por qué en los tweets? La razón es sencilla aunque compuesta: por un lado, porque las redes sociales actualmente sirven de magnífico soporte para conocer la realidad *de la "calle"* y por otro lado, porque simplemente era posible, es decir, la disponibilidad de un corpus como TASS a partir del cual aprender, permitía trasladar lo aprendido al sistema y después valorar con un elevado grado de confianza. Ambas cosas formaban una posibilidad muy atractiva.

Por lo tanto, el objetivo final es la investigación, el desarrollo y la evaluación de un sistema capaz de determinar la polaridad de los tweets sobre la base de un conjunto de características observadas desde el punto de vista de la lingüística. Este proceso

en fases, se aprovechará también para la generación de un corpus anotado que pueda ser útil para futuros clasificadores.

Metodología y evaluación:

1. Estudio del estado del arte.
 - i. Obtención de pautas dadas y recolección de buenos resultados en el pasado.
2. Análisis del dominio:
 - i. Realizar una exploración en busca de los rasgos o características que se desean considerar a la hora de analizar los tweets. Implica inicialmente un análisis del dominio de la plataforma en particular y del lenguaje de las redes sociales en general.
3. Examen del corpus:
 - i. Búsqueda y análisis de las características relevantes para la explotación del corpus y la obtención de la información y conocimiento (basada en los resultados del punto anterior).
4. Generación del sistema.
 - i. Análisis de las herramientas necesarias (librerías, diccionarios,...)
 - ii. Creación de módulos de procesamiento. Cada módulo se encargará de procesar una de las características consideradas clave para definir la polaridad. Además, dará como resultado el etiquetado de la parte correspondiente del corpus.

5. Ejecución y etiquetado de polaridad. Cada tweet del corpus TASS será finalmente anotado automáticamente con la polaridad obtenida del sistema y la tarea de la creación del nuevo corpus se dará por concluida.
6. Evaluación del sistema. Gracias a los juicios de relevancia contenidos en el corpus inicial será posible fijar la precisión del sistema.
7. Desarrollo de conclusiones finales. Análisis personal sobre los puntos fuertes y débiles de acuerdo a los resultados obtenidos en el paso anterior y profundización en la exploración de futuras mejoras.

PARTE 2: TRABAJO REALIZADO Y EXPERIMENTOS

Los enfoques basados en lingüística requieren normalmente un procesamiento por fases que abarca desde las más simples tareas de preparación o limpieza de los textos hasta los más complejos procesos de comprensión formal.

Algunas de dichas tareas se encuentran más vinculadas al PLN mientras que otras son dependientes de la plataforma en la que se trabaja. Así, las primeras podrían entenderse como estándares estables en este ámbito y las segundas como tareas cambiantes dada su estrecha vinculación al entorno. Para trasladar todas ellas a la realidad operacional se ha hecho uso de módulos. Aquí vemos *grosso modo* la arquitectura general del sistema.

Arquitectura general

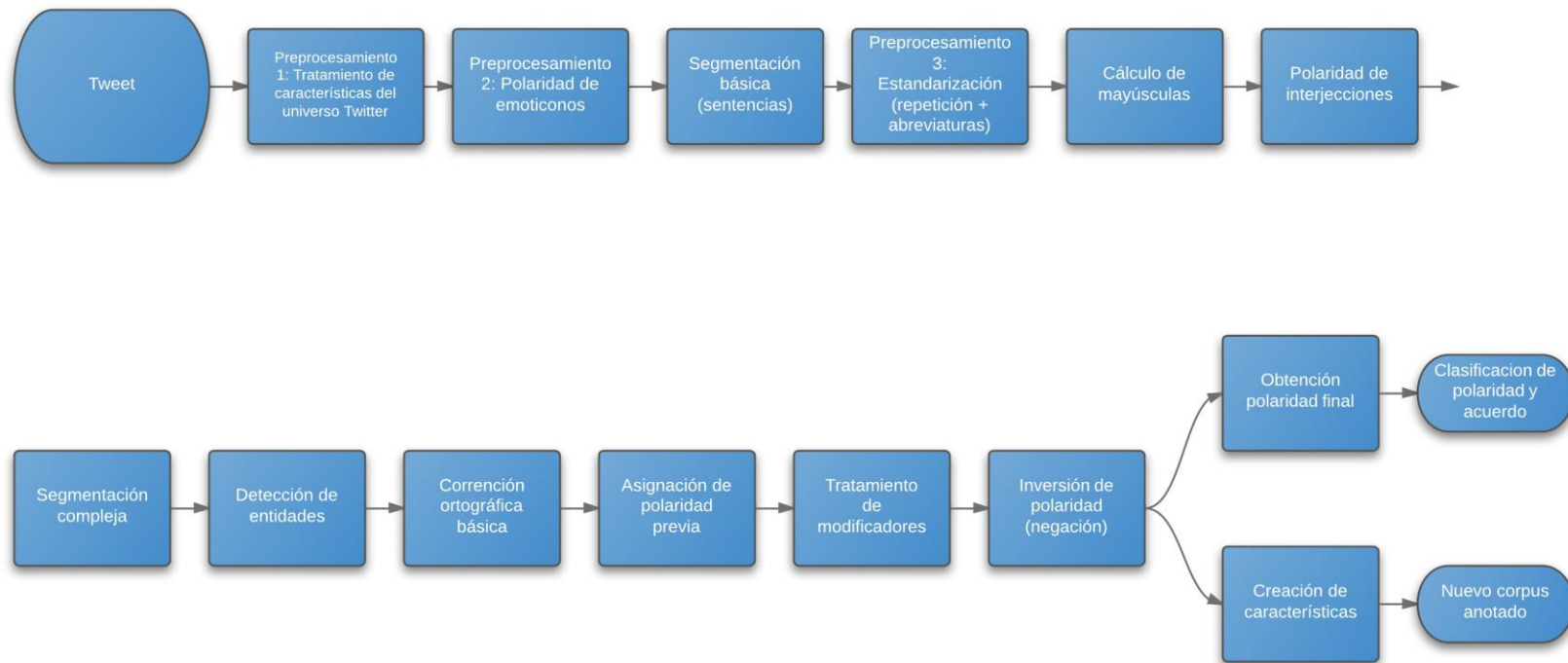


Figura 3: Arquitectura general.

Cada módulo se ha preparado por separado, lo que implica beneficiarse de la programación estructurada: legibilidad, manejabilidad y versatilidad para una futura traslación al tratamiento de problemas similares. Sin embargo, también es esencial concienciarse de que ciertos procedimientos constituyen la base sobre la cual se asientan otros y que la eliminación de algunos, tales como las tareas de preprocesamiento o las de segmentación, puede influir en el rendimiento y eficacia del sistema.

5. Funcionalidades básicas

En este apartado, uno a uno se presenta cada módulo y lo que su aplicación constituye y aporta al funcionamiento general en este trabajo.

5.1 Preprocesamiento 1: Tratamiento de características del universo Twitter



Técnica a aplicar:
Expresiones regulares.

Figura 4: Preprocesamiento 1.

Este es uno de los módulos más relacionado con las peculiaridades de Twitter. Su objetivo es extraer las URL"s, los hashtag y las menciones, dado que en siguientes procesos, sobre todo los dos últimos, constituyen motivo de error dada su inexistencia en el lenguaje natural. Cuando se hace uso de herramientas de PLN es crucial tener en consideración este tipo de excepciones del lenguaje ya que pueden resultar muy perjudiciales para las herramientas lingüísticas y arruinar los resultados.

Para este proceso se emplean expresiones regulares que permiten lidiar con cualquier tipo de instancia sea cual sea su estructura (siempre que cumplan el estándar oficial conveniente).

5.2 Preprocesamiento 2: Extracción de polaridad de emoticonos

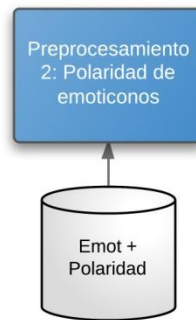


Figura 5: Preprocesamiento 2.

Otro elemento que hace ilegibles los mensajes si nos basamos en la gramática real y no en la ajustada, que es la que se usa en las redes sociales actualmente, son los emoticonos. Si nos fijamos detenidamente en ellos (en su forma escrita, no gráfica) nos encontramos signos de puntuación que se combinan pero que independientemente han perdido su utilidad real. Así, el punto y coma no marca una pausa sino que simbolizan un guiño. Hasta que se estandaricen estos usos, forman parte de una jerga que no se contempla en las herramientas de PLN y ante la cuál presentan serias dificultades. No obstante, a su vez, estos símbolos juegan un papel, a veces incluso determinante, en la detección de la polaridad de un tweet ya que portan mucha semántica y a veces pueden ser el único síntoma evidente de las sensaciones del autor.

Para este módulo se emplea un diccionario que contiene entradas como la siguiente:

Emot	Polaridad
: -)	P

Tabla 2. Ejemplo de entrada en el diccionario de emoticonos.

La polaridad puede encontrarse dentro de este dominio {N+, N, P, P+, None} aunque muchas veces, para dar una importancia más relativa a estos elementos, sólo se manejan las polaridades N, P y None.

5.3 Segmentación básica

Una de las razones de éxito en este tipo de enfoques es la granularidad de la segmentación realizada. Pero esta afirmación la argumentaré más adelante dado que el funcionamiento de este módulo es mucho más sencillo conceptualmente.

La segmentación básica es la llamada segmentación por oraciones o sentencias. Es el producto de un *tokenizador* junto con un separador de oraciones. Lo que se obtienen son *tokens* o palabras asignadas a oraciones. ¿Para qué realizamos este proceso? La explicación es bien sencilla. Los módulos que vienen a continuación crean o identifican características afiliadas a las palabras. Quizás sea el momento de comentar brevemente que la polaridad global de un tweet es una composición de las polaridades de sus oraciones, que a su vez necesitan de las polaridades de las proposiciones y estas del nivel inmediatamente inferior y así sucesivamente hasta el nivel de las palabras. Por ello, lo primero es comenzar desde lo más pequeño para ir creciendo según vaya siendo evidente el requerimiento y la capacidad de generalizar.

También podemos preguntarnos: ¿Por qué no segmentar a todos los niveles ya que estamos en ello? Sin duda, no es por una falta de intención sino un problema de imposibilidad. *Freeling*, que es la librería que emplearemos para la segmentación, no realiza por sí sola el proceso ineludible de normalización (necesario por el lenguaje especial que se usa en los tweets). Por lo tanto, se necesitan antes otros procesos.

5.4 Preprocesamiento 3: Estandarización



Figura 1: Preprocesamiento 3.

La estandarización es un proceso que incluye otras de las tareas obligadas cuando el lenguaje al que nos enfrentamos dista mucho de adecuarse a la gramática y al léxico de la lengua empleada.

Por un lado, como mecanismo de enfatización, la repetición abunda en los tweets y debe ser corregida para un posterior proceso lingüístico. Por ello, la información relativa a la repetición de caracteres queda almacenada internamente en ese *token* en forma de número de letras repetidas y de porcentaje sobre el número de letras total. Debemos recordar que el fin de este sistema no es sólo una polaridad, sino también un corpus que debe ser anotado con todas aquellas características susceptibles de ser buenas emisoras de semántica. Y la repetición es una de ellas:

1. Él lo hace fatal.
2. Él lo hace fataaaaaaaal.

Para evitar detectar como destacado un término donde la repetición es producto de un hecho fortuito, sólo se tienen en consideración aquellas repeticiones que sobrepasen los dos caracteres (por ejemplo: "quierooo") a excepción de las conjunciones "ll", "cc" y "rr" donde se necesita hallar 4 caracteres iguales para considerarlo repetición intencionada. En resumen, el hallazgo de un solo carácter repetido no será tomado en cuenta (aunque sí corregido).

Por otro lado, como resultado de la espontaneidad y dinamismo de esta forma de lenguaje coloquial, aparecen sucesivamente abreviaturas que deben ser traducidas. Para este cometido se emplea un diccionario con las abreviaturas más populares que igual que cualquiera de los diccionarios agregados admiten modificaciones, bajas o nuevas inserciones.

5.5 Cálculo de mayúsculas

Este módulo se separa del anterior porque el objetivo es diferente. En el anterior se pretendía preparar el contenido para futuros procesos de PLN (aunque además se obtuviera alguna característica). En este caso, el objetivo final es el cálculo de otra de las características que consideramos clave para el cálculo de la polaridad, pero sin realizar ninguna corrección.

El uso indebido de mayúsculas es un indicio de enfatización. De hecho, en el *protocolo* informal de internet, su aparición tiene como significado "gritar lo que uno está diciendo". Así, no tienen el mismo valor informativo las dos frases siguientes:

1. Él lo hace fatal.

2. Él lo hace FATAL. → La mayúscula enfatiza, otorga mayor contundencia.

El único inconveniente reside en la facilidad de cometer errores en los mensajes de micro-blogging donde la velocidad de la interacción entre los usuarios suele motivar la presencia de fallos accidentales. Como filtro para estos casos, se anotan sólo aquellas palabras con un porcentaje de letras en mayúsculas mayor a 60%.

Esta característica es también relevante para el corpus y futuros trabajos dado que muchas investigaciones, sobre todo con enfoques basados en Minería de Datos, reportan buenos resultados al incluirla.

5.6 Polaridad de interjecciones



Figura 6: Extracción de polaridad de interjecciones.

Al igual que ocurría con los emoticonos, las interjecciones son grandes portadoras de sentimiento. Fijémonos en el siguiente ejemplo y comparémoslo para distinguir qué aporta la interjección a la polaridad final:

1. No tenía razón. → Sin polaridad (None).
2. Porras! No tenía razón. → Polaridad negativa (N).

Por lo que podemos deducir que, de hecho, al menos algunas interjecciones comunes tienen una polaridad de difícil discusión (aunque es cierto que algunas son ambiguas). Éstas son las que forman parte del diccionario de interjecciones aunque otras sin polaridad son también añadidas como mecanismo de defensa ante posibles fallos en la ejecución de herramientas de PLN. Cuando llegue el momento de explicar el módulo encargado de los procesamientos gramaticales, estas divagaciones o aparentes inconsistencias serán defendidas.

El formato del diccionario es similar al empleado en el diccionario de emoticonos, y como ocurría en aquel, las entradas son expresiones regulares que permitirán que la búsqueda esté menos sujeta a la estaticidad de correspondencias exactas:

Interjección	Polaridad
(i)*Ánimo(!)*	P

Tabla 3. Ejemplo de entrada en el diccionario de intersecciones.

5.7 Segmentación compleja

Existen muchas cuestiones que deben resolverse cuando uno se pregunta cómo rescatar la polaridad de 140 caracteres. Una inicial podría ser: ¿qué elementos son pistas clave que recogen sentimiento? Tras esta primera cuestión, se debe abordar una segunda: ¿cómo se combinan estos elementos o cómo deben tratarse conjuntamente si se quiere llegar a la polaridad general? Si consideramos la posibilidad de que en un tweet es factible hallar diferentes polaridades, es necesario fragmentarlo de alguna forma considerando la siguiente cuestión: ¿Cuál es la mayor estructura sintáctica compuesta donde todos los sentimientos son coherentes? Nuestra respuesta: **la proposición simple**.

Gramaticalmente hablando, la proposición es la unidad de lenguaje que tiene sujeto y predicado y la menor sobre la que es posible decidir su verdad o falsedad. Es por ello, por lo que entendemos que una proposición es UNA IDEA, un sentimiento. Por lo tanto, a la hora de estudiar un texto, consideraremos el sentimiento general sobre la base de los hallados en las proposiciones y estableceremos si hay discrepancias o no en base a ese conjunto.

¿Y cuál es la estructura compuesta mínima capaz de albergar ideas? Una primera aproximación tomada en algunos trabajos fue trocear en proposiciones y simplemente quedarse con esos grandes segmentos. Supondría considerar entonces que la proposición es la unidad compuesta mínima transportadora de sentimiento, pero entonces damos con algún ejemplo como el siguiente que plantea serias dudas:

- [La chica que vive cerca es guapa] [aunque anda raro]. → 2 términos con polaridad en dos proposiciones. Este es el caso que ya quedaría recogido con el enfoque de proposiciones.
- [La chica guapa] tiene perro. → Una proposición con un sintagma nominal donde reside el único elemento con polaridad. El elemento compuesto mínimo pasa a ser el sintagma nominal.

Como ya he mencionado, algunos trabajos se limitan a realizar segmentos sobre la base de proposiciones y lo cierto es que también nosotros podríamos adaptarnos a su decisión, pero entonces otras cuestiones de complejidad operacional aparecen: en

caso de hallarnos frente a una negación ¿cuál es el ámbito de dicha negación?; si aparece un modificador (por ejemplo, 'muy'), ¿a qué términos afecta?,...

