Minería de textos y análisis de sentimientos en ***sanidadysalud.com***

Proyecto Final de Master

Autor: Sergio Rincón García

Director: Antonio Pareja-Lora

Máster en Minería de Datos e Inteligencia de Negocios, 2015 - 2016

Universidad Complutense de Madrid

Facultad de Estudios Estadísticos
Avda. Puerta de Hierro s/n
Ciudad Universitaria
28040-MADRID

Contenido

1. Introducción

Si atendemos a la segunda definición de *sentimiento* incluida en la Real Academia Española de la Lengua[1] (véase la Ilustración 1), observamos que es un estado del ánimo alterado por causas determinadas.

Sentimiento.
1. *m. Acción y efecto de sentir o sentirse.*
2. *m. Estado afectivo del ánimo producido por causas que lo impresionan vivamente.*
3. *m. Estado del ánimo afligido por un suceso triste o doloroso.*

Ilustración 1: Definición de sentimiento (RAE, 2015)

En particular, las opiniones leídas o escuchadas pueden producir cambios en las creencias y en los estados afectivos, es decir, producir un sentimiento. De hecho, cuando tenemos que tomar una decisión somos nosotros mismos los que buscamos otras opiniones para dar soporte a la misma.

Tradicionalmente hemos buscado las opiniones de nuestro círculo más cercano compuesto por familiares y conocidos; sin embargo, en la era de la globalización, nuestro círculo se ha ampliado enormemente y ahora, para cada decisión a la que nos enfrentamos, podemos encontrar (por ejemplo, en la web) miles de opiniones para consultar antes de tomar la decisión.

Dichas opiniones habitualmente se componen de contenido subjetivo y están asociadas a sentimientos, que pueden ser positivos, negativos o neutros, en función de las creencias o experiencias de cada usuario.

Un claro ejemplo de ello, lo encontramos a la hora de hacer una compra a través de comercio electrónico, ya que es muy habitual revisar las opiniones asociadas al objeto de la compra (véase la Ilustración 2).

[1] http://dle.rae.es/?id=XbTu91V (Recuperado el 16 de diciembre de 2015)

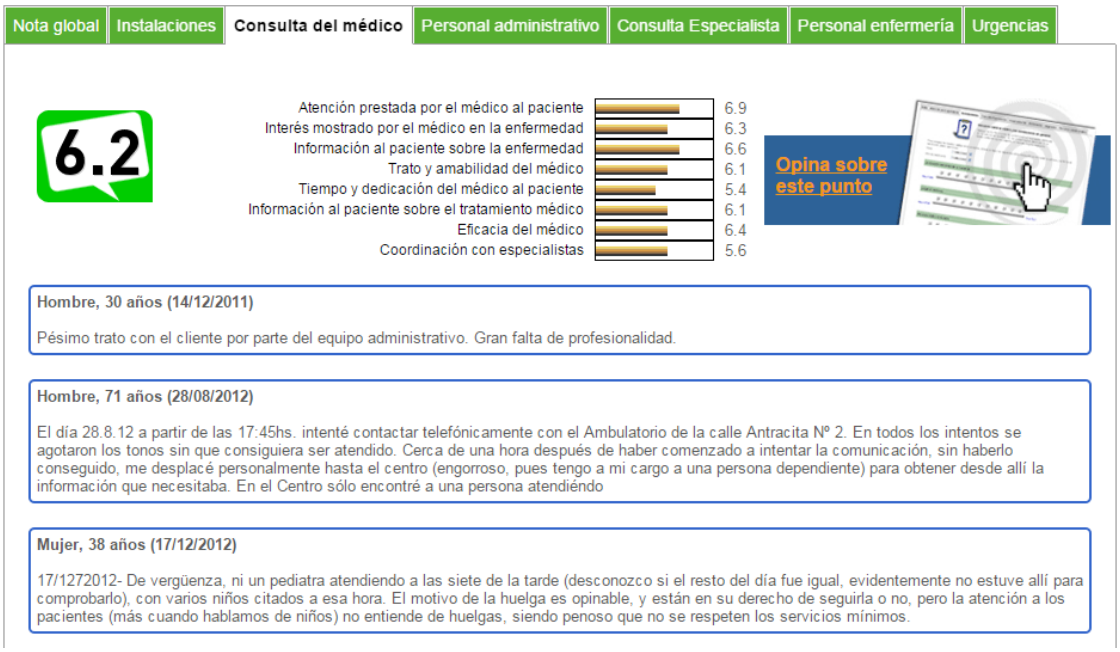
Las opiniones de cliente más útiles

6 de 7 personas piensan que la opinión es útil
Excelente Edición
Por Euriloco en 20 de octubre de 2012
Formato: Versión Kindle
Una edición perfecta con abundantes notas a pie de página que ayudan a entender la lectura.
Incluye los grabados de Doré que añaden al ebook más calidad.
Es sin duda un ejemplo a seguir por otras editoriales.
¿Esta opinión te ha parecido útil? Sí No Informar de un abuso