No debemos olvidar que uno de los cometidos de este trabajo es también el de crear un corpus consistente que pueda ser el principal input de futuros trabajos sobre algoritmia u otros enfoques lingüísticos. Por lo que, finalmente se decidió fragmentar en base a 3 estructuras gramaticales:

- Propositiones.
- Sintagmas nominales.
- Sintagmas preposicionales. Esta fragmentación es necesaria para la negación.

Obviamente, cada uno de los trozos puede a su vez formar parte de otro de nivel superior y crear estructuras recursivas muy complejas. El típico caso donde se puede observar esto es al encontrarnos con oraciones de relativo. Estas proposiciones van enmarcadas dentro de un sintagma nominal y delimitar éste último nos ayuda a definir el ámbito de las polaridades halladas en su interior.

Prop [SN [La chica

Prop[que odiaba a los niños]]

solía pasear con su

SN [simpática amiga]].

No es parte de este trabajo definir qué conceptos existen dentro de los mensajes y qué sentimientos van vinculados a cada uno de ellos, pero esta tarea de segmentación es una clave bastante fiable para llevar a cabo dicha labor. El corpus anotado dejará patente esta idea y permitirá futuras indagaciones en esta línea de investigación abierta.

En el anexo se ha añadido un diagrama de las clases que conforman los segmentos para poder visualizar esta compleja estructura recursiva.

Para llevar a cabo esta segmentación se ha empleado la herramienta *Freeling*² que es una librería de código abierto para el procesamiento multilingüe automático. Proporciona varios servicios de anotación, entre los que se incluyen análisis morfológico y sintáctico. Cuenta además con módulos encargados de la detección de entidades, fechas, números y *multipalabras*, además de uno encargado de la desambiguación. Todo esto conforma un paquete bastante completo que facilita las

² <http://nlp.lsi.upc.edu/freeling/>

tareas de PLN. Cabe destacar que es una herramienta en construcción permanente y que su naturaleza de código abierto permite realizar modificaciones o ampliaciones, cambios de diccionarios, etc...

Existen otras herramientas disponibles para este tipo de análisis como Stylus de Daedalus³ o CLIC⁴ del Centro de Lenguaje y Computación de Barcelona pero la mayoría no eran software libre y no permitían acceder internamente al código ni a los diccionarios. La potencia de *Freeling* no radica sólo en su destreza, sino en la posibilidad de adentrarte en su código y aprender. Asimismo, como se podrá leer más adelante, tuve ciertos problemas con la herramienta que rápidamente puse en conocimiento de la comunidad y para los que pude hallar soluciones. Una oportunidad que no tiene cabida cuando adquieres una herramienta cerrada.

5.7.1 Modificaciones sugeridas a *Freeling*

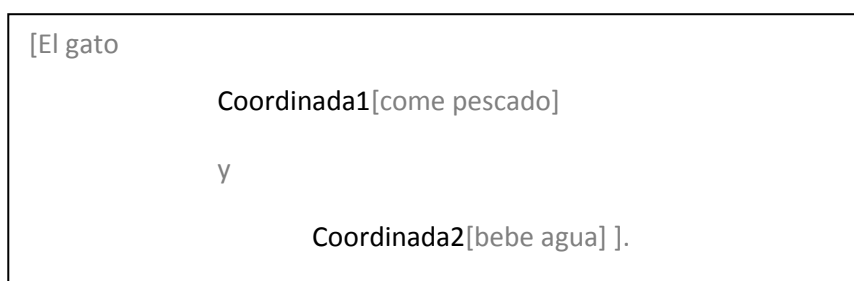
Este módulo ha sido uno de los que más tiempo y dedicación ha requerido para su desarrollo. No tanto por la complejidad de su programación, sino más bien por la dependencia mantenida con la librería *Freeling* cuyos resultados de anotación no cumplían siempre con las expectativas.

Cuando diseñas un sistema automático, una de las cualidades que debes intentar suplir es el determinismo: que ante una misma verdad, el sistema actúe de la misma forma. El inconveniente de usar desarrollos de terceros es que esto no se puede prometer.

Uno de los cometidos de mi módulo era segmentar por proposiciones, de modo que ante la aparición de una coordinada (etiqueta “*coor-vb*”), igual que con una subordinada, se creaba una nueva proposición dependiente de la que la contuviera. Veamos el comportamiento de *Freeling* en dos casos distintos:

Caso 1: El gato come pescado y bebe agua.

Sintácticamente, podríamos visualizarlo de esta forma:



³ <http://www.mystilus.com>

⁴ <http://cllc.ub.edu/>

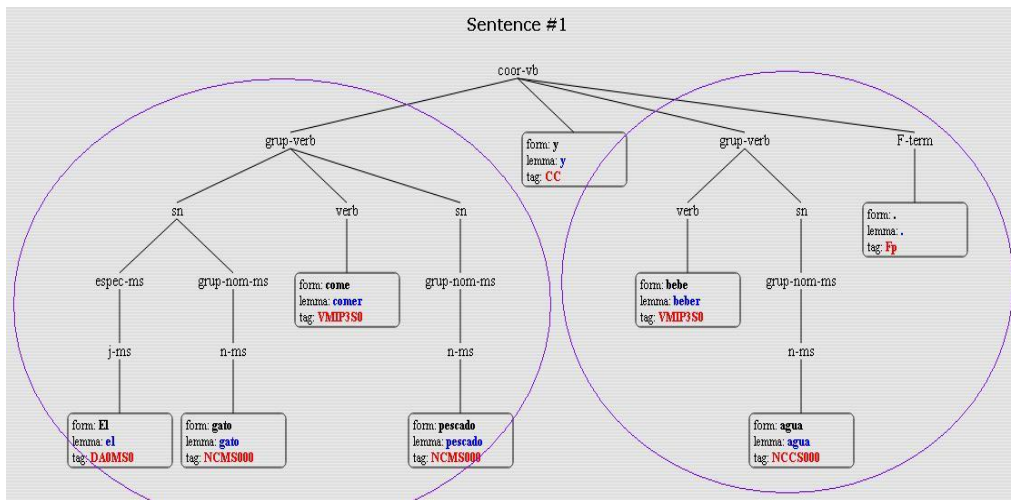
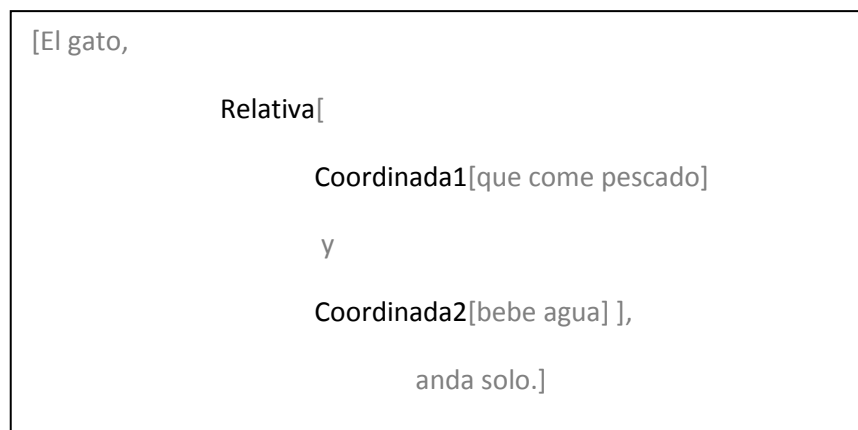


Figura 7: Árbol sintáctico para el caso 1.

Caso 2: El gato, que come pescado y bebe agua, anda solo.



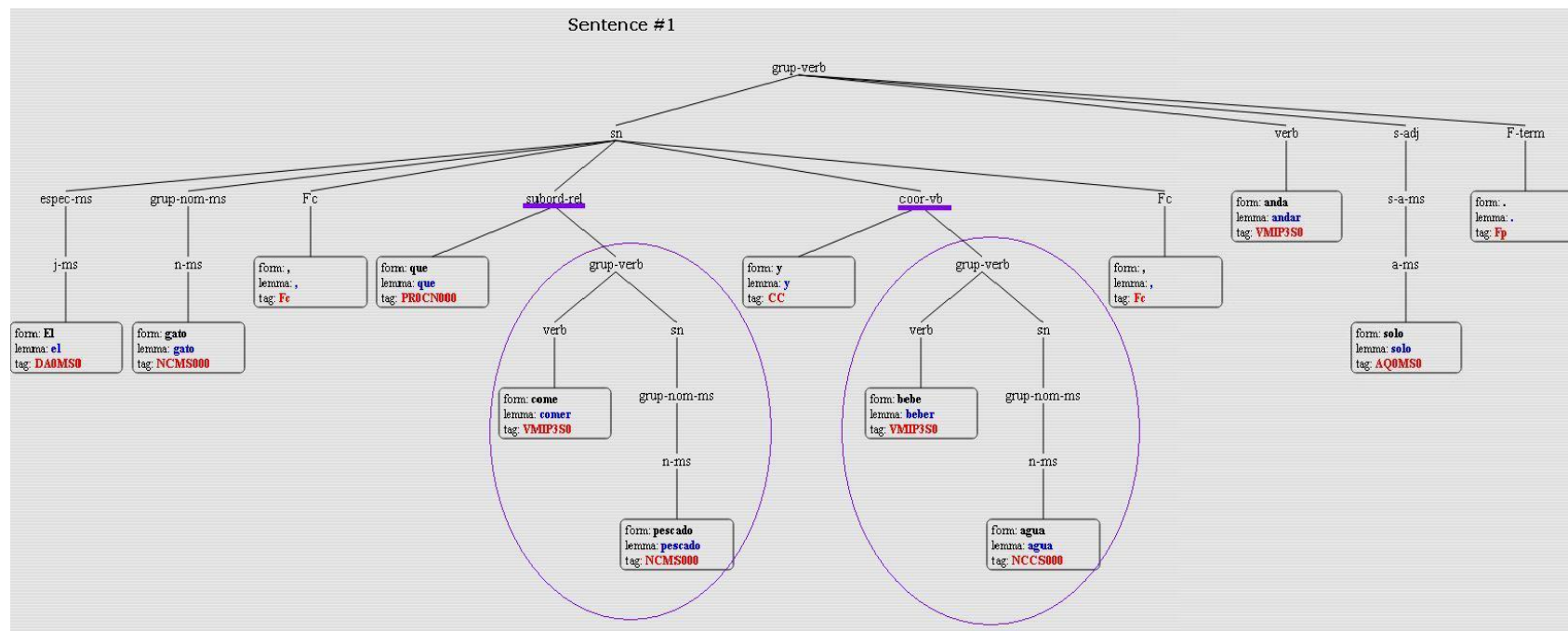


Figura 8: Árbol sintáctico para el caso 2.

En la primera imagen podemos comprobar cómo se crean correctamente las coordinadas (dos “group-verb” que cuelgan de “coor-vb”). Sin embargo, en el segundo caso, la etiqueta “coor-vb” no está incluida dentro de la subordinada. Sintácticamente, debería haber dos coordinadas debajo de la relativa. De hecho, no existe ningún caso en la lengua castellana en el que dentro de un sintagma nominal (“sn”) pueda existir una oración a menos que se trate de una relativa. En este segundo caso, *Freeling* hace una equivocada distinción entre la relativa y la coordinada (que debería ser parte de la relativa). El error proviene del archivo de dependencias de la herramienta que permite la construcción de estos árboles. Simplemente, no se contempla la posibilidad de encontrarnos una coordinada dentro de una relativa (tampoco existe esta posibilidad para las subordinadas). Todas las entradas en este archivo que contemplan las subordinadas relativas tienen este formato:

```
6001 - - (prel-fp,grup-verb) top_left RELABEL subord-rel:- % las libretas (que(habéis utilizado))
```

Sin entrar en mucho detalle sobre la sintaxis de dichas entradas, podemos observar cómo la única posibilidad que se considera como rama en la formación de la relativa es “**grup-verb**”, es decir, una proposición simple. En el caso de existir oraciones compuestas dentro de la relativa (ya anotadas en un proceso anterior como “coor-vb”, “subord-vb”,...), según este archivo de dependencias, se quedarían fuera de esa rama de proposición relativa. Esto constituía un grave problema para mí, ya que de repente aparecía una coordinada suelta añadida a un sintagma nominal y una gran parte de mi análisis posterior no era válido.

Otra cuestión que llamó mi atención fue el análisis para los sintagmas preposicionales. Una vez que descubrí la posibilidad de imbuir una negación en estas estructuras, me propuse anotarlas y cuál fue mi sorpresa al detectar que no *Freeling* no era capaz de abordar eficientemente los sintagmas que contenían verbos en forma de infinitivo. Veamos visualmente el problema:

Caso 1: [En vez de guapa], ella es bastante fea.

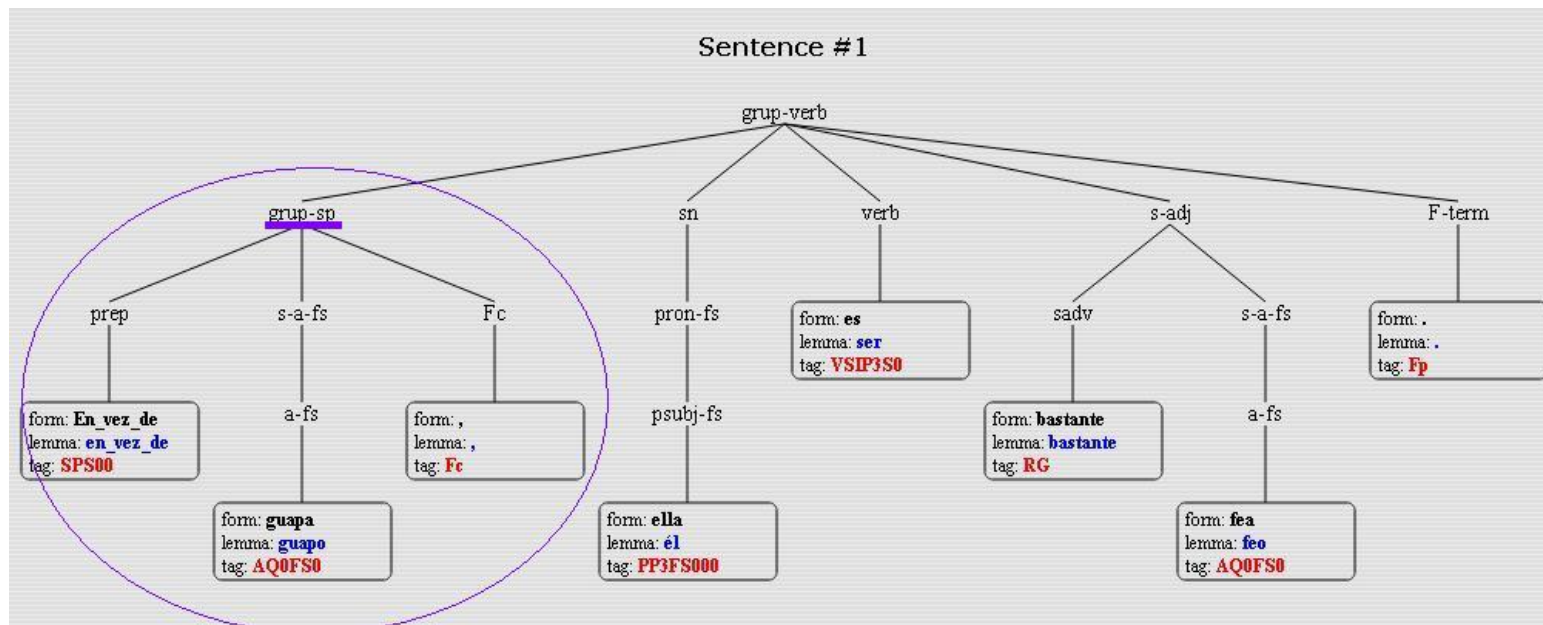


Figura 9: Árbol sintáctico para el caso 1.

Caso 2: En vez de cantar bien las canciones, lo hace mal.