4 de 5 personas piensan que la opinión es útil
Buena edición, exceptos por las notas
Por Felipe Santa-Cruz (escritor) en 23 de septiembre de 2013
Formato: Versión Kindle | Compra verificada
La edición, muy buena. Y leer el Quijote en Kindle es una gozada. La primera vez que leí esta obra casi tenía que usar una muñequera metacarpiana para paliar el dolor que me producía sostener cada tomo.
Lo único que no me ha terminado de convencer son las notas. Por dos motivos:
1. En muchas ocasiones desvelan lo que va a ocurrir, lo cual le llena a uno de rabia y de pensamientos homicidas.
2. Cuando ya una palabra o algún punto susceptible de ser comentado ha sido anotado anteriormente y se vuelve a repetir, la editorial opta por introducir una nota que nos señala que se habló de ello en tal o cual capítulo. Y, claro, si no lo recuerdas (que es lo normal en un libro de tal envergadura), tienes que ir al índice del libro, buscar el capítulo y encontrar la nota. Esto, con la tecnología que ofrece el formato eBook, es un despropósito. Lo más lógico sería que enlazaran con la nota primitiva.
¿Esta opinión te ha parecido útil? Sí No Informar de un abuso

Ilustración 2: Opinión sobre un libro recogida en www.amazon.es

Otro ámbito típico de consulta de opiniones, es el sector sanitario, donde es habitual consultar las opiniones recogidas en Internet sobre los centros o profesionales sanitarios antes de contratar sus servicios (véase Ilustración 3).



Nota global | Instalaciones | **Consulta del médico** | Personal administrativo | Consulta Especialista | Personal enfermería | Urgencias

6.2

Atención prestada por el médico al paciente 6.9
Interés mostrado por el médico en la enfermedad 6.3
Información al paciente sobre la enfermedad 6.6
Trato y amabilidad del médico 6.1
Tiempo y dedicación del médico al paciente 5.4
Información al paciente sobre el tratamiento médico 6.1
Eficacia del médico 6.4
Coordinación con especialistas 5.6

Opina sobre este punto

Hombre, 30 años (14/12/2011)
Pésimo trato con el cliente por parte del equipo administrativo. Gran falta de profesionalidad.

Hombre, 71 años (28/08/2012)
El día 28.8.12 a partir de las 17:45hs. intenté contactar telefónicamente con el Ambulatorio de la calle Antracita Nº 2. En todos los intentos se agotaron los tonos sin que consiguiera ser atendido. Cerca de una hora después de haber comenzado a intentar la comunicación, sin haberlo conseguido, me desplacé personalmente hasta el centro (engorroso, pues tengo a mi cargo a una persona dependiente) para obtener desde allí la información que necesitaba. En el Centro sólo encontré a una persona atendiéndo

Mujer, 38 años (17/12/2012)
17/1272012- De vergüenza, ni un pediatra atendiendo a las siete de la tarde (desconozco si el resto del día fue igual, evidentemente no estuve allí para comprobarlo), con varios niños citados a esa hora. El motivo de la huelga es opinable, y están en su derecho de seguirla o no, pero la atención a los pacientes (más cuando hablamos de niños) no entiende de huelgas, siendo penoso que no se respeten los servicios mínimos.

Ilustración 3: Opinión sobre servicios sanitarios recogida en www.sanidadysalud.com

Más allá de nuestro punto de vista como individuos, desde una perspectiva empresarial, monitorizar y analizar las opiniones de los usuarios o clientes se convierte en algo casi esencial en un entorno de alta competitividad. Sin embargo, en mayor o menor medida, las empresas se enfrentan al problema de analizar un gran volumen de opiniones relativas a productos o servicios, propios o de la competencia, y de múltiples fuentes, como pueden ser blogs, artículos de opinión o redes sociales; lo que hace inviable abordar manualmente el análisis de esas opiniones, siendo necesario establecer procesos automáticos que se encarguen de la tarea.

En este entorno, el **Procesamiento del Lenguaje Natural (PLN),** como subrama de la Inteligencia Artificial y la Lingüística Computacional, cobra una relevancia clave, con el objetivo de extraer masivamente el significado residente en el lenguaje natural o humano.

Es por ello que, sin pérdida de generalidad, en este trabajo se propone una aproximación para recoger y clasificar los sentimientos[2] de las opiniones asociadas a los centros sanitarios españoles, recogidas en el portal web www.sanidadysalud.com, por medio de técnicas de minería de textos y análisis de sentimientos.

Mediante dichas técnicas, los textos con opiniones dejan de ser datos desestructurados para convertirse en datos estructurados, aptos para su posterior análisis y clasificación. De esta forma, pueden ser integrados en los análisis cuantitativos habituales de cualquier negocio.

En particular, los análisis y modelos desarrollados en el presente trabajo tienen como propósito conseguir un agente software con el que **automatizar la revisión de las opiniones de los usuarios que se publican en el mencionado portal web**. Actualmente, dicha tarea se está efectuando de forma manual, con el fin de evitar la publicación de opiniones que, o bien no cumplen el propósito general del portal web, o bien son lesivas para el honor de los profesionales que trabajan en los centros sanitarios.

Este documento consta de **cuatro secciones adicionales a la presente,** más la bibliografía consultada. En la siguiente sección se hará un repaso del **estado del arte** de la minería de textos y opiniones. A continuación, se fijarán el **ámbito, el alcance y los objetivos** del proyecto, incluyendo sus criterios de éxito. Después se entrará en el **desarrollo** propiamente del sistema informático, describiendo cada una de las fases que lo componen desde un punto de vista de la metodología CRISP-DM. Finalmente, se hará una valoración general del proyecto a modo de **conclusión** y se sentarán las bases de los trabajos futuros.

2. Estado del Arte

Desde el punto de vista del análisis cuantitativo de datos, el lenguaje natural que usamos los humanos para comunicarnos a menudo es incluido dentro de la categoría de “datos no estructurados”. Por ello, la mayoría de las subtareas incluidas en el Procesamiento el Lenguaje Natural (PLN) incluyen la conversión de los “datos no estructurados” en “datos estructurados”.

En particular, si nos restringimos al lenguaje natural escrito, la **minería de textos** se refiere al conjunto de procesos que consiguen extraer significado (datos estructurados) de un conjunto de textos (datos no estructurados)

Como ejemplo de algunas tareas típicas de la minería de textos tenemos (*SemEval 2015, 2015;* véanse también, en el apartado de Referencias, las entradas de Wikipedia relativas a *Document classification, Automatic summarization, Named entity recognition y Sentiment Analysis*):

- **Categorización de textos** (*Text categorization/Document classification*)

  Busca asignar un documento, de forma genérica, o un texto, de forma particular, a una o más categorías o clases.

  Algunas de sus aplicaciones serían: filtrado de spam, identificación del idioma, análisis de sentimientos, etc.

[2] Entendidos como el posicionamiento de quién expresa su opinión respecto a un producto o servicio.

- **Similaridad semántica de textos** (*Semantic Text similarity*)

  Captura el grado de similaridad semántica de dos porciones de texto.

  Una de sus principales aplicaciones son los buscadores de internet, dentro del campo de Búsqueda y Recuperación de Información.

- **Resumen de documentos** (*Document summarization*)

  Intenta encontrar un conjunto de datos representativo de todo el documento, identificando para ello sus partes más informativas. El problema también se puede generalizar para hacer un resumen de múltiples documentos.

  De nuevo, una de sus principales aplicaciones son los buscadores de internet (por ejemplo la creación del *PageRank* de las webs más representativas de un tema). No obstante, también es habitual utilizar esta técnica para elegir imágenes representativas de una colección; o incluso a nivel de fotogramas de un vídeo.

- **Extracción de conceptos o entidades** (*Named entity recognition*)

  Busca localizar y clasificar los elementos de un texto dentro de unas categorías predefinidas, como pueden ser nombres de personas, nombres de empresas, lugares, porcentajes, expresiones de tiempo, etc.

  Una aplicación actual de la técnica es la identificación de entidades o expresiones importantes en los textos para crear vínculos directamente con *Wikipedia*[3] *(entity linking)*.

- **Desambiguación del significado** (*Word Sense Disambiguation*)

  Aborda el problema de seleccionar el sentido en el que una palabra es usada en una frase, cuando la palabra tiene múltiples significados (por ejemplo, banco puede referirse a banco de peces, al banco como entidad financiera, etc.)

  De nuevo, una de las principales aplicaciones actuales son los motores de búsqueda.

- **Análisis de sentimientos** (*Sentiment Analysis/Opinion mining*)

  Tiene como objetivo determinar la polaridad general de un documento, una frase o un aspecto del mismo, tratando de detectar la actitud del creador en base a las posibles emociones, juicios o evaluaciones contenidas en el documento. Las etiquetas más extendidas para clasificar la polaridad son: positiva, negativa o neutra.

  Con el crecimiento de los medios de comunicación social (web 2.0) en Internet, las aplicaciones del Análisis de Sentimientos cobran especial relevancia, fundamentalmente desde el punto de vista empresarial, de cara a analizar todo tipo de expresiones online: análisis de revisiones y opiniones de productos, gestión de la reputación, identificación de nuevas oportunidades de negocio, etc.

Desde el año 2000 *(Liu, 2012)*, la investigación sobre el análisis de opiniones personales y sentimientos ha crecido de forma sensible, debido fundamentalmente al gran número de campos de aplicación existentes en la actualidad, así como al crecimiento exponencial de opiniones recogidas en Internet (redes sociales, blogs, foros, etc.) que hacen posible disponer

[3] www.wikipedia.org

de datos suficientes para avanzar en la investigación. Sin embargo, a día de hoy es un campo abierto con múltiples retos por resolver, al igual que el resto de ramas del Procesamiento del Lenguaje Natural con las que está íntimamente relacionado: extracción de entidades y conceptos, desambiguación del significado, resumen de documentos, etc.