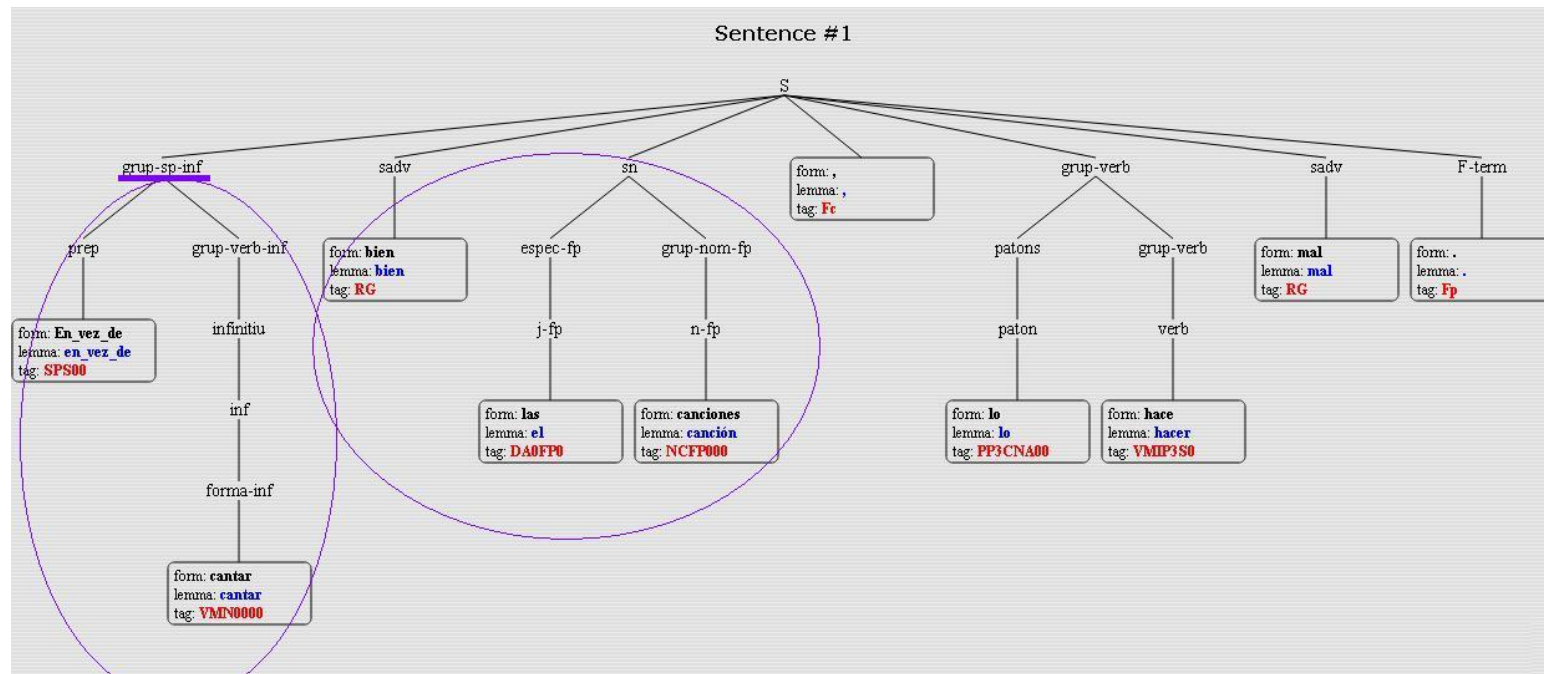


Figura 10: Árbol sintáctico para el caso 2.

El problema reside en las entradas destinadas a definir el sintagma preposicional con infinitivo en el archivo que define la gramática. Todas ellas tienen una composición similar a la siguiente:

`grup-sp-inf ==> +prep, grup-verb-inf`

donde la definición de `group-verb-inf` (más abajo en el mismo archivo) es:

`grup-verb-inf ==> +infinitiu.`

Y `infinitiu`, como su nombre señala, es un verbo en su forma impersonal de infinitivo.

Entonces, ¿qué ocurre si el sintagma contiene más elementos? La respuesta se encuentra en lo que hemos podido observar. Simplemente, queda separada del sintagma.

En la última versión de *Freeling* (3.0) siguen realizándose cambios y la gramática cambia a un ritmo vertiginoso. El archivo que utiliza el *chunker*, que es el módulo encargado de comenzar a crear el árbol uniendo los trozos más básicos es el que menos modificaciones sufre. En cambio, el archivo de dependencias, que reúne estos trozos ya creados para finalizar el árbol en constituyentes de mayor nivel, sí que ha sido alterado en muchas más ocasiones. Este archivo de dependencias está basado en reglas heurísticas y por esa razón, las dependencias de larga distancia son difíciles de codificar. Si se quieren seguir haciendo pruebas en la demo de la página oficial se podrá comprobar como el caso que he comentado, dependiendo de en qué posición coloques el sintagma dentro de la oración, a veces acierta y otras no. Esto es porque la gramática de dependencias a veces sí que es capaz de juntar los trozos del *chunker*, y otras no.

La solución que yo planteé es modificar el archivo del *chunker*, permitiendo que el sintagma preposicional pudiera estar formado por más constituyentes. Esto permitió que la gramática de dependencias se librara de tener que lidiar con estos asuntos y convirtió al sistema en un poco más determinista.

Un ejemplo de modificación (existen más dado que para cada tipo de preposición existe una entrada de este tipo) es el que añado a continuación:

`grup-sp-inf ==> +prep, grup-verb-inf | +prep, grup-verb-inf, sn | +prep, grup-verb-inf, sadv | +prep, grup-verb-inf, s-adj.`

La última opción pertenece a los casos de predicados nominales con verbos copulativos (por ello lleva un constituyente cuyo núcleo es un adjetivo detrás del sintagma verbal).

Otra modificación forzosa que tuve que realizar tuvo lugar al encontrarnos este error:

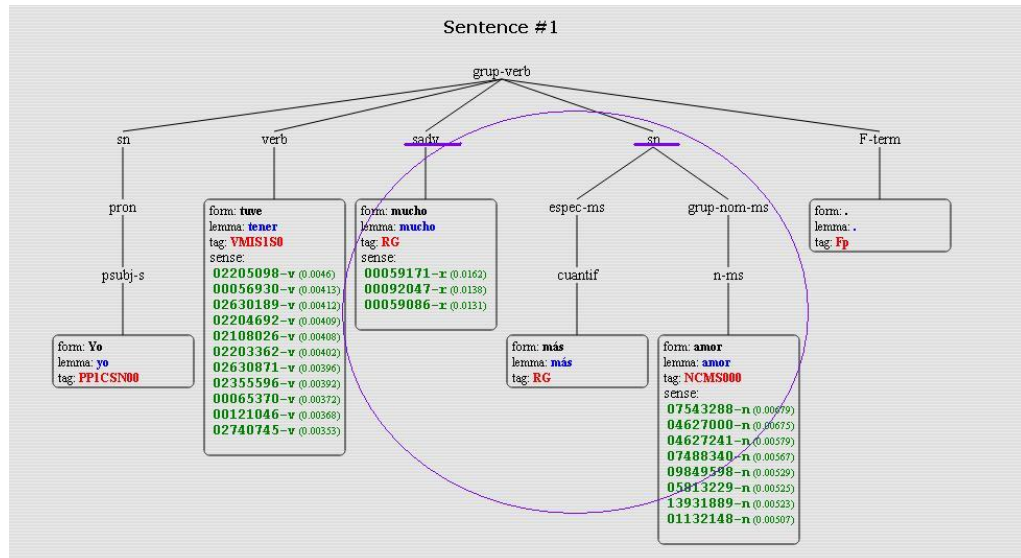


Figura 11: Árbol sintáctico para el caso 3.

La palabra “más” es un determinante indefinido, pero el *tagger* sólo es capaz de etiquetarla con la categoría de adverbio.

Para empezar, el conjunto de etiquetas que utiliza *Freeling* es EAGLES y podemos observar en su descripción⁵, en el apartado de determinantes indefinidos, como no existe entrada para “más”. Es un error muy común en la lengua castellana considerar este término siempre como adverbio, pero también un error poco racional.

En el ejemplo que he expuesto queda patente de forma clara que dicha palabra acompaña al nombre (de hecho, *Freeling* dentro su árbol sintáctico también lo estima así). De todas formas, ante posibles dudas es muy útil revisarse el artículo ⁶del filólogo Manuel LH donde además añade un ejemplo similar como uso de determinante:

“Se dice que son determinantes cuando acompañan al nombre o sustantivo y toman de él su género y su número: había muchas personas, vi pocos modelos, estuvieron demasiadas veces, compraron bastantes libros, limpiaron toda la casa, trajeron más botellas, dejaron menos cartas. Sólo cuando acompañan al nombre se puede decir que son determinantes.”

⁵ <http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html>

⁶ <http://suite101.net/article/determinantes-y-pronombres-indefinidos-o-adverbios-de-cantidad-a67708#axzz2JMAFlrQw>

Indagando en el proceso de la librería para el etiquetado gramatical llegamos a un archivo denominado “dicc.src” que sirve para asignar posibles etiquetas a los términos que más tarde serán desambiguados por el *tagger* (que como elección personal configuré para seleccionar el etiquetador basado en modelos ocultos de Markov HMM). Si buscamos la línea que contiene la palabra “más” nos encontramos con lo siguiente:

más más RG

Donde la segunda palabra se refiere a la raíz gramatical y la tercera a las etiquetas posibles. Por lo tanto, podemos deducir que el *tagger* nunca tendrá opción de asignar otra categoría a dicho término. Como primera intuición, pensé que según la teoría de Markov aplicada a la gramática castellana, existiría mayor probabilidad de encontrarse un bigrama compuesto por un determinante más un sustantivo que por un adverbio más un sustantivo, con lo que si conseguía modificar el diccionario y añadirle la posibilidad de que “más” fuera un determinante indefinido, seguramente el *tagger* se decantaría por la opción de determinante frente a la de adverbio al hallar una estructura como la del ejemplo.

El cambio realizado dejó la entrada en el diccionario de esta manera:

más más RG **más DIOCSO** (un cambio similar se aplicó a la palabra “menos”)

Mi intuición fue acertada y finalmente se consiguió que el etiquetado funcionara correctamente.

La única cuestión que quedaba sin resolver era que el *chunker* (el responsable de ir juntando trozos y formar la base del árbol sintáctico) no añadía el adverbio “mucho” al sintagma permitiendo que el sintagma quedara así:

Sintagma Nominal [mucho más amor]

Para remediar esta situación, se volvió a retocar la gramática incluyendo las nuevas posibilidades.

La definición del sintagma nominal es la siguiente:

sn ==> espec-ms, +grup-nom-ms | espec-mp, +grup-nom-mp
| espec-fs, +grup-nom-fs | espec-fp, +grup-nom-fp.

Donde parece que “espec-XX” (donde se incluyen los posibles acompañantes del nombre) debería ser la etiqueta a modificar. En nuestro caso, más internamente, la estructura se correspondía con un elemento compuesto denominado en un nivel inferior “grup-complex-spec”:

espec-XX ==> **grup-complex-spec-XX** | cuantif | num-XX
| dem-XX | pos-XX | int-XX | exc-XX
| indef-XX | num-XX | j-XX.

Esta categoría contiene numerosas entradas dada la variedad posible. Una de ellas se asemejaba mucho gramaticalmente a nuestra situación:

grup-complex-spec-fs ==> RG(muy), +indef-fs.

Finalmente se decidió añadir una nueva línea que contemplaba la estructura requerida:

grup-complex-spec-ms ==> RG(mucho), +indef-ms. %mucho más, mucho menos,...

Con este cambio, se alcanzó el propósito y el *chunker* unió el adverbio dentro del sintagma.

5.8 Detección de entidades

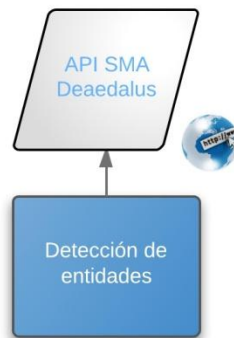


Figura 12: Detección de entidades.

La detección de entidades es uno de los procesos a ejecutar que más consenso tienen en la comunidad científica que aborda estos temas. Para este trabajo se barajó la posibilidad de trabajar con la librería de *Freeling* que también contenía un

módulo dedicado a esta tarea. Finalmente, y dadas las mejoras obtenidas, se decantó por el API SMA (Social Media Analytics) de Daedalus⁷.

Daedalus propone un API que permite mediante servicio web obtener un XML con la extracción realizada de entidades tales como organizaciones, personas, lugares, expresiones de tiempo,... Esta distinción nos permitirá anotar el corpus con dichas entidades, además de evitar que estos elementos sean procesados por los módulos siguientes. Esta distinción evita que si la entidad pudiera ser confundida con un término común (por ejemplo: BELLEZA S.A.) llegara a ser etiquetada erróneamente con polaridades previas, uso de mayúsculas o cualquier otro módulo que detectara cualquier característica con la que la palabra mantuviera correspondencia.

Al igual que ya comenté en el apartado anterior para los conceptos, la detección de entidades nombradas es un proceso clave esencial cuando se quieren asignar polaridades a elementos del texto. Además, facilita enormemente el descubrimiento de temas o *topics* en los tweets, que es otro de los asuntos fundamentales en los que los investigadores están aunando esfuerzos.

5.9 Corrección ortográfica

Existen trabajos que usan el algoritmo de la distancia mínima de edición contra un diccionario para corregir posibles palabras erróneamente escritas [1]. Este es un punto en el que nos hubiera gustado ahondar, pero por el momento nos hemos conformado con una simple corrección ortográfica.

Pudiera parecer curiosa la localización de este módulo, pero esta decisión de diseño se corresponde con la situación del módulo anterior. Es posible encontrarnos con muchos mensajes donde el uso de mayúsculas no es apropiado según las reglas que rigen la lengua española, pero tampoco era una opción acertada trasladar toda mayúscula a minúscula ya que las entidades nombradas tienen derecho a emplearla. Es por esto por lo que es necesario esperar al procesamiento anterior. Una vez que se detectan todas las entidades y que se ha anotado el uso de mayúsculas como síntoma de enfatización, podemos pasar a corregir las palabras indebidamente escritas.

⁷ <http://api.daedalus.es/sma-info>

5.10 Asignación de polaridad previa



Figura 13: Asignación de polaridad previa.

La polaridad previa es una característica íntimamente vinculada a palabras con un alto grado de subjetividad. Obviamente, la polaridad de las palabras no determina de forma absoluta la polaridad de un texto, sobre todo si tenemos en cuenta que el contexto es un factor concluyente y que las palabras observadas de forma aislada deben ser relegadas a un segundo plano.

¿De qué elementos se puede inferir sentimiento en un mensaje? La respuesta es compleja e interfieren muchos factores. Inicialmente, obviamente, las palabras aportan su parte, pero existen otras variables internas y externas que pueden desde alterar conceptualmente los significados hasta invertir completamente las polaridades.

Uno de estos factores nace en los individuos. Cuando se establece una comunicación, indistintamente hablada o escrita, las experiencias de los hablantes y los oyentes además de sus estados de ánimo juegan un papel muy relevante. Sin embargo, se puede hacer caso omiso a estas conjeturas aunque en lo que al escritor se refiere, sí deberíamos de hecho tener en cuenta tanta información como fuera posible. Es necesario mencionarlo para advertir que la polaridad de las palabras no son valores absolutos.

*SentiWordnet*⁸ es una herramienta creada a partir de clasificadores entrenados con métodos semi-supervisados con el fin de asignar una polaridad previa, no a las palabras como formas léxicas, sino a las palabras como conceptos. Esta discriminación por conceptos hace alusión a la noción de synset. Por lo tanto, esta desambiguación en base al significado es mucho más precisa que si tenemos en cuenta únicamente la palabra como conjunto de letras.

⁸ <http://sentiwordnet.isti.cnr.it/>

#	POS	ID SYNSET	=	PosScore	NegScore	SynsetTerms	Gloss
a		00011665		0.125	0.375	too-greedy#1, overgreedy#1,excessively	gluttonous

Tabla 4. Ejemplo de entrada en SentiWordnet

1. Ella es terriblemente orgullosa. → La palabra orgullosa en este contexto se refiere a la arrogancia y tiene una valoración negativa.
2. Una madre siempre está orgullosa de sus hijos. → En este caso, el orgullo es sinónimo de satisfacción y la polaridad es positiva.

Otra vez el contexto juega un rol muy importante. ¿No es por tanto indiscutible la necesidad de ‘comprender’ lo que se está diciendo? ¿De entender el significado de los términos? Aquí es donde los synsets y la desambiguación se hacen esenciales. Por esto mismo, aprovechando estos procesos ejecutados por la librería *Freeling* en un módulo anterior, somos capaces de asignar polaridades previas con una cierta seguridad de que cada término puede ser valorado según su verdadero significado.