La simple tarea de clasificar el sentimiento de un texto como positivo o negativo es a veces tan complicada, que diferentes personas pueden no ponerse de acuerdo para otorgarle una clasificación definitiva. Esto es debido a que un mismo texto puede ser interpretado de forma diferente en función de factores culturales, de dominio, de idioma o incluso personales. Es decir, las opiniones y los sentimientos son ***subjetivos.*** Es por ello que, para poder extraer información objetiva lo primero es disponer de un conjunto suficientemente amplio de opiniones.

Además, debido a esa subjetividad, la clasificación de sentimientos es muy dependiente del dominio de las opiniones o documentos, y de hecho el principal reto abierto es conseguir un buen rendimiento de clasificadores independientes del dominio o generales.

La Clasificación de Sentimientos probablemente sea la tarea más estudiada dentro del análisis de sentimientos. Su formulación habitual es un problema de clasificación con dos clases, ***positiva*** y ***negativa.***

Este problema de clasificación, como otros tantos, puede abordarse con **aprendizaje automático supervisado o no supervisado.**

El **enfoque supervisado** parte de un **conjunto de datos anotado con la polaridad del documento**. Esta anotación de cada documento puede hacerse de forma automática, por ejemplo, cuando los documentos llevan asociada una evaluación numérica, como ocurre en las opiniones de productos en los portales de compra en Internet, o puede hacerse de forma manual por medio de anotadores humanos.

La ventaja de la anotación automática de documentos es que se puede hacer de forma masiva, y de esta forma conseguir un mayor conjunto de datos para entrenar el algoritmo. Sin embargo, esta anotación pierde precisión frente a la anotación manual. Por el contrario, la anotación manual gana en precisión, pero pierde en capacidad anotadora, por lo que el conjunto de datos para entrenamiento tiende a ser más pequeño.

Debido a que es, en definitiva, un problema de clasificación de texto *(Liu, 2012)*, cualquier método de aprendizaje automático supervisado puede ser utilizado, aunque, *Support Vector Machines (SVM)* o *Naive Bayes* se encuentran entre los más populares.

Pang, Lee y Vaithyanathan *(2002)* fue el primer estudio que aplicó este enfoque para clasificar opiniones de películas en dos clases, positiva y negativa. En este caso, los textos eran representados como vectores de palabras (*unigramas*) que, junto con sus frecuencias, se usaban como variables independientes para la clasificación, que era efectuada con *SVM* y *Naive Bayes*.

Posteriormente, se han seguido otras muchas líneas de investigación en cuanto a otros algoritmos de aprendizaje automático, y sobre todo, a diferentes maneras de formar un conjunto de variables independientes, o características, efectivo, ya que esto se ha demostrado que es clave *(Liu, 2012)* para el éxito de la clasificación. Por ejemplo, como variables independientes habitualmente se usan:

- Términos del documento junto con su frecuencia y/o su posición y/o su relevancia medida en términos de su valor *TF-IDF (Term frequency – Inverse document frequency)* o calculada mediante técnicas estadísticas.
- Categorías gramaticales de las palabras: adjetivo, nombre, verbo, etc.
- Expresiones y palabras con sentimiento directo asociado: *bueno*, *malo*, *horrible*, *cuesta un ojo de la cara*, etc.
- Palabras modificadoras y/o inversoras de la polaridad: *no*, *nunca, poco, nada, nadie, podría, debería, apenas,* etc.

Si, por el contrario, atendemos al **enfoque no supervisado o semántico** *(Turney, 2002)* fue una de las primeras aproximaciones al respecto. Básicamente, la idea consistía en comparar las palabras consecutivas (bigramas) con patrones sintácticos prefijados, por ser los más proclives a ser usados para expresar opiniones. Estos patrones se formaban en base a la categoría gramatical de las palabras. Después, a estos bigramas se les asociaba una orientación semántica o polaridad en base a su distancia a la palabra positiva "excelente" y a su distancia a la palabra negativa "pobre", y por último se calculaba la polaridad del documento como la media de todas las polaridades conseguidas para los bigramas.

Otros enfoques más modernos están basados en el uso de ***lexicones***, consistentes en una colección de términos y frases anotados con su polaridad e intensidad. En este caso, la clave para realizar la clasificación consiste en la identificación de dichos términos en el documento; a partir de ahí, se completa la técnica con un tratamiento de las palabras que modifican la polaridad o directamente la invierten. Una vez hecho esto, a grandes rasgos, se computa la polaridad general del documento en función de las polaridades de los términos encontrados.

Por sus características, **tanto el enfoque supervisado como el no supervisado tienen sus pros y sus contras**; es por ello que ambas líneas de investigación siguen abiertas.