SentiWordnet devuelve dos valores concernientes a la polaridad positiva y a la polaridad negativa. Si restamos ambas cifras obtenemos la polaridad real, delimitada por el intervalo $[-1, 1]$, donde -1 es negativa al máximo y $+1$ extremadamente positiva. Según el corpus TASS los rangos posibles entran dentro de esta clasificación {N+, N, Neu, P, P+, None}. Internamente realizamos una asignación de polaridad dentro de esa clasificación comparando los valores numéricos con las siguientes fórmulas lógicas:

Valor numérico	Polaridad
>0.6	P+
$>0.2 \ \&\& \ <0.6$	P
$>-0.2 \ \&\& \ <0.2$	None o Neu
$<-0.2 \ \&\& \ >-0.6$	N
<-0.6	N+

Tabla 5. Límites de las polaridades.

‘Neu’ se corresponde con la polaridad neutral y es un caso especial que no es posible asignar a un único término. Es decir, es necesario que su valor numérico esté en ese intervalo (cercano a 0) pero siempre porque en el texto se han encontrado polaridades de los dos extremos de la clasificación. Se asume que la polaridad Neutral es sinónimo de indecisión.

5.11 Tratamiento de modificadores



Figura 14: Tratamiento de modificadores.

Cuando hablamos de modificador, nos referimos a aquel elemento dentro de la oración capaz de alterar la polaridad de algún término o inclusive de la proposición en su conjunto.

Un modificador típico es el adverbio “muy”. ¿Qué ocurre con la palabra ‘bonita’ cuando va precedida de este modificador? Sin tener que meternos en materia técnica ni lingüística, un oyente es capaz de distinguir perfectamente la diferencia y el valor de la palabra ‘bonita’ en solitario y cuando va acompañada de esta forma léxica. El resultado de aplicar un modificador es una alteración de la polaridad que puede revelarse como una reducción de grado o como un aumento. Por lo tanto, son también útiles para pasar de clasificaciones con polaridades elementales ({N, Neu, P, None}) a clasificaciones de mayor rango donde se tienen en consideración sentimientos más extremos (P+ y N+).

Dos ejemplos sencillos:

- Esta cartera es **muy** [bonita/Polaridad: P]. → La palabra ‘bonita’, de por sí, tiene una polaridad previa positiva (ya extraída en el módulo precedente) que junto con el modificador ‘muy’ debería ser aumentada.
- Esta cartera es **poco** [bonita/Polaridad: P]. → En este caso, el modificador hace justo lo contrario, convirtiendo algo positivo en algo negativo.

Veamos primero qué es ese diccionario de modificadores. Cada entrada en este diccionario tiene el siguiente formato:

Modificador	Grado de aumento/reducción de polaridad
-------------	---

extremadamente	++
Bastante	+
menos	-

Tabla 6: Ejemplo de entrada del diccionario de modificadores.

El grado significa qué nivel de repercusión tiene sobre el foco de la modificación. Determinar el foco es el asunto clave en este módulo.

Gramaticalmente, los modificadores son adverbios, así que la pregunta es clara: ¿a qué puede acompañar un adverbio?

1. A un adjetivo (Ejemplo: **Muy** guapa).
2. A otro adverbio (Ejemplo: **Realmente muy** guapa).
3. A un verbo (Ejemplo: Insulta **mucho**).

Dado que hemos anotado con la librería *Freeling* cada palabra con su etiqueta gramatical, no parece difícil distinguir los 3 casos. El primer caso, de hecho, es el más sencillo y la aplicación del modificador funciona como expliqué con los ejemplos anteriores (aumentando o reduciendo la polaridad del siguiente término).

En el segundo caso nos encontramos ante algo que hemos denominado ‘agrupación de modificadores’. Una agrupación, como su nombre indica, es un conjunto de modificadores que finalmente actuará como uno solo. La combinación de sus grados se evalúa teniendo en cuenta que el primer modificador no actúa realmente sobre el adjetivo/verbo, sino sobre el segundo modificador.

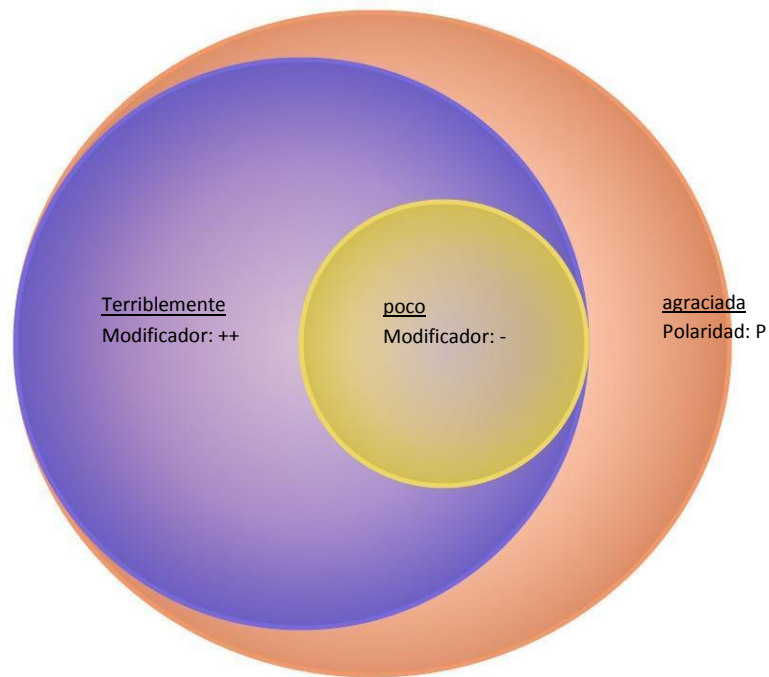


Figura 15: Ámbito de los modificadores.

En este ejemplo podemos observar cómo actúan las agrupaciones de modificadores. La bola naranja es el conjunto global de modificadores más el adjetivo al que modifican, lo que constituirán finalmente un único conglomerado con una sola polaridad. La bola morada, en cambio, implica a la agrupación y significa que el primer modificador actúa sobre el que contiene (bola amarilla). El resultado de la conjunción de ambos modificadores es algo similar a esto:

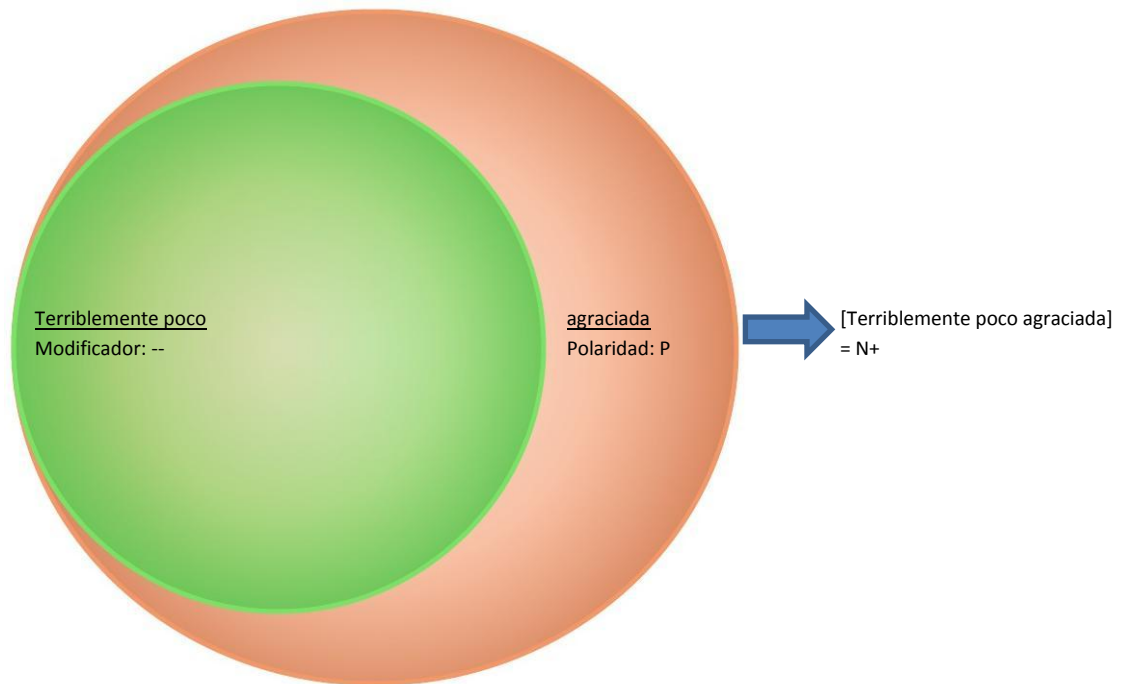


Figura 16: Agrupación de modificadores.

Por lo que podemos deducir algunas expresiones lógicas que adjunto en el cuadro siguiente:

Mod 1	Mod 2	Polaridad agrupación	Polaridad núcleo	Polaridad global del conjunto	Ejemplo
+ Or ++	+ Or ++	++	P Or P+	P+	Realmente muy guapa
+ Or ++	+ Or ++	++	N Or N+	N+	Realmente muy fea
+ Or ++	- Or --	--	P Or P+	N+	Realmente poco guapa
+ Or ++	- Or --	--	N Or N+	P+	Realmente poco fea

Tabla 7. Cálculo de los modificadores

El primer modificador aumenta o disminuye el valor del segundo y cuando el valor de la agrupación es calculado pueden ocurrir dos cosas:

1. Si la polaridad de la agrupación es ++, aumentará el grado de la polaridad del núcleo.

2. Si la polaridad de la agrupación es --, la polaridad del núcleo es invertida.

Volviendo a los casos, nos queda por comentar el caso 3 cuando el modificador afecta a la proposición. La repercusión de los modificadores en este caso es mucho más obvia:

“Me gustaba **terriblemente**” → Si “gustar” ya contiene polaridad positiva, “gustar terriblemente” tendrá polaridad doblemente positiva. El resultado de tener un modificador adverbial que acompaña al verbo provoca que la polaridad de toda la proposición, que es el ámbito de dicho verbo (si recordamos bien qué significa el término ‘proposición’), se vea alterada siguiendo las pautas ya mencionadas.

Por último, debo comentar que aunque se ha generalizado haciendo alusión sólo a los adverbios, se han introducido también modificadores formados por locuciones adverbiales:

[**Sin duda alguna**] ella es guapa → pasa de polaridad P a P+ gracias al modificador.

5.12 Inversión de polaridad – Negación

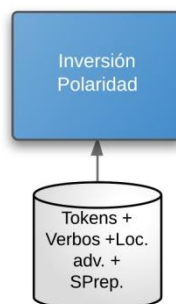


Figura 17: Inversión de la polaridad.

La negación tiene un efecto contundente sobre la polaridad de un elemento: la inversión. Esto quiere decir que la polaridad hallada hasta ahora es totalmente invertida pasando de positiva a negativa y a la inversa en cualquiera de los grados en los que se encuentre la clasificación ({N+, N, P, P+}). Fijémonos en este comportamiento:

1. Ella es muy guapa. → P+
2. Ella **no** es muy guapa. → N

Donde la parte del texto que está subrayada es la que contiene realmente la polaridad.

Podemos comprobar como la polaridad ha cambiado totalmente aunque debemos comentar que no se realiza una inversión total ya que en el segundo ejemplo, la polaridad no es extremadamente negativa. Este hecho se puede observar también en el caso contrario.

1. Ella es muy fea. → N+
2. Ella **no** es muy fea. → P (sería una falacia considerar que la mujer es una belleza).

La negación puede detectarse en distintos elementos. El más común es en un conjunto de *tokens* o palabras como “no, nadie, nada, nunca,...” pero existen otros elementos con igual poder de inversión como locuciones adverbiales, sintagmas preposicionales o verbos.

Respecto a las locuciones adverbiales podemos afirmar que es un caso particular del conjunto de *tokens* negativos con la peculiaridad de ser construcciones compuestas. Ejemplos típicos que podemos encontrarnos son: jamás en la vida, nunca jamás,...

En cuanto a los sintagmas preposicionales, podríamos nombrar el más común: ‘en vez de’. Esta formación niega al completo todo lo entra en el ámbito del sintagma. Veámoslo con un caso práctico:

3. La chica es guapa/Polaridad: P. → P
4. La chica,
[en vez de ser guapa:Polaridad: P]/Polaridad: N,
es fea/Polaridad: N. → N

En el punto 4 podemos comprobar como la traducción del sintagma es algo similar a una relativa del estilo a: “que no es guapa” y por tanto, es necesario aplicarle los procesos diseñados para el tratamiento de la negación. Entrados en materia, podemos ahora entender la razón por la que la segmentación también tenía en cuenta este tipo de sintagmas. Ésta era su verdadera utilidad.

Un comportamiento que ha podido llamar la atención es cómo dos negaciones no alteran la percepción recogida. En castellano, al contrario que en inglés, una

negación no invierte el efecto de otra negación⁹ si se topan en el mismo ámbito (en nuestro caso, hemos decidido que el ámbito general de las negaciones es la proposición). Simplemente la recalcan:

No voy nunca al dentista. ≠ Sí que voy al dentista.

No vino nadie. ≠ Sí vino alguien.

Los sintagmas preposicionales son una excepción para este fenómeno dado que realmente son una alternativa a una proposición y por tanto tienen entidad propia separada. Esto quiere decir que si hallamos un elemento de negación dentro del sintagma, sólo tendrá efectos en su ámbito y una vez fuera, en la proposición, esta negación no tendrá ningún efecto inversor sobre la proposición. Se puede comprobar este funcionamiento retomando un ejemplo anterior:

5. [La chica,

[**en vez de**/Elemento de negación

Ámbito negación[ser guapa/Polaridad: P,]

] /**Polaridad SP: N**

es fea/Polaridad: N.] → Polaridad proposición: N

Por lo tanto, si la proposición ya contuviera un elemento capaz de negar su polaridad, no existiría un caso de doble negación. En resumen, el ámbito del sintagma preposicional es invisible para la proposición.

6. [He vivido

[**sin**/Elemento de negación

Ámbito negación[amor/Polaridad:P]

] /**Polaridad SP: N**

.] → Polaridad proposición: N

7. [**Nunca**/Elemento de negación

he vivido

[**sin**/Elemento de negación

Ámbito negación[amor/Polaridad:P]

] /**Polaridad SP: N**

.] → Polaridad proposición: P

⁹

[http://www.rae.es/rae/gestores/gespub000018.nsf/\(voAnexos\)/arch8100821B76809110C12571B80038BA4A/\\$File/CuestionesparaelfAQdeconsultas.htm#ap8](http://www.rae.es/rae/gestores/gespub000018.nsf/(voAnexos)/arch8100821B76809110C12571B80038BA4A/$File/CuestionesparaelfAQdeconsultas.htm#ap8)

En cuanto a los verbos, debemos hacer distinción sintácticamente entre los verbos cuyo objeto negado estará en el ámbito de la proposición y aquellos verbos que pueden precisar de otra nueva proposición para representar lo que están negando.

8. Esa chica carece de belleza.

9. Proposición principal [No creo

Proposición Sustantiva [que esa chica sea bella]

.]

En el caso (8) el tratamiento es similar al hallazgo de *tokens* negativos, pero en el siguiente caso vemos una excepción a la norma inicial de que la proposición sería el ámbito de la negación. De hecho, a efectos de polaridad: *la chica ya no es bella* y esto provoca que se traslade el ámbito de la negación a un conjunto complejo compuesto por las dos proposiciones.

Para poder identificar estos casos se ha empleado el módulo del estilo indirecto (que viene a continuación), dado que esta forma de proposición sustantiva es la que permite el diálogo en estos términos. Veamos como el ejemplo (9) es un giro a una frase en estilo directo:

No creo que esa chica sea bella → Esa chica no es bella. (Pensado y dicho por el mismo sujeto)

Es necesario hacer hincapié en que la persona reflejada en el verbo (en este caso 1ª persona) es un elemento clave a la hora de determinar este fenómeno y finalmente la polaridad. Esta afirmación se comprenderá en los siguientes párrafos.

5.14 Procesamiento del estilo directo e indirecto



Figura 18: Procesamiento del estilo directo e indirecto.

En la lengua castellana, existen dos estilos para reproducir los mensajes.

El estilo directo es aquel en que se expresa a modo de cita las palabras de otro individuo:

- (1) Ella dijo: “el perro huele mal”.

El estilo indirecto es el empleado por el narrador para reproducir las palabras de otro sujeto o de sí mismo:

- (2) Ella dijo que el perro olía mal.

- (3) Yo dije que el perro olía mal.

- (4) Ella no dijo que el perro oliera mal. → ¿Huele bien? No lo sabemos, es un comentario otorgado por un tercero.

Cuando tratamos con una cita o con un mensaje emitido por otra persona, no se debería tener en cuenta la polaridad de las palabras en dicho comentario dado que no es una sensación del redactor, sino de una tercera persona (ejemplos 1 y 2) y el propósito del que las transcribe no es otro más que proporcionar **objetivamente** una información. En este módulo, al detectar estos usos, automáticamente la polaridad de la cita o proposición sustantiva se vuelve *None*.