Además del método de aprendizaje automático utilizado, hoy en día se pueden **considerar tres niveles de profundidad** *(Liu, 2012)* **para el análisis de sentimientos:**

- **Documento:** El problema que se aborda en este nivel es como **clasificar la opinión general del documento**. Éste era el enfoque, en los trabajos iniciales de *Pang, Lee y Vaithyanathan, (2002) y* Turney (2002). Se asume que cada documento expresa opiniones sobre una única entidad u objeto, así como que cada opinión es emitida por un solo emisor o autor. Por ejemplo, las opiniones sobre un producto en concreto. Por ello, cuando se comparan múltiples entidades, este enfoque es insuficiente.
- **Frase:** En este nivel de análisis se **considera cada frase como una unidad** independiente y se asume que cada frase sólo debería contener una opinión. Esta tarea está muy relacionada con otra denominada Clasificación de la Subjetividad *(Wiebe, Bruce y O'Hara, 1999)*, consistente en determinar si una frase es objetiva o subjetiva.
- **Entidad y Aspecto:** Este es el nivel de análisis de grano más fino de las líneas de investigación actuales, ya que es el que consigue extraer más información de las opiniones. En vez de atender a las construcciones de lenguaje (documentos, párrafos, frases, etc.) se centra en la opinión directamente, bajo la idea de que **una opinión está compuesta de un sentimiento (positivo o negativo) y un objetivo.** La forma habitual de representar estos objetivos es por medio de **entidade**s y/o sus **atributos, que juntos forman el aspecto a analizar** y serán la mínima unidad con polaridad**.** Por ejemplo, la opinión *"por teléfono es casi imposible contactar con el centro, por internet muy bien*" se podría referir a los atributos "contacto telefónico" y "contacto por internet" de la entidad "personal administrativo", teniendo una polaridad negativa el

primero y positiva el segundo, respectivamente. Así, este problema de análisis típicamente se subdivide en varias subtareas: (i) identificar las entidades y atributos, (ii) clasificar sus polaridades y (iii) clasificar la polaridad general. Dichas subtareas siguen siendo problemas abiertos dentro del Procesamiento del Lenguaje Natural, es por lo que este nivel de análisis es el que plantea mayores retos actualmente.

3. Ámbito y objetivos

En el presente apartado principalmente se hace una revisión del ámbito o dominio de actuación del presente trabajo, a la vez que se fija su alcance. Además, se detallan los objetivos que persigue el mismo, incluyendo los criterios de éxito del proyecto.

3.1. Dominio y alcance

El dominio seleccionado para realizar el presente trabajo sobre minería de textos y análisis de sentimientos son las **opiniones** sobre centros de salud, hospitales y farmacias españolas **recogidas** en el portal web ***sanidadysalud.com.***

Dicho portal web tiene como propósito recoger las opiniones de los usuarios de los centros y servicios sanitarios, tanto para promocionar los puntos fuertes como para identificar los aspectos susceptibles de mejora de dichos centros y servicios (véase la Ilustración 4).

Ilustración 4: Portal web sanidadysalud.com

Para ello, el portal web pone a disposición de los usuarios una serie de cuestionarios de satisfacción o encuestas sobre diferentes aspectos de los centros y servicios sanitarios, que pueden ser completados libremente.

La configuración de dichos cuestionarios (véase la Ilustración 5) está diseñada en función del tipo de centro al que están asociados. Así, por ejemplo, si un usuario quiere completar un cuestionario de satisfacción sobre su hospital de referencia, el cuestionario que aparecerá estará compuesto por indicadores diferentes a los que aparecerían si el objeto de su opinión fuera su farmacia de referencia.