Caso diferente es cuando el emisor del mensaje es una primera persona (3) y las palabras reflejan el sentimiento del redactor. En esta ocasión, las valoraciones sí son tomadas en cuenta y la proposición subordinada que le sigue puede tornarse un caso de la negación:

(7) Yo creía que el perro olía mal.

(8) Yo **no** creía que el perro oliera mal. → Parece que olía al menos medianamente bien.*

(9) Yo **no** creía que el perro oliera bien. → Tampoco implica que huela fatal.*

En los 3 casos, el narrador opina subjetivamente sobre un hecho, pero existen diferencias en el tratamiento de las 3 oraciones. En el ejemplo (7), ambas proposiciones son afirmativas y no existe ningún elemento que haga tambalearse la semántica de ambas. En los dos ejemplos posteriores, sin embargo, el hallazgo de una negación en la proposición principal nos obliga a realizar un tratamiento más tedioso en el módulo anteriormente descrito. Básicamente, es necesaria una inversión de la polaridad de la proposición subordinada entendiéndolo como inversión no un cambio radical de signo, sino una inversión “suavizada”*.

Para el procesamiento de estas funcionalidades se usa un diccionario que contiene los verbos (realmente almacena *synsets* para abarcar con menos entradas más términos dado que se agrupan por sinónimos) más comunes que dan lugar a este tipo de diálogo (“creer”, “pensar”, “decir”, “considerar”...). También se le añadieron posteriormente construcciones más complejas como: “No veo posible que...”, “Es improbable que...”,...

5.13 Obtención de la polaridad final

Una vez que hemos anotado los elementos susceptibles de ser pistas clave para la obtención de la polaridad final sólo queda reunirlos todos en un único algoritmo que en última instancia nos dé las clasificaciones pertinentes del tweet completo.

Dichas clasificaciones se enmarcan dentro de las esperadas para el corpus TASS: la clasificación de polaridad y la clasificación de acuerdo. La última indica el grado de consenso de las polaridades obtenidas. Es extremadamente útil para determinar cuándo la polaridad neutral aparece por la aparición de valoraciones positivas y subjetivas al mismo tiempo.

Sobre la primera, el algoritmo va agrupando elementos con sus polaridades finales de forma recursiva, del elemento más pequeño a las grandes asociaciones como son las oraciones. Cuando el procesamiento llega a este módulo, realmente lo que tenemos como entrada son proposiciones para las que ya se han tratado independientemente sus casos peculiares de modificación o negación. Lo que falta por tanto es aglutinar estas polaridades y devolver un único valor dentro de la

clasificación que hemos ya ido manejando en los módulos predecesores ({N+, N, Neu, P, P+, None}). Esta aglutinación no es más que el cálculo de la media de los elementos. Pero no es un cálculo básico sino una media donde los elementos son ponderados según su nivel en el árbol sintáctico. En el sistema, el nivel 1 corresponde al nivel principal y que más poder debiera tener, el nivel 2 a las proposiciones subordinadas, el 3 a subordinadas dentro de subordinadas, y así sucesivamente.

```
while (thereAreMoreSintacticUnits)
{
    if (sintacticUnitPolarity!=NEU)
        if (sintacticUnitPolarity!=NONE)
        {
            valueOfLevel = 1 - ( propositionLevel / 10 );
            sumScores += sumScores+(scoreOfParticularUnit *
valueOfLevel) ;
        }
}
scoreOfNominalPhrase=sumScores/numSintacticUnitsWithPolarity;
```

¿Cuál es la teoría que sustenta esta operación? Según [25], se debería considerar cada palabra como mayor o menor portadora de polaridad según su nivel en el árbol sintáctico. Nosotros hemos trasladado este concepto a las proposiciones, fundamentando la teoría en que la proposición principal tiene mucha más relevancia en la determinación de la polaridad y que ya que las subordinadas, por definición, no tienen personalidad individual propia, su polaridad debe tener distinto valor o influencia en el global.

Podemos consolidar esta estrategia analizando la siguiente cita:

[El periódico

[que suele ser pesimista] /Polaridad: N/Nivel: 2

ha dado hoy estupendas noticias.] ¿¿P??/Nivel: 1

¿Deberíamos dar el mismo valor a la subordinada relativa que a la principal? ¿Cuál es la idea principal? Nuestro criterio indica que la proposición principal de nivel 1 es la que tiene más responsabilidad en la transmisión del mensaje y que por ello su valor debe ser superior. Según nuestro algoritmo, el resultado en el ejemplo sería un sentimiento positivo.

Es obvio que si todas las proposiciones en el tweet contienen polaridades positivas, o por el contrario negativas de forma consensuada, al agregarlas el resultado no variará. Todos los sentimientos positivos convergerán en uno positivo y del mismo modo los negativos lo harán hacia un resultado negativo. El problema surge cuando

hay discordancia. Si se han hallado proposiciones con polaridad positiva y negativa mezcladas podemos ratificar que el twittero no tiene una opinión firme. Ésta se debe precisar con la valoración neutral (Neu) y la definición de desacuerdo.

La clasificación de acuerdo/desacuerdo (*{Agreement/Disagreement}*) permite distinguir estos casos, cuando no todos los sentimiento viajan en el mismo sentido. En nuestro sistema, siguiendo el enfoque del grupo de Daedalus [17], empleamos la desviación estándar para descubrir fluctuaciones notables alrededor de la media y poder certificar el desacuerdo o *disagreement*. Matemáticamente, este valor es una medida del grado de dispersión de los datos con respecto al valor promedio. Es decir, constituye simplemente la variación con respecto a la media aritmética. Cuando la desviación se encuentra por encima del 0.2 podemos considerar que existe desavenencia.

5.14 Creación de características

Uno de los puntos fuertes de este trabajo es que pretende que todos los procesos anteriores no dejen de tenerse en cuenta aunque los juicios de relevancia indiquen finalmente que el sistema no es óptimo (incluso aunque no se pretendiera).

Inicialmente, cuando se llevó a cabo el *brainstorming*, no estaba muy claro qué enfoque seguiría este trabajo. Por ello, se realizaron investigaciones siguiendo diversas líneas entre las que también se encontraban los trabajos realizados con Minería de Datos. Lo cierto, es que ambas perspectivas comparten muchas de las características como por ejemplo los indicios de enfatización (repetición de letras, uso de mayúsculas,...) o incluso la polaridad previa. Es por ello por lo que dentro del planteamiento se incluyó la idea de anotar el corpus TASS y crear un nuevo corpus que permitiera en un futuro probar nuevas teorías e inclusive cambiar radicalmente de enfoque.

En el anexo 3, al final de la memoria, se pueden localizar ejemplos del corpus anotado.

6. Experimentación y pruebas

Para la experimentación se ha diseñado un marco de pruebas que contiene diferentes configuraciones para comprobar el proceso y aporte de algunos de los módulos más relevantes.

La programación estructurada permite activar o desactivar los módulos para dicha tarea aunque no todos ellos son conceptualmente anulables. Esto quiere decir que dependiendo de las tareas que les conciernen no siempre es admisible desactivarlos. Por ejemplo, los tratamientos considerados dentro del preprocesamiento son inapelables y su no ejecución trae consecuencias muy desafortunadas sobre los resultados (hasta el punto de que la herramienta *Freeling* no es capaz de crear sentencias en el módulo de segmentación básica y todo el proceso siguiente es cancelado).

Sin embargo, existen otros módulos cuya labor no se basa en la adecuación de los textos para su consiguiente procesado, sino que tienen un carácter enriquecedor para la precisión. Por ello, hemos diseñado este marco de pruebas que nos permitirá cerciorarnos de la contribución de cada proceso, además de facilitarnos la localización de los puntos débiles del entorno. Descubrir las flaquezas nos posibilitará el análisis de nuevas vías en investigaciones futuras.

- Configuración 1: activación del tratamiento de modificadores.
- Configuración 2: activación del tratamiento de inversión de polaridad /negación.
- Configuración 3: activación de la detección de objetividad en estilo indirecto y enunciados de citas.
- Configuración 4: activación de todos los módulos.

Nuestro *baseline* consistirá en una configuración básica sin ninguno de los módulos aquí arriba citados activados. Cada configuración permitirá activar únicamente uno de ellos y analizar su aporte y funcionamiento de forma aislada.

Para evaluar los resultados se usarán los juicios de relevancia proporcionados por Daedalus. Aunque la polaridad obtenida en cada tweet se debería cotejar dentro de la clasificación de 6 niveles, las evaluaciones también incluirán las métricas considerando únicamente los 4 niveles elementales ({P, Neu, N, None}), realizando el mismo procedimiento que en el *workshop* de TASS.

6.1 Resultados

En la siguiente tabla se presentan los resultados obtenidos:

Conf ig.	Prec. 4 cat	Prec. 6 cat	P+	P	Neu	N	N+	None
Base line	0,50822	0,35401	0,1923414	0,366848	0,161389	0,205146	0,082836	0,656345
1	0,508861	0,359024	<u>0,2063818</u>	0,36079	0,154326	0,202592	<u>0,091696</u>	0,656748
2	0,507789	0,354011	<u>0,188195</u>	0,366168	0,159073	0,213427	<u>0,080186</u>	0,656721
3	0,50774	0,354001	0,192107	0,366168	0,159073	0,20358	0,082836	<u>0,657549</u>
4	0,507847	0,359048	0,201791	0,360109	0,149649	0,209689	0,089254	0,65833

Tabla 8: Resultados del trabajo para la medida de precisión.

Antes de pasar a comentar los resultados, debo apuntar que finalmente y tras muchos intentos, se tuvo que desactivar la detección de entidades ya que el servidor de Daedalus no podía atender la gran cantidad de peticiones generadas por las ejecuciones en paralelo desde un cliente con licencia para investigación. Se mantuvo contacto con el administrador de la herramienta pero desgraciadamente, su puesta a punto no estuvo a tiempo para incluir el módulo en las pruebas. Aunque sí que es posible mostrar la bajada de precisión tras este incidente gracias a una prueba anterior que se llevó a cabo activando el módulo de detección de entidades para la última configuración.

Los resultados de aquel experimento para 20.000 tweets fueron los siguientes:

Config.	Prec. 4 cat	Prec. 6 cat	P+	P	Neu	N	N+	None
4 + DE.	0,525614	0,3639219	0,2024523	0,354581	0,1635883	0,212798	0,094265	0,6653763

Tabla 9: Precisión para 20.000 tweets con la 4ª configuración y detección de entidades.

Para comprobar el aumento de precisión tras utilizar una herramienta como la de Daedalus, se realizó la siguiente prueba: se recogió el mismo conjunto de tweets de su prueba equivalente (la 4ª configuración con la activación de todos los módulos) y se ejecutó la evaluación. La precisión obtenida se recoge en la tabla siguiente:

Config.	Prec. 4 cat	Prec. 6 cat	P+	P	Neu	N	N+	None
4 + DE.	0,511224	0,351203	0,201938	0,336049	0,166667	0,213183	0,087558	0,651722

Tabla 10: Precisión para 20.000 tweets con la 4ª configuración sin detección de entidades.

Como se puede comprobar, la precisión descendió más de un punto en un entorno similar al bloquear el uso de la detección de entidades. Es por tanto ineludible mencionar que mis resultados están condicionados por esta restricción y que no son los óptimos, entendiendo óptimo como aquellos resultantes del sistema si su funcionamiento estuviera al completo.

Volviendo a la primera tabla, nos encontramos con que los resultados parecen no variar sea cual sea la configuración aplicada. La explicación a esto la hallaremos haciendo una pequeña revisión sobre el corpus empleado.

Aparentemente al menos, resulta obvio que el entorno creado parece no adecuarse a las características del corpus. Esto quiere decir que aunque los módulos ejecutan su función, el contexto de los tweets no encaja con el modelo creado o al menos está algo lejos de poder ser recogido por estas funcionalidades (detección de modificadores, negación y estilo directo e indirecto).

La clase peor representada en los resultados es la clase N+. La razón por la que esto ocurre es porque si revisamos estos tweets nos encontramos con que la mayoría de ellos son ejemplos como el siguiente:

```
<twit>
  <twitid>149521816954740738</twitid>
  <user>sanchez_sonia</user>
  <content><![CDATA["Vengo como un miembro más", dice
Chacón al llegar a la reunión de
#muchoPSOEporthacer]]></content>
  <date>2011-12-21T17:09:13</date>
  <lang>es</lang>
</twit>
```

La pregunta es: ¿podría una máquina detectar tal cantidad de subjetividad sujeta a un contexto? Al menos en el caso de este entorno, la respuesta es no. En el conjunto de tweets pertenecientes a la categoría de economía o política, la valoración de los jueces goza de una mayor subjetividad y requiere de un contexto e inclusive un conocimiento tan amplio que una buena categorización se hace casi inviable. La realidad es que la percepción podría ser totalmente alterada si este tweet lo escribiera y lo leyera un militante del PSOE. Por lo tanto, en esta coyuntura participan 3 personajes: el escritor, el juez y la tercera persona (yo) adentrada en el intento de obtener una respuesta que complazca o se ajuste a la del segundo (aunque debería ser más conveniente que encajara con la del primero).

Por otro lado, la siguiente clase con peor precisión es la clase neutral. Es posible que la razón radique en una falta de definición exacta sobre lo que esta clase implica. Según la teoría en la que este entorno se ha desarrollado, esta clase se aplica a los

tweets que mantienen disconformidad por contener sentimientos dispares. En cambio, buscando ejemplos en el corpus nos encontramos con casos como éste:

```
<twit>
  <twitid>143696388260237312</twitid>
  <user>alfonsomerlos</user>
  <content><![CDATA['Hoy abordaremos en @13tv_aldia
@13television la gestión privada de los hospitales públicos.
¿qué ganamos los ciudadanos? ¿qué perdemos?']]></content>
  <date>2011-12-05T15:21:03</date>
  <lang>es</lang>
  <sentiments>
  <polarity><value>NEU</value><type>AGREEMENT</type></polar
ity>
  </sentiments>
</twit>
```

Observando la valoración en cuanto a la polaridad nos encontramos con un tweet que aun siendo neutral se considera como miembro del tipo *agreement*. La consecuencia de estos hallazgos fue la suposición de que quizás o bien no estaba bien definida la clase por nuestra parte o bien el corpus tenía alguna falta de coherencia. Respecto a la polaridad asignada también existen serias dudas: ¿es neutral porque las dos preguntas del final contienen conceptos positivos y negativos respectivamente? Desde el punto de vista del creador y por consiguiente de este entorno, existe o debería existir una distinción clara entre que los conceptos individualmente puedan pertenecer a una determinada categoría y que el sentimiento detrás del contexto se acople también a lo recogido por dichos conceptos. De hecho, este tweet se ajusta más a un texto meramente informativo y por consiguiente objetivo. En este caso, parece posible que dichos términos hayan podido confundir la categorización. Este tweet es perfecto para establecer la barrera entre lo que se considera sentimiento u opinión (vinculado a las percepciones personales) y aquello puramente narrativo e imparcial.

Si nos centramos en la activación de módulos, comprobamos que el corpus parece indiferente a estas funcionalidades. En el caso del tratamiento de modificadores, sí que detectamos una muy ligera mejora en las clases afectadas al aplicarlos (debería existir una mejora de precisión al menos en las clases N+ y P+), pero es casi inexistente. La razón es que aunque el módulo ejecuta su función a la perfección (dentro de sus límites teóricos), no se encuentran apenas ejemplos en el corpus donde se pueda aplicar. Las clases más extremas lo son por otras razones ajenas a los modificadores y su clasificación está normalmente más vinculada a cuestiones de contexto.