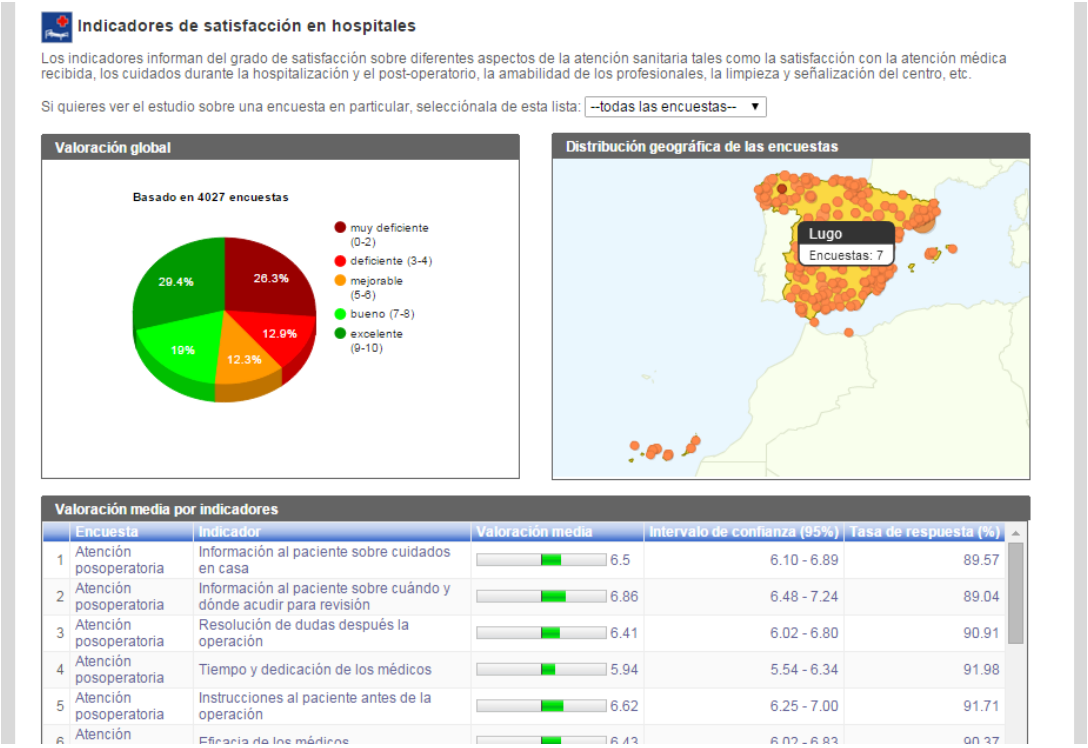


Ilustración 5: Indicadores contenidos en los cuestionarios de satisfacción de un Hospital

Es por ello que, para acceder a los cuestionarios, los usuarios primero deben seleccionar el centro sobre el que quieren expresar su opinión. Una vez en la ficha de ese centro, el usuario puede completar los cuestionarios que tiene asociados el centro, correspondientes a su tipo (véase la Ilustración 6).

Ilustración 6: Ficha de un centro de tipo Hospital

Como pueden verse en la Ilustración 7, cada **cuestionario** está compuesto por:

- **Preguntas socio-demográficas**.
- Preguntas o **Indicadores de calidad del servicio recibido**, con respuestas equidistantes en escala *Likert* de 0 a 10, según el nivel de conformidad con la pregunta.
- Un **campo de texto, para expresar una opinión abierta**.

En concreto, este campo de texto de los cuestionarios, donde se recoge la opinión abierta de los usuarios, será el que se utilizará en el presente trabajo para realizar la minería de textos y el análisis de sentimientos.

Por último, cabe reseñar que los cuestionarios del mencionado portal también presentan una serie de restricciones[4] a tener en cuenta, que en resumen son:

- El usuario debe responder a, al menos, un indicador de calidad para que el cuestionario se considere completado y válido.
- Un usuario sólo puede completar un único cuestionario del mismo tipo para un mismo centro por semana.
- Las **opiniones de texto tienen un límite máximo de 500 caracteres**.

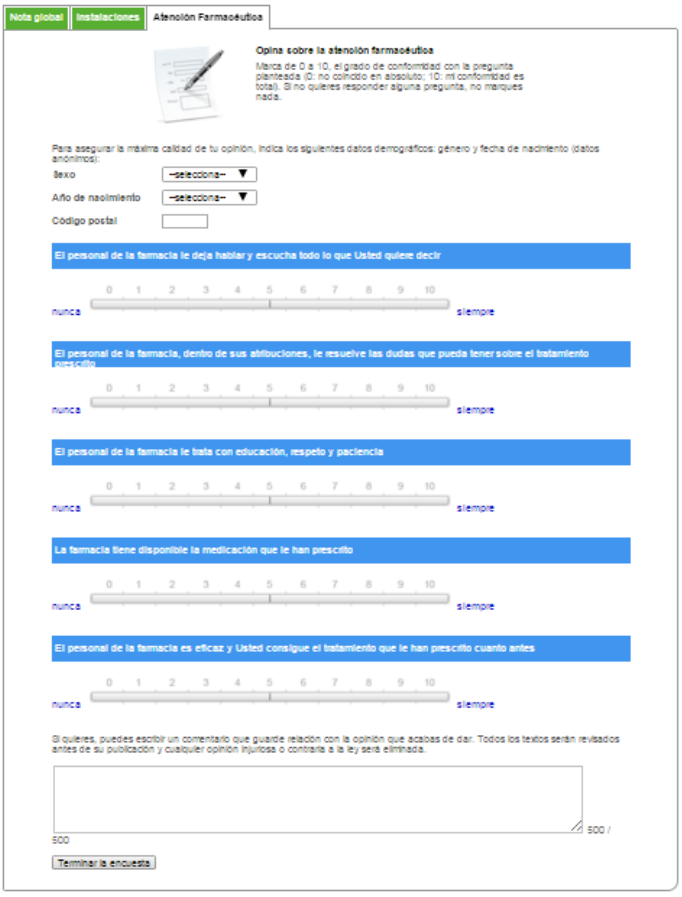


Ilustración 7: Cuestionario de satisfacción relativo a la Atención farmacéutica para un centro de tipo Farmacia

[4] Referencia: http://www.sanidadysalud.com/estudio-de-confianza-metodologia.php

3.2. Hipótesis de trabajo

Generalmente las opiniones de cualquier tipo de usuario tienen una fuerte carga de sentimientos y/o de subjetividad. El ámbito de internet, por supuesto, no es una excepción; y, en particular, las opiniones recogidas en el portal web ***sanidadysalud.com*** presentan las mismas características. De esta forma, para **evitar la publicación de opiniones negativas no constructivas o lesivas** para el honor de los profesionales sanitarios que trabajan en los centros, **actualmente todas las opiniones son revisadas manualmente** antes de ser publicadas.