En lo referente a la negación, los resultados de estas mismas categorías empeoran y la explicación a este fenómeno puede deberse al uso de la ironía o a la aparición de sentencias interrogativas en los que este sistema no repara.

```

<twit>
  <twitid>144470614957240321</twitid>
  <user>Albert_Rivera</user>
  <content><![CDATA['Por qué no se exige a los bancos
intervenidos con dinero público una línea de crédito para PYMES,
autónomos y nuevas empresas que contraten?']]></content>
  <date>2011-12-07T18:37:33</date>
  <lang>es</lang>
  <sentiments>
    <polarity><value>N+</value><type>AGREEMENT</type></polarity>
  </sentiments>
  <topics>
    <topic>economía</topic>
  </topics>
</twit>

<twit>
  <twitid>143990880381059072</twitid>
  <user>LosadaPescador</user>
  <content><![CDATA['A los okupas no les gusta el edificio de
Concepcion Jeronima xq esta en obras. Les gustan mas los
edificios a estrenar, pero sin pagar...']]></content>
  <date>2011-12-06T10:51:15</date>
  <lang>es</lang>
  <sentiments>
<polarity><value>NEU</value><type>DISAGREEMENT</type></polarity>
  </sentiments>
  <topics>
    <topic>economía</topic>
    <topic>política</topic>
  </topics>
</twit>

<twit>
  <twitid>144466101076627457</twitid>
  <user>rafam_simancas</user>
  <content><![CDATA['La novia de Cristiano Ronaldo, Irina de
nombre, prefiere "jugar" sin camiseta, se vé que no aprecia
tanto los colores. http://t.co/7hNdwoDD']]></content>
  <date>2011-12-07T18:19:37</date>
  <lang>es</lang>
  <sentiments>
    <polarity><value>P+</value><type>AGREEMENT</type></polarity>
  </sentiments>
  <topics>
    <topic>fútbol</topic>
  </topics>
</twit>

```

Revisando los anteriores ejemplos podemos concluir que no es una labor sencilla categorizarlos con una polaridad y, de hecho, cabe la posibilidad de que varios lectores difirieran en la clasificación. El aprendizaje de estas observaciones, sin embargo, debe basarse en el entendimiento de que la ironía u otras herramientas lingüísticas no deben pasarse por alto y deben tratarse como indicios más que evidentes de que las valoraciones basadas simplemente en términos nunca darán precisiones notorias.

Respecto al último módulo de detección de estilo directo e indirecto, debemos comentar que, de nuevo echando un vistazo al corpus, no parece adecuarse con las valoraciones otorgadas. Es posible que se deban tener en cuenta conceptos más complejos a la hora de determinar si una citación es objetiva o subjetiva. Si nos remitimos al ejemplo anterior sobre Chacón, donde este procesamiento dictaminaría un NONE, nos topamos con que, a pesar de existir un claro estilo indirecto, es valorado como negativo. Esto sucede porque en este proceso se entremezclan varias variables que juegan un papel determinante: la situación y el contexto socio-cultural del momento, el conocimiento del juez y su propia subjetividad (de hecho, es complicado poder valorar este tweet de forma contundente) y por último, la ideología del tweetero. Como tónica general, se ha observado que el estilo indirecto es igualmente evaluado y tomado en consideración. Ahondando en este tema, se ha encontrado un artículo [4] que afirma esta conclusión asegurando que al menos en el mundo de la política, entre un 62% y un 77% de usuarios (con mayor tendencia los de *derechas*) tienen a citar comentarios de usuarios de la ideología contraria. Las posibles mejoras se verán en el siguiente apartado.

Sin embargo, y a pesar de la poca efectividad demostrada por esta primera versión los módulos desarrollados, así como la simple combinación, el punto más relevante de esta evaluación es la gran precisión obtenida simplemente con el *baseline*, lo que le convierte en una buena base para futuros trabajos. La potencia de esta configuración tan aparentemente sencilla radica en el estudio y aplicación de una segmentación compleja (para lo que hubo que modificar la gramática de Freeling). Esto ratifica que estructurar y valorar las proposiciones independientemente y darles un peso relativo según su poder dentro del conjunto de la oración, es una elección acertada y puede constituirse como un punto clave a partir del cual desarrollar nuevas mejoras eficientes.

7. Conclusiones y futuras líneas de trabajo

Una vez vistas las deficiencias y los puntos fuertes del trabajo realizado, es un buen momento para el análisis profundo de los resultados anteriores y pasar a la concreción.

La falta de eficiencia de las funcionalidades propuestas es consecuencia de varios hechos:

- La ironía tergiversa la semántica de las proposiciones haciéndose necesario un tratamiento más sofisticado de las expresiones.
- El contexto y otros condicionantes pragmáticos involucrados en la comunicación alteran los significados. Dentro de este conjunto se encuentra el contexto así como las relaciones del escritor con el mundo que le rodea.
- Las redes sociales se han erigido como plataformas preferidas para la discusión y la protesta. El estudio del estilo directo e indirecto es una funcionalidad muy válida para detectar la objetividad, pero no para dar por hecho que una réplica o citación es siempre impersonal. En la mayoría de los casos en este corpus, los enunciados han sido valorados. La razón principal puede residir en la motivación de una persona al reescribir lo dicho por otra. Normalmente, la incitación vendrá dada bien o por ratificar lo expuesto o por condenarlo, por lo tanto no cabe lugar para la indiferencia.
- La posible ineficacia de alguno de los nuevos módulos con posibles errores gramaticales, lo que exige una prueba de la calidad de los resultados para cada módulo individualmente.

Uno de las grandes carencias del enfoque de este trabajo es el aislamiento del texto con el mundo que le rodea. La comunicación es un acto en el que se ven inmersos muchos factores sin los cuales las palabras no tendrían apenas valor. Por lo tanto, a partir de ahora, el camino se centra en buscar este conjunto óptimo de relaciones que afectan decisivamente en el acto comunicativo.

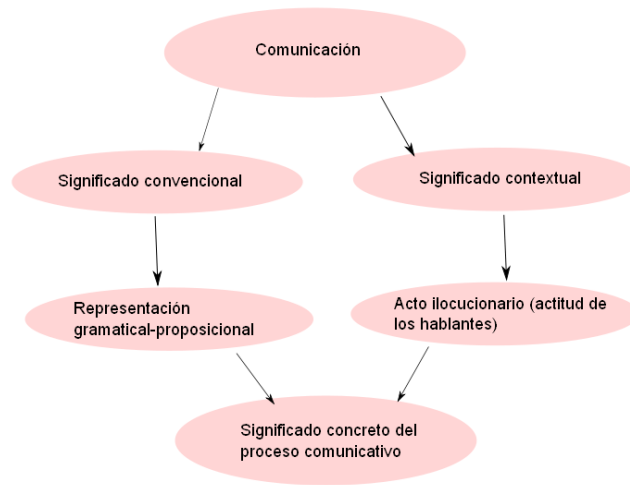


Figura 19: Acto comunicativo

Si nos fijamos en la figura se puede observar como la rama izquierda se adecúa perfectamente (aunque de forma muy general) al trabajo llevado a cabo. Ahora queda por explorar la rama de la derecha.

Un buen comienzo podría ser empezar por distinguir qué es subjetivo y qué no lo es. Para aquello puramente informativo, la rama derecha quedaría podada mientras que para los textos transportadores de percepciones viajarían por ambas dando un mayor peso a la rama derecha.

De cara a definir el contexto serían necesarias dos armas:

- Información sobre el escritor.
- Información sobre los temas que se abordan en el mensaje.

Si se fuera capaz de determinar la personalidad del autor e indagar sobre sus inclinaciones, se construiría más fácilmente un grafo teórico sobre el que asentar posibles hipótesis. Es decir, no es suficiente sólo con definir construcciones gramaticales susceptibles de constituir por ejemplo ironía ya que hemos comprobado como una mera citación puede ser símbolo de una formulación sarcástica. Investigar el perfil de los tweeteros (afinidades, personas que siguen, temas sobre los que suele comentar,...) sería un buen punto de partida. Básicamente, significa añadir una recolección previa de variables involucradas que permita, una vez definido el tema del mensaje, generar hipótesis (tanto para la semántica convencional como para la contextual) que naveguen por el grafo anterior para recombinarse y construir el significado concreto del proceso comunicativo. Si por ejemplo se detectara un perfil

con inclinaciones comunistas, en el punto en que converge el grafo se podría identificar con bastante precisión una ironía si gramaticalmente (análisis en la rama izquierda) el texto se ajustara a una construcción de tal tipo y si la rama derecha estuviera alertando de tales inclinaciones políticas no encajando con el sentido convencional del mensaje.

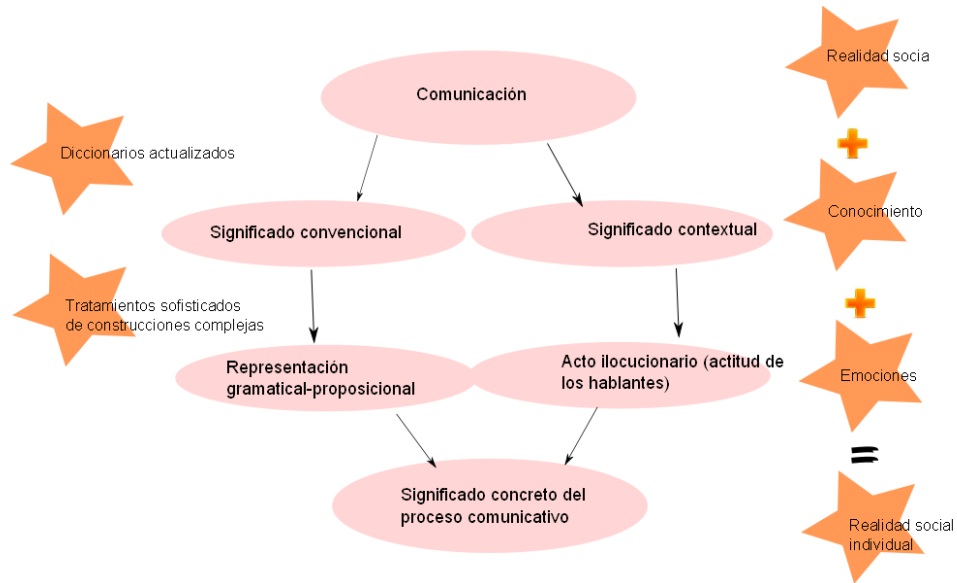


Figura 20: Acto comunicativo y necesidades de análisis.

Es cierto que inclusive la representación de la gramática y la semántica se han visto mermadas en este trabajo y requieren algunos reajustes.

Inicialmente sería necesario estudiar nuevas herramientas dado que, como ocurre con la librería *Freeling*, algunas de estas utilidades parecen no ser aun lo suficientemente estables y deterministas en sus resultados. En lo referente a esta librería, algunas de sus deficiencias se encuentran en su incapacidad de realizar una buena segmentación cuando las oraciones son bastante complejas generando un encadenamiento incorrecto de las proposiciones:

Prop 1[No me me parece correcta

SN [la acción

Prop Relativa [que hizo mi vecina

Prop 2[porque lo hizo desde el orgullo] → Subordinada dentro de la relativa. ERROR.

]].]

En algunos casos no afecta al rendimiento, pero cuando es necesario un tratamiento de la negación, su ámbito es erróneamente delimitado y los resultados fallan. De la misma forma, la librería presenta dificultades para identificar proposiciones yuxtapuestas con lo que el alcance de la negación traspasa fronteras y de nuevo nos encontramos ante un desenlace equivocado:

Prop 1[No me me parece correcta

SN [la acción

Prop Relativa [que hizo mi vecina]

]

, lo hizo desde el orgullo] → Este trozo debería ser una proposición independiente, de lo contrario el ámbito del término 'no' alcanzará a esta parte invirtiendo también su polaridad.

.]

También cabría la posibilidad de completar todos los diccionarios con nuevos conceptos e incluso generar otros lexicones con la ayuda de lingüistas profesionales.

De cara a futuras ampliaciones, en trabajos futuros deberían considerarse nuevas funcionalidades.

Una de las más importantes es el tratamiento de la modalidad irreal advertible por el uso del subjuntivo. Estos usos lingüísticos deben valorarse como afirmaciones no factuales donde los elementos tienen un significado más inadvertido (posiblemente generándose excepciones a reglas anteriores de otros módulos):

- Si no fuera estúpida, no tendríamos problemas. → La negación no invierte en este caso la polaridad, sino que funciona como un enfatizador que junto con la condicional muestra rotundidad ante un hecho considerado real por el interlocutor.
- No es sólo que fuera a pensar mal de mí, sino que podría pensar que lo he hecho a propósito. → otro caso de negación que es sensible de asignar erróneamente una polaridad positiva al primer fragmento de la oración.

- Le hubiera gustado contárselo. → la modalidad irreal puede convertirse en ocasiones en un inversor de polaridad ya que presenta situaciones o pensamientos de hechos que no han llegado a cumplirse. Suele ser portador de sentimientos negativos como la tristeza o desazón.

Otra funcionalidad interesante sería el tratamiento de frases interrogativas que ha resultado una de las herramientas más usuales en los tweets para la intromisión de la denuncia o demostración de la inconformidad. El caso más típico podría ser el siguiente:

- ¿**No** sería más fácil si el gobierno escuchara a los gobernados?
- ¿**No** debería la comunidad ponerse en pie y defender sus derechos?

Por otro lado, la ironía también emplea construcciones favoritas para recalcar su labor de reclamo o agitador de conciencias:

- Parece que a los funcionarios **no les gusta** que les rebajen el sueldo, **prefieren** rebajar su “trabajo”.
- A los okupas **no les gusta** el edificio de Concepción Jerónima porque está en obras. **Les gustan más** los edificios a estrenar, pero sin pagar...
- ¿**Qué puede esperarse** de un hombre que lo único que hizo fue casarse con una famosa?

Las comparaciones, el uso de comillas o algunas interrogativas forman claros síntomas de esta figura retórica.

Todos estos indicios, desde los implicados en la definición contextual hasta los más puramente lingüísticos, deben ser recogidos por nuevas funcionalidades en un trabajo futuro. Básicamente, se trata de reforzar ambas ramas del grafo para establecer un aceptable nivel de confianza en los resultados obtenidos.

Estos mensajes también han demostrado que rara vez en esta plataforma un texto debe interpretarse sólo desde el punto de vista del significado más corriente. El universo de las redes sociales y su utilidad como defensor de la difusión del pensamiento libre han creado un movimiento social que bien podría apodarse: *digo lo que quiero y como quiero* y que requiere de herramientas muy sofisticadas capaces de entender más de contextos que de formas léxicas.

8. BIBLIOGRAFÍA

- [1] Balbachan, F., Dell’Era, D. Análisis Automatizado de Sentimiento en Textos Breves de la Plataforma Twitter Resumen.
- [2] Na, J.C., Sui, H., Khoo, C., Chan, S., & Zhou, Y. (2004). Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference (pp. 49-54). Wurzburg, Germany: Ergon Verlag.
- [3] Kim and E. Hovy. 2007. Crystal: Analyzing predictive opinions on the web.
- [4] Pang, B., Lee, L., Vaithyanathan, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. En EMNLP-CoNLL.
- [5] T. Mullen, R. Malouf. 2006. A preliminary investigation into sentiment analysis of informal political discourse. In Proceedings of the AAI Symposium on Computational Approaches to Analyzing Weblogs, 159–162.
- [6] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In Proceedings of SIGKDD, 2011.
- [7] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. Sentiment analysis of twitter data. In Proceedings of the ACL 2011 Workshop on Languages in Social Media (2011), pp. 30–38.
- [8] Pak, A., Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In *Proc. of LREC*, 2010.
- [9] E. Riloff, S. Patwardhan, and J. Wiebe, Feature subsumption for opinion analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2006.
- [10] Go, A., R. Bhayani, and L. Huang. Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.
- [11] Go, A., R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. Technical report, Stanford Digital Library Technologies Project, 2009.
- [12] Riloff, Ellen and Janyce Wieb. Learning extraction patterns for subjective expressions. In EMNLP, 2003.
- [13] Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M. Learning subjective language. Technical Report TR-02-100, 2002.
- [14] Tumasjan, A., Sprenger, T. O., Sandner, P. G.; and Welpe, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Fourth International AAI Conference on Weblogs and Social Media, 2010.
- [15] Taboada, M., Brooke, J., Toloski, M., Voll, K., Stede, M. Lexicon-based Methods for Sentiment Analysis. *Journal of Computational Linguists*, 2010.
- [16] Kouloumpis, E., Wilson, T., Moore, J. Twitter Sentiment analysis: The good the bad and the omg! In Proceedings of the ICWSM (2011).