En este contexto, se plantea como **hipótesis de trabajo principal que, mediante técnicas de minería de textos y análisis de sentimientos se puede construir un sistema automático que clasifique correctamente en negativas y no negativas las opiniones escritas en castellano en *sanidadysalud.com* sobre centros sanitarios españoles.**

De esta forma, las opiniones clasificadas como no negativas pueden ser publicadas automáticamente, dejando la revisión manual exclusivamente para las clasificadas como negativas.

Sin embargo, para conseguirlo, el sistema primero debe hacer frente al **reto de diferenciar las opiniones** recogidas que entran dentro de propósito general del portal, y por tanto, son **relevantes para su publicación**. Serán éstas, las opiniones relevantes, las que deberá **clasificar como negativas o no negativas** para su revisión.

3.3. Sistema de evaluación y criterios de éxito

En primer lugar, cabe resaltar que los datos disponibles de *sanidadysalud.com* abarcan desde el año 2010, año en el que se puso producción el sistema de cuestionarios, hasta el momento actual. En particular, la extracción de datos tuvo lugar el día 3 de Febrero de 2016.

Por tanto, se propone **dividir en base al año** el conjunto de datos para entrenar y evaluar los modelos conseguidos. Así, **el conjunto de datos o colección de textos que servirá para evaluar el rendimiento de los modelos conseguidos será el formado por las opiniones escritas en el año 2015 o 2016.**

Para monitorizar el rendimiento del sistema, y sus subsistemas, y la consecución de los objetivos, se utilizarán las siguientes métricas típicas de los problemas de clasificación binaria (Sokolova y Lapalme, 2009):

- **Precisión (*Precision*)** = $\frac{tp}{tp+fp}$
- **Cobertura (*Recall*)** = $\frac{tp}{tp+fn}$
- **Exactitud (*Accuracy*)** = $\frac{tp+tn}{tp+fp+tn+fn}$
- **Medida F_β (*F_β Measure*)** = $(1+\beta^2)\frac{precision*recall}{\beta^2*precision+recall}$
- **Tasa de Verdaderos Positivos (*tpr*)** = $\frac{tp}{tp+fn}$
- **Tasa de Falsos Positivos (*fpr*)** = $\frac{fn}{tp+fn}$
- **Área Bajo la Curva (*AUC*)** = $\frac{1}{2}\left(\frac{tp}{tp+fn}+\frac{tn}{tn+fp}\right)$

donde ***tp*** es el número de **observaciones de la clase positiva clasificadas correctamente**, ***tn*** el número de **observaciones de la clase negativa clasificadas correctamente**, ***fp*** las observaciones de la clase negativa clasificadas incorrectamente y ***fn*** las observaciones de la clase positiva clasificadas incorrectamente.

3.3.1. Clasificador de Relevancia

En primer lugar, como hemos comentado en el punto anterior, el sistema deberá **clasificar las opiniones recogidas en relevantes o no relevantes**, según si cumplen el propósito general de *sanidadysalud.com.* Éste es un paso importante para asegurar la calidad del análisis de sentimientos posterior, ya que en este paso se filtrará la información no relevante o no publicable y que, por tanto, añadiría ruido al clasificador de sentimientos.

La hipótesis inicial es que la mayoría de las opiniones son relevantes, representando las opiniones no relevantes un porcentaje muy pequeño. Por tanto, probablemente el punto de corte de probabilidad para determinar si una opinión es considerada relevante o no relevante necesitará ser optimizado. Por ello**, la métrica propuesta para determinar el rendimiento del clasificador de relevancia es el AUC, y se considerará que el primer objetivo estará cumplido si se consigue alcanzar un AUC de 0.9 en la colección de textos de evaluación.**

3.3.2. Clasificador de Polaridad

Respecto al análisis de sentimientos, éste se hará exclusivamente sobre las opiniones clasificadas como relevantes en el paso anterior.

Este subsistema debe servir para **maximizar la detección de las opiniones con polaridad negativa, minimizando la clasificación manual actual**. Por tanto, queremos evitar (i) que el sistema clasifique erróneamente opiniones negativas como no negativas, para que no se publique automáticamente una posible opinión lesiva, y (ii) clasificar erróneamente opiniones no negativas como negativas, para reducir el número de opiniones a revisar manualmente.

En este caso **la hipótesis inicial es que el número de opiniones negativas y no negativas está suficientemente balanceado** ya que parece lógico pensar que las personas dejamos constancia de nuestra opinión en internet cuando la satisfacción es alta o cuando es baja**.** En este caso también se propone también **usar el AUC como métrica de evaluación**, debido a que se busca un clasificador que separe muy bien las clases objetivo.