- [17] Villena Román, J., Lana-serrano, S., Moreno García, C. DAEDALUS at RepLab 2012: Polarity Classification and Filtering on Twitter Data.
- [18] Yang, C., Bhattacharya, S., Srinivasan, P. Lexical and Machine Learning approaches toward Online Reputation Management.
- [19] The Kelsey group. Online consumer-generated reviews have significant impact on offline purchase behavior. Press Release, November 2007. <http://www.comscore.com/press/release.asp?press=1928>.
- [20] Han, X., Zhao, J. Named entity disambiguation by leveraging wikipedia semantic knowledge. In Proceeding of the 18th ACM conference on Information and knowledge management, CIKM '09, pages 215–224.
- [21] De Albornoz, J.C., Chugur, I., Amigó, E. Using an Emotion-based Model and Sentiment Analysis Techniques to Classify Polarity for Reputation.
- [22] T. Mullen, R. Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs, 159–162, 2006.
- [23] Grossman, D.A., Frieder, O. Information retrieval: algorithms and heuristics. Boston: Kluwer Academia Publishers, 1998.
- [24] Urizar, X.S. Fundazioa, Elhuyar. TASS: Detecting Sentiments in Spanish Tweets.
- [25] Roser Morante and Eduardo Blanco. *SEM 2012 shared task: Resolving the scope and focus of negation. In Proceedings of the First Joint Conference on Lexical and Computational Semantics, Montreal, 2012.
- [26] Szarvas, G., Vincze, V., Farkas, R., Csirik, J. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In Proc. Of BioNLP 2008, pages 38–45, Columbus, Ohio. ACL.
- [27] Morante, R., A. Liekens, and W. Daelemans. 2008. Learning the scope of negation in biomedical texts. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 715–724.
- [28] Pang, B., Lee, L. Opinion Mining And Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008.
- [29] Batista, F., Ribeiro, R. The L2F Strategy for Sentiment Analysis and Topic Classification.

PARTE III: ANEXOS

ANEXO 1: Estructura de datos

En este anexo se pretende dilucidar la complejidad estructural de la segmentación compleja. Como ya expliqué en el apartado correspondiente, la segmentación realizada contemplaba ciertas unidades sintácticas (clase que se puede observar en el diagrama siguiente) en el siguiente orden de mayor a menor según su complejidad:

- Oración
- Proposición
- Sintagma Nominal
- Sintagma preposicional
- Token o término

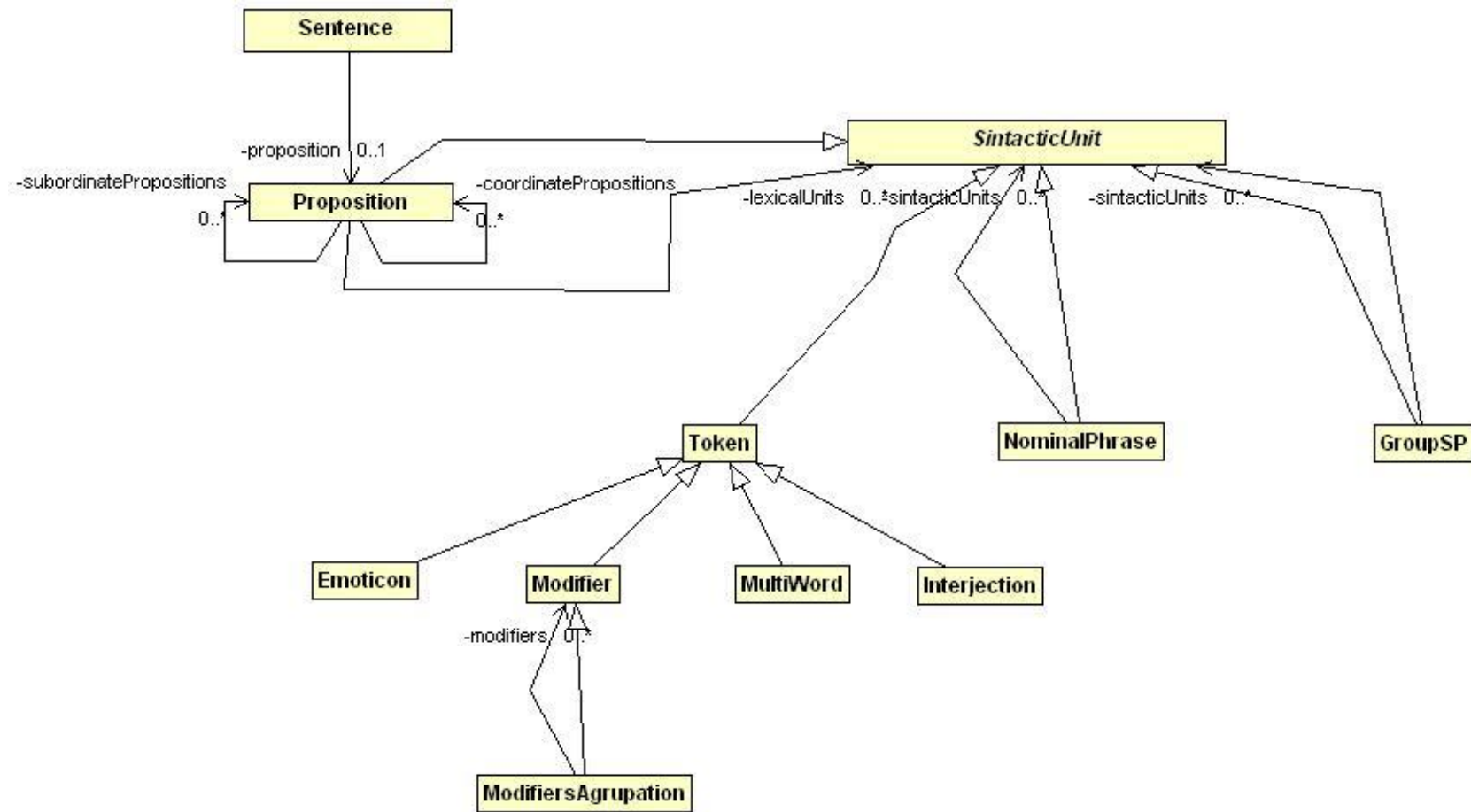


Figura 21: Diagrama de clases de elementos sintácticos.

La clase unidad sintáctica es aquella destinada a abarcar cualquier tipo de estructura capaz de ser albergada dentro de un árbol sintáctico para una sentencia. Esta generalización de clases (dado que todas las clases excepto sentencia heredan de ella) contiene las propiedades que todas ellas comparten y es un mecanismo de programación para permitir inclusiones de cualquier tipo en clases o estructuras de datos contenedoras. Así, por ejemplo, una proposición contiene unidades sintácticas que pueden ser cualquier tipo de especificidad dentro de esa clase (tokens, modificadores, multipalabras, interjecciones, otras proposiciones...) y lo mismo ocurre con los sintagmas nominales o preposicionales: los primeros pueden encerrar una proposición relativa mientras que los segundos proposiciones en cualquier grado de encadenamiento.

Además, una proposición puede contener coordinadas o subordinadas interiormente yendo desde el nivel más general dentro de su estructura hasta el nivel más interno. Esta adecuación constituyó una de las labores más complicadas en cuanto a su manejo ya que obligaba a la utilización de métodos recursivos que recorrieran todos los niveles en la búsqueda y cálculo de características.

ANEXO 2: Cuestiones técnicas sobre la experimentación

Para llevar a cabo la experimentación también se tuvieron que tener en cuenta factores ajenos al trabajo de investigación pero muy concluyentes en la ejecución y puesta en marcha del entorno. La consideración más preocupante fue la velocidad.

Una vez que finalizó la tarea de programación se pasó a realizar pequeñas pruebas para comprobar el correcto funcionamiento de los módulos y de todo el conjunto en general. Las anotaciones o marcas realizadas en el código para testar los tiempos dieron la increíble suma de 250 horas para la ejecución completa del código.

Existían diversos factores que contribuían a dicha situación: el procesador del ordenador, la tardanza en las llamadas al servicio web de Daedalus y el carácter secuencial del propio sistema.

Para cada uno de los obstáculos se propusieron una serie de alternativas:

1. Migrar de un ordenador de sobremesa y su limitada capacidad a un mainframe u computadora más potente. La UNED puso a disposición su máquina soporte con 12 núcleos en 6 CPU's que permitiría reducir los tiempos de procesamiento.
2. Las llamadas a Daedalus eran el claro cuello de botella del sistema, añadiendo entre cada llamada unos 6 segundos de desfase. Tras varios ajustes conseguimos reducir este tiempo evitando que el servidor almacenara nuestra sesión y encolara las peticiones (gestión que suelen tramitar con el fin de protegerse a sí mismos de un ataque masivo).
3. El carácter secuencial del procesamiento obligaba a mantener el orden para que todos los módulos se mantuvieran en constante sincronía. Pero por otro lado, el procesamiento para el corpus TASS era totalmente repetitivo. Exactamente 60.798 veces de procesamiento conceptualmente independiente. Para pasar de lo conceptualmente a lo físicamente independiente, se probó a lanzar hilos concurrentes como mecanismo de paralelización. El problema residía en que la librería Freeling no era *thread-safe* así que el intento fue en vano (aunque una parte del código finamente fue ejecutado con hilos). La otra opción tangible era el multiproceso, creando un pool de procesos donde cada uno estaba encargado de un trozo del corpus. De esta forma se lanzaban concurrentemente x procesos controlando que en cuanto uno finalizara uno nuevo entrara con la asignación de otro subconjunto del corpus. De esta forma, se mantenían siempre en ejecución un número constante de procesos.

Configuración de ejemplo:

- ❖ Tamaño del *pool* de procesos: **10** → se lanzarán concurrentemente 10 procesos.
- ❖ Número de tweets asignados para procesarse en cada proceso: **1.000**

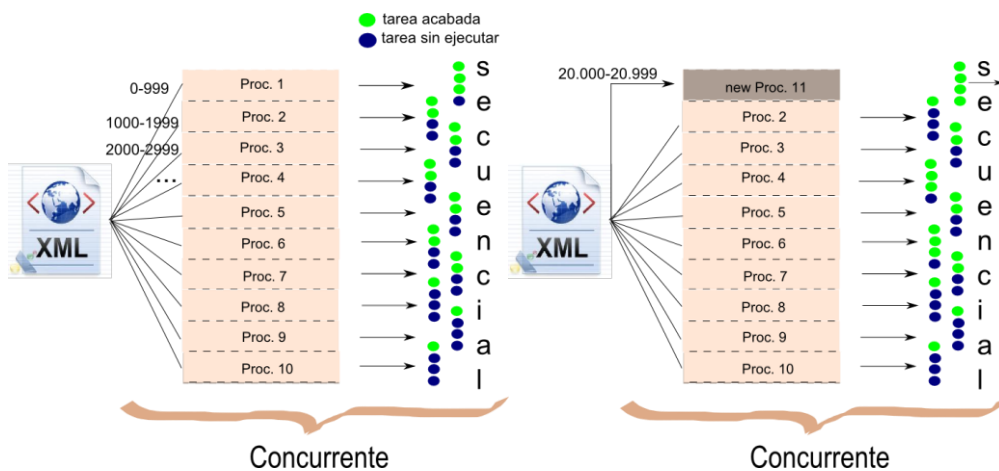


Figura 22: Pool de procesos

Gracias a esto el procesamiento se aligeraba sustancialmente aunque, visualizando el tratamiento de cada tweet, se podía contemplar como la llamada a la API de Daedalus seguía consumiendo la mayor franja de tiempo. Para propulsar una mayor velocidad se delimitó este pedazo de código que conformaba el cuello de botella y se siguió un enfoque multihilo. Para lograr este objetivo tuvo que cortarse el código secuencial en 3 partes: la anterior a dichas peticiones a la API, las propias llamadas y el resto de procesamiento abarcando hasta el final. De esta forma, el primer trozo se procesaba secuencialmente para todos los tweets cargados en un proceso y el segundo trozo lanzaba llamadas concurrentes (1 llamada para cada tweet) para acabar de nuevo procesando en línea la última parte (también para todos los tweets destinados al proceso). Esta resultó ser la única forma de añadir algo de concurrencia respetando un código puramente lineal.

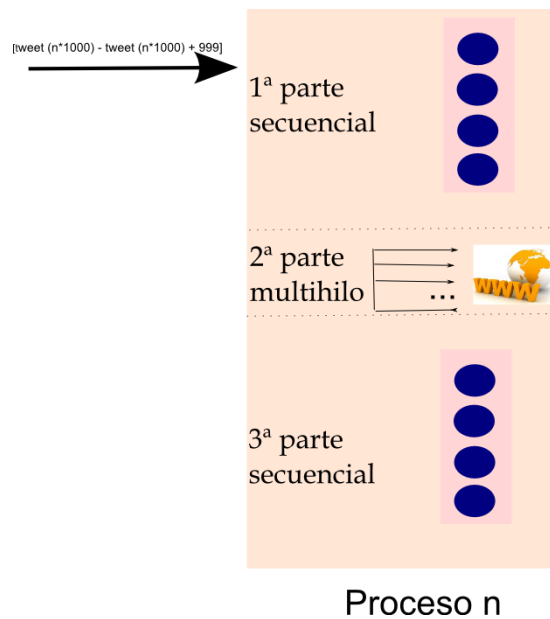


Figura 23: Proceso

Como ya he comentado en el apartado de Experimentación y Pruebas, finalmente no se pudieron aplicar las mejoras de activar la detección de entidades, con lo que el tiempo inicialmente esperado para la ejecución se redujo sustancialmente.

Dado que se plantearon 5 configuraciones, la mejor opción fue ejecutar paralelamente el código en 5 máquinas diferentes. La marca de tiempo medio obtenido fue de 10 horas por máquina.

ANEXO 3: Ejemplos sobre el nuevo corpus anotado

Para comprobar los resultados del procesamiento y del módulo de Generación de Corpus Anotado, vamos a pasar a la visualización de ejemplos en las diferentes configuraciones.

En estas anotaciones aparecen todos los síntomas susceptibles de ser buenos recolectores de semántica tanto para este enfoque como para enfoques basados en Minería. Estos datos etiquetados servirán como entrada para futuras pruebas e inclusive permitirán valorar y analizar qué características pueden ser mejores pistas para detectar ciertos fenómenos (que categoría gramatical abunda en los mensajes objetivos, cuándo se usa más la enfatización, qué grado de confianza o peso tiene un emoticono en un tweet,...) en un futuro.

Se han remarcado con diferentes colores las etiquetas más relevantes. De todas formas, se aconseja realizar un recorrido por todas las imágenes para cerciorarse del correcto funcionamiento de los algoritmos. Aun así, daré algunos breves consejos para incitar a fijarse en las características remarcables.

Baseline: al no haberse activado ningún módulo puede comprobarse como la polaridad sólo tiene en cuenta los términos que contienen polaridad previa. Es por esta razón por lo que el resultado es completamente erróneo.

```

<xml>
  <twits>
    <tweet>
      <twitid>142378325086715906</twitid>
      <user>jesusmarana</user>
      <date>2011-12-02T00:03:32</date>
      <originalContent>No me gustó la alegría que mostraba la directora del Banco Santander.
      Debería estar muuuuuy triste!!</originalContent>
      <sentences>
        <proposition>
          <level>1</level>
          <lexicalUnits>
            <token>No/Lemma:no/POSTag:RN</token>
            <token>me/Lemma:me/POSTag:PP1CS000</token>
            <token>gustó/Lemma:gustar/Synset:01777210/POSTag:VMIS3S0/PreviousPolarity:1.0
            StrongP/FinalPolarity(UPC+RL+Mod):1.0 StrongP</token>
            <nominalPhrase>
              <token>la/Lemma:el/POSTag:DA0FS0</token>
              <token>
                alegría/Lemma:alegría/Synset:05829782/POSTag:NCFS000/PreviousPolarity:0.375
                P/FinalPolarity(UPC+RL+Mod):0.375 P</token>
              <token>que/Lemma:que/POSTag:PROCN000</token>
              <relativeClause>
                <level>2</level>
                <lexicalUnits>
                  <token>mostraba/Lemma:mostrar/Synset:00943837/POSTag:VMII3S0</token>
                  <nominalPhrase>
                    <token>la/Lemma:el/POSTag:DA0FS0</token>
                    <token>directora/Lemma:director/Synset:10164233/POSTag:NCFS000</token>
                    <token>de/Lemma:de/POSTag:SPS00</token>
                    <nominalPhrase>
                      <token>el/Lemma:el/POSTag:DA0MS0</token>
                      <token>Banco Santander/Entity</token>
                      <polarity>0.0 NONE</polarity>
                    </nominalPhrase>
                    <polarity>0.0 NONE</polarity>
                  </nominalPhrase>
                  <token>./Lemma:./POSTag:Fp</token>
                </lexicalUnits>
                <repeatedLettersPercentage>0.0</repeatedLettersPercentage>
                <uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
                <finalPolarity>0.0 NONE</finalPolarity>
                <agreement>AGREEMENT</agreement>
              </relativeClause>
              <polarity>0.375 P</polarity>
            </nominalPhrase>
          </lexicalUnits>
          <repeatedLettersPercentage>0.0</repeatedLettersPercentage>
          <uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
          <finalPolarity>0.6875 StrongP</finalPolarity>
          <agreement>AGREEMENT</agreement>
        </proposition>
        <proposition>
          <level>1</level>
          <lexicalUnits>
            <token>Debería/Lemma:deber/Synset:02731814/POSTag:VMIC3S0</token>

```

```

<token>estar/Lemma:estar/Synset:02655135/POSTag:VAN0000</token>
<token>muuuuuy/StandardForm:muy/Lemma:muy/Synset:00014285/POSTag:RG/RL:57.142857
</token>
<token>triste/Lemma:triste/Synset:01361863/POSTag:AQ0CS0/PreviousPolarity:-0.625
StrongN/FinalPolarity(UPC+RL+Mod):-0.625 StrongN</token>
<token>!/Lemma:!/POSTag:Fat</token>
<token>!/Lemma:!/POSTag:Fat</token>
</lexicalUnits>
<repeatedLettersPercentage>14.0</repeatedLettersPercentage>
<uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
<finalPolarity>-0.625 StrongN</finalPolarity>
<agreement>AGREEMENT</agreement>
</proposition>
</sentences>
<TweetFinalPolarity>0.03125 NEU</TweetFinalPolarity>
<agreement>AGREEMENT</agreement>
</twit>
</twits>
</xml>

```

Configuración 1: el módulo de tratamiento de modificadores está activado, por lo tanto el modificador “muy” es detectado y la polaridad final de aquello que se ve afectado es alterada.

```

<xml>
  <twits>
    <twit>
      <twitid>142378325086715906</twitid>
      <user>jesusmarana</user>
      <date>2011-12-02T00:03:32</date>
      <originalContent>No me gustó la alegría que mostraba la directora del Banco Santander.
      Debería estar muuuuuy triste!!</originalContent>
      <sentences>
        <proposition>
          <level>1</level>
          <lexicalUnits>
            <token>No/Lemma:no/POSTag:RN</token>
            <token>me/Lemma:me/POSTag:PP1CS000</token>
            <token>gustó/Lemma:gustar/Synset:01777210/POSTag:VMIS3S0/PreviousPolarity:1.0
            StrongP/FinalPolarity(UPC+RL+Mod):1.0 StrongP</token>
            <nominalPhrase>
              <token>la/Lemma:el/POSTag:DA0FS0</token>
              <token>
                alegría/Lemma:alegría/Synset:05829782/POSTag:NCFS000/PreviousPolarity:0.375
                P/FinalPolarity(UPC+RL+Mod):0.375 P</token>
              <token>que/Lemma:que/POSTag:PR0CN000</token>
              <relativeClause>
                <level>2</level>
                <lexicalUnits>
                  <token>mostraba/Lemma:mostrar/Synset:00943837/POSTag:VMII3S0</token>
                  <nominalPhrase>
                    <token>la/Lemma:el/POSTag:DA0FS0</token>
                    <token>directora/Lemma:director/Synset:10164233/POSTag:NCFS000</token>
                    <token>de/Lemma:de/POSTag:SPS00</token>
                    <nominalPhrase>
                      <token>el/Lemma:el/POSTag:DA0MS0</token>
                      <token>Banco Santander/Entity</token>
                      <polarity>0.0 NONE</polarity>
                    </nominalPhrase>
                    <polarity>0.0 NONE</polarity>
                  </nominalPhrase>
                  <token>./Lemma:./POSTag:Fp</token>
                </lexicalUnits>
                <repeatedLettersPercentage>0.0</repeatedLettersPercentage>
                <uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
                <finalPolarity>0.0 NONE</finalPolarity>
                <agreement>AGREEMENT</agreement>
              </relativeClause>
              <polarity>0.375 P</polarity>
            </nominalPhrase>
          </lexicalUnits>
          <repeatedLettersPercentage>0.0</repeatedLettersPercentage>
          <uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
          <finalPolarity>0.6875 StrongP</finalPolarity>
          <agreement>AGREEMENT</agreement>
        </proposition>
        <proposition>
          <level>1</level>
          <lexicalUnits>
            <token>Debería/Lemma:deber/Synset:02731814/POSTag:VMIC3S0</token>

```

```

<token>estar/Lemma:estar/Synset:02655135/POSTag:VAN0000</token>
<token>
muuuuuy/StandardForm:muy/Lemma:muy/Synset:00014285/POSTag:RG/RL:57.142857/Modifier
{++}</token>
<token>triste/Lemma:triste/Synset:01361863/POSTag:AQ0CS0/PreviousPolarity:-0.625
StrongN/FinalPolarity(UPC+RL+Mod):-1.0 StrongN</token>
<token>!/Lemma:!/POSTag:Fat</token>
<token>!/Lemma:!/POSTag:Fat</token>
</lexicalUnits>
<repeatedLettersPercentage>14.0</repeatedLettersPercentage>
<uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
<polarityBeforeModifiers>-0.625 StrongN</polarityBeforeModifiers>
<finalPolarity>-1.0 StrongN</finalPolarity>
<agreement>AGREEMENT</agreement>
</proposition>
</sentences>
<TweetFinalPolarity>-0.15625 NEU</TweetFinalPolarity>
<agreement>AGREEMENT</agreement>
</twit>
</twits>
</xml>

```

Configuración 2: la detección de la negación desencadena cambios bruscos en las polaridades provocando que los resultados se acerquen más a los correctos. Este hecho se puede comprobar en las etiquetas denominadas *finalPolarity* en ambas proposiciones. También es visible la detección del elemento causante de la negación en el primer fragmento (anotación *NegationElement*).

```

<xml>
  <twits>
    <twit>
      <twitid>142378325086715906</twitid>
      <user>jesusmarana</user>
      <date>2011-12-02T00:03:32</date>
      <originalContent>No me gustó la alegría que mostraba la directora del Banco Santander.
      Debería estar muuuuuy triste!!</originalContent>
      <sentences>
        <proposition>
          <level>1</level>
          <lexicalUnits>
            <token>No/Lemma:no/POSTag:RN/NegationElement</token>
            <token>me/Lemma:me/POSTag:PP1CS000</token>
            <token>gustó/Lemma:gustar/Synset:01777210/POSTag:VMIS3S0/PreviousPolarity:1.0
            StrongP/FinalPolarity(UPC+RL+Mod):1.0 StrongP</token>
            <nominalPhrase>
              <token>la/Lemma:el/POSTag:DA0FS0</token>
              <token>
                alegría/Lemma:alegría/Synset:05829782/POSTag:NCFS000/PreviousPolarity:0.375
                P/FinalPolarity(UPC+RL+Mod):0.375 P</token>
              <token>que/Lemma:que/POSTag:PR0CN000</token>
            <relativeClause>
              <level>2</level>
              <lexicalUnits>
                <token>mostraba/Lemma:mostrar/Synset:00943837/POSTag:VMII3S0</token>
                <nominalPhrase>
                  <token>la/Lemma:el/POSTag:DA0FS0</token>
                  <token>directora/Lemma:director/Synset:10164233/POSTag:NCFS000</token>
                  <token>de/Lemma:de/POSTag:SPS00</token>
                  <nominalPhrase>
                    <token>el/Lemma:el/POSTag:DA0MS0</token>
                    <token>Banco Santander/Entity</token>
                    <polarity>0.0 NONE</polarity>
                  </nominalPhrase>
                  <polarity>0.0 NONE</polarity>
                </nominalPhrase>
                <token>./Lemma:./POSTag:Fp</token>
              </lexicalUnits>
              <repeatedLettersPercentage>0.0</repeatedLettersPercentage>
              <uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
              <finalPolarity>0.0 NONE</finalPolarity>
              <agreement>AGREEMENT</agreement>
            </relativeClause>
            <polarity>0.375 P</polarity>
          </nominalPhrase>
        </lexicalUnits>
        <repeatedLettersPercentage>0.0</repeatedLettersPercentage>
        <uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
        <finalPolarity>-0.2 N</finalPolarity>
        <agreement>AGREEMENT</agreement>
      </proposition>
    </proposition>
  </sentences>
  <level>1</level>
  <lexicalUnits>
    <token>Debería/Lemma:deber/Synset:02731814/POSTag:VMIC3S0</token>
  </lexicalUnits>
</twit>
</twits>
</xml>

```



```

<token>estar/Lemma:estar/Synset:02655135/POSTag:VAN0000</token>
<token>muuuuuy/StandardForm:muy/Lemma:muy/Synset:00014285/POSTag:RG/RL:57.142857
</token>
<token>triste/Lemma:triste/Synset:01361863/POSTag:AQ0CS0/PreviousPolarity:-0.625
StrongN/FinalPolarity(UPC+RL+Mod):-0.625 StrongN</token>
<token>!/Lemma:!/POSTag:Fat</token>
<token>!/Lemma:!/POSTag:Fat</token>
</lexicalUnits>
<repeatedLettersPercentage>14.0</repeatedLettersPercentage>
<uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
<finalPolarity>-0.625 StrongN</finalPolarity>
<agreement>AGREEMENT</agreement>
</proposition>
</sentences>
<TweetFinalPolarity>-0.4125 N</TweetFinalPolarity>
<agreement>AGREEMENT</agreement>
</twit>
</twits>
</xml>

```

Configuración 3: El tratamiento de los estilos en la comunicación provoca que la proposición subordinada pase de una polaridad a la ausencia total de sentimiento (NONE). La primera está marcada como *polarityAsIndependentProposition* y se ve afectada por la detección del estilo indirecto en su proposición principal. Obviamente, al mantener el módulo de la negación desactivado, la polaridad no es la correcta.

```

<xml>
  <twits>
    <twit>
      <twitid>142378325086715906</twitid>
      <user>jesumarana</user>
      <date>2011-12-02T00:03:32</date>
      <originalContent>Obama asistió ayer al Congreso y dijo que no le gustaba la actuación
de nuestro Gobierno.</originalContent>
      <sentences>
        <proposition>
          <CoordinateProposition>
            <level>1</level>
            <lexicalUnits>
              <token>Obama/Entity</token>
              <token>asistió/Lemma:asistir/Synset:02612762/POSTag:VMIS3S0</token>
              <token>ayer/Lemma:ayer/POSTag:RG</token>
              <groupSP>
                <token>a/Lemma:a/POSTag:SPS00</token>
                <nominalPhrase>
                  <token>el/Lemma:el/POSTag:DA0MS0</token>
                  <token>Congreso/Entity</token>
                  <polarity>0.0 NONE</polarity>
                </nominalPhrase>
                <polarity>0.0 NONE</polarity>
              </groupSP>
              <token>y/Lemma:y/POSTag:CC</token>
            </lexicalUnits>
          </CoordinateProposition>
          <level>1</level>
          <lexicalUnits>
            <token>dijo/Lemma:decir/Synset:00941990/POSTag:VMIS3S0</token>
            <token>que/Lemma:que/POSTag:CS</token>
          </lexicalUnits>
          <repeatedLettersPercentage>0.0</repeatedLettersPercentage>
          <uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
          <SubordinateProposition>
            <level>2</level>
            <lexicalUnits>
              <token>no/Lemma:no/POSTag:RN</token>
              <token>le/Lemma:le/POSTag:PP3CSD00</token>
              <token>
gustaba/Lemma:gustar/Synset:01777210/POSTag:VMII3S0/PreviousPolarity:1.0
StrongP/FinalPolarity(UPC+RL+Mod):1.0 StrongP</token>
              <nominalPhrase>
                <token>la/Lemma:el/POSTag:DA0FS0</token>
                <token>actuación/Lemma:actuación/Synset:09765278/POSTag:NCFS000</token>
                <token>de/Lemma:de/POSTag:SPS00</token>
              </nominalPhrase>
              <token>nuestro/Lemma:nuestro/POSTag:DP1MSP</token>
              <token>Gobierno/Entity</token>
              <token>./Lemma:./POSTag:Fp</token>
              <polarity>0.0 NONE</polarity>
            </nominalPhrase>
            <polarity>0.0 NONE</polarity>
          </nominalPhrase>
        </lexicalUnits>
      </sentences>
    </twit>
  </twits>
</xml>

```

```

    <repeatedLettersPercentage>0.0</repeatedLettersPercentage>
    <uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
    <polarityBeforeModifiers>1.0 StrongP</polarityBeforeModifiers>
    <polarityAsIndependentProposition>1.0 StrongP</polarityAsIndependentProposition>
    <finalPolarity>0.0 NONE</finalPolarity>
    <agreement>AGREEMENT</agreement>
  </SubordinateProposition>
  <finalPolarity>0.0 NONE</finalPolarity>
  <agreement>AGREEMENT</agreement>
</CoordinateProposition>
<finalPolarity>0.0 NONE</finalPolarity>
<agreement>AGREEMENT</agreement>
<finalPolarity/>
<agreement>AGREEMENT</agreement>
</proposition>
</sentences>
<TweetFinalPolarity>0.0 NONE</TweetFinalPolarity>
<agreement>AGREEMENT</agreement>
</twit>
</twits>
</xml>

```

Configuración 4: Una vez que activamos todos los módulos, la solución final a la categorización es la siguiente (pudiéndose comprobar como la suma del tratamiento de modificadores junto con la negación arrojan un resultado negativo pero más extremo que con la configuración 2):

```

<xml>
  <twits>
    <twit>
      <twitid>142378325086715906</twitid>
      <user>jesusmarana</user>
      <date>2011-12-02T00:03:32</date>
      <originalContent>No me gustó la alegría que mostraba la directora del Banco Santander.
      Debería estar muuuuuy triste!!</originalContent>
      <sentences>
        <proposition>
          <level>1</level>
          <lexicalUnits>
            <token>No/Lemma:no/POSTag:RN/NegationElement</token>
            <token>me/Lemma:me/POSTag:PP1CS000</token>
            <token>gustó/Lemma:gustar/Synset:01777210/POSTag:VMIS3S0/PreviousPolarity:1.0
            StrongP/FinalPolarity(UPC+RL+Mod):1.0 StrongP</token>
          <nominalPhrase>
            <token>la/Lemma:el/POSTag:DA0FS0</token>
            <token>
              alegría/Lemma:alegría/Synset:05829782/POSTag:NCFS000/PreviousPolarity:0.375
              P/FinalPolarity(UPC+RL+Mod):0.375 P</token>
            <token>que/Lemma:que/POSTag:PR0CN000</token>
          <relativeClause>
            <level>2</level>
            <lexicalUnits>
              <token>mostraba/Lemma:mostrar/Synset:00943837/POSTag:VMII3S0</token>
            <nominalPhrase>
              <token>la/Lemma:el/POSTag:DA0FS0</token>
              <token>directora/Lemma:director/Synset:10164233/POSTag:NCFS000</token>
              <token>de/Lemma:de/POSTag:SPS00</token>
            <nominalPhrase>
              <token>el/Lemma:el/POSTag:DA0MS0</token>
              <token>Banco Santander/Entity</token><polarity>0.0 NONE</polarity>
            </nominalPhrase>
            <polarity>0.0 NONE</polarity>
          </nominalPhrase>
          <token>./Lemma:./POSTag:Fp</token>
        </lexicalUnits>
        <repeatedLettersPercentage>0.0</repeatedLettersPercentage>
        <uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
        <finalPolarity>0.0 NONE</finalPolarity>
        <agreement>AGREEMENT</agreement>
      </relativeClause>
      <polarity>0.375 P</polarity>
    </nominalPhrase>
  </lexicalUnits>
  <repeatedLettersPercentage>0.0</repeatedLettersPercentage>
  <uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
  <finalPolarity>-0.2 N</finalPolarity>
  <agreement>AGREEMENT</agreement>
</proposition>
<proposition>
  <level>1</level>
  <lexicalUnits>
    <token>Debería/Lemma:deber/Synset:02731814/POSTag:VMIC3S0</token>
    <token>estar/Lemma:estar/Synset:02655135/POSTag:VAN0000</token>
  </lexicalUnits>
</proposition>

```

```
<token>
muuuuuy/StandardForm:muy/Lemma:muy/Synset:00014285/POSTag:RG/RL:57.142857/Modifier
{++}</token>
<token>triste/Lemma:triste/Synset:01361863/POSTag:AQ0CS0/PreviousPolarity:-0.625
StrongN/FinalPolarity(UPC+RL+Mod):-1.0 StrongN</token>
<token>!/Lemma:!/POSTag:Fat</token>
<token>!/Lemma:!/POSTag:Fat</token>
</lexicalUnits>
<repeatedLettersPercentage>14.0</repeatedLettersPercentage>
<uppercaseLettersPercentage>0.0</uppercaseLettersPercentage>
<polarityBeforeModifiers>-0.625 StrongN</polarityBeforeModifiers>
<finalPolarity>-1.0 StrongN</finalPolarity>
<agreement>AGREEMENT</agreement>
</proposition>
</sentences>
<TweetFinalPolarity>-0.6 StrongN</TweetFinalPolarity>
<agreement>AGREEMENT</agreement>
</twit>
</twits>
</xml>
```