Se considerará que **el objetivo se habrá cumplido si el clasificador alcanza un AUC de 0.9 en la colección de textos de evaluación.**

4. Desarrollo de un sistema de clasificación de sentimientos: Senti-SyS

4.1. Enfoque

La fuente de datos para la construcción del sistema clasificador será utilizar exclusivamente la información contenida en las opiniones abiertas recogidas en el portal. Por tanto, como no se va a utilizar la metainformación asociada a los cuestionarios donde se ubican los textos para la construcción de los modelos, se intentará extraer la máxima información de éstos por medio de diversas técnicas de minería de textos.

Como el **propósito principal es detectar las opiniones negativas**, la polaridad negativa será priorizada sobre la polaridad no negativa. Así, **cualquier opinión que contenga una valoración negativa sería clasificada como negativa, a pesar de que también contenga valoraciones**

positivas. Por ello, **el análisis de sentimiento** se hará **a nivel de frase,** y por tanto será necesario previamente descomponer los textos en sus frases constituyentes. De esta forma, **el análisis de sentimientos será más ajustado** que el que se conseguiría clasificando la polaridad general de la opinión completa.

Sin embargo, la clasificación de la polaridad de la opinión es el final del proceso. Debido a que tratamos con datos provenientes de internet**, debemos asumir que dentro de nuestra colección de opiniones habrá textos que contienen información no relevante o son en realidad irrelevantes (*spam)***, y por tanto no se ajustan al propósito general de *sanidadysalud.com*. Estos textos no son aptos para su publicación, y hay que identificarlos, filtrarlos y excluirlos de la fase de análisis. Así, nos enfrentamos a un caso concreto de lo que se denomina en la literatura del área ***Detección de Opiniones Irrelevantes*** *(Liu, 2012)*.

Por consiguiente es **necesario** previamente **hacer una serie de comprobaciones sobre el texto recibido en forma de opinión**:

1. **Determinar si el texto está escrito en castellano**.
2. **Determinar si el texto es una opinión o valoración relevante** o apta para su análisis antes de su publicación, ya que los textos recibidos pueden ser mensajes incompletos o con información no relevante para determinar el nivel del servicio ofrecido en el centro sanitario.

Así, una vez determinado que el texto está escrito en castellano y es una opinión relevante, apta para su análisis antes de su publicación, ya sí podrá continuarse con la fase de análisis de la polaridad.

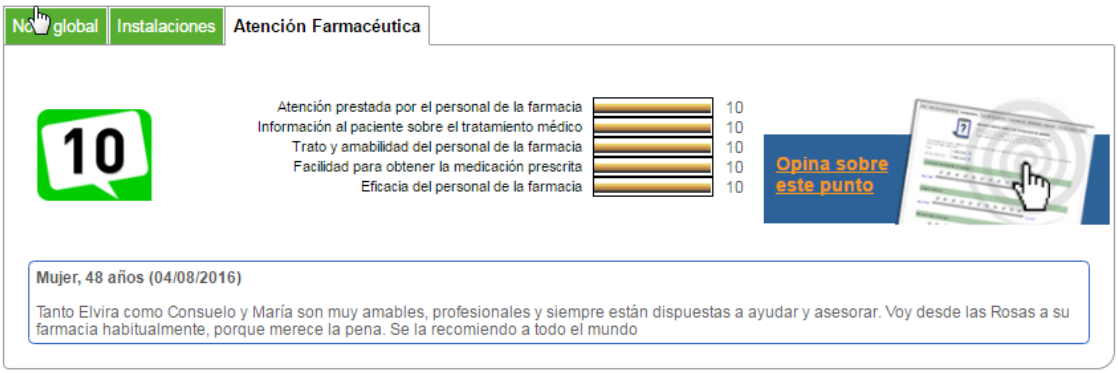


Ilustración 8: Captura de pantalla de una opinión no-negativa recogida en sanidadysalud.com

En **un análisis de sentimientos a nivel de frase**, en primer lugar el texto suele dividirse en frases para **después establecer una clasificación de polaridad para cada una** con un subsistema construido mediante aprendizaje automático.

Por ejemplo, si se pretendiera analizar el sentimiento de la opinión de Ilustración 8 asociada a una farmacia, las frases y sus polaridades serían:

- *"*Tanto Elvira como Consuelo y María son muy amables, profesionales y siempre están dispuestas a ayudar y asesorar*",* con polaridad **no negativa.**
- **"**Voy desde las Rosas a su farmacia habitualmente, porque merece la pena.**"**, con polaridad **no-negativa.**
- **"**Se la recomiendo a todo el mundo**"** con polaridad **no negativa.**

Por tanto, **la opinión general** también **sería clasificada como no negativa.**

En la Ilustración 9 pueden observarse algunos **ejemplos de opiniones consideradas negativas**.

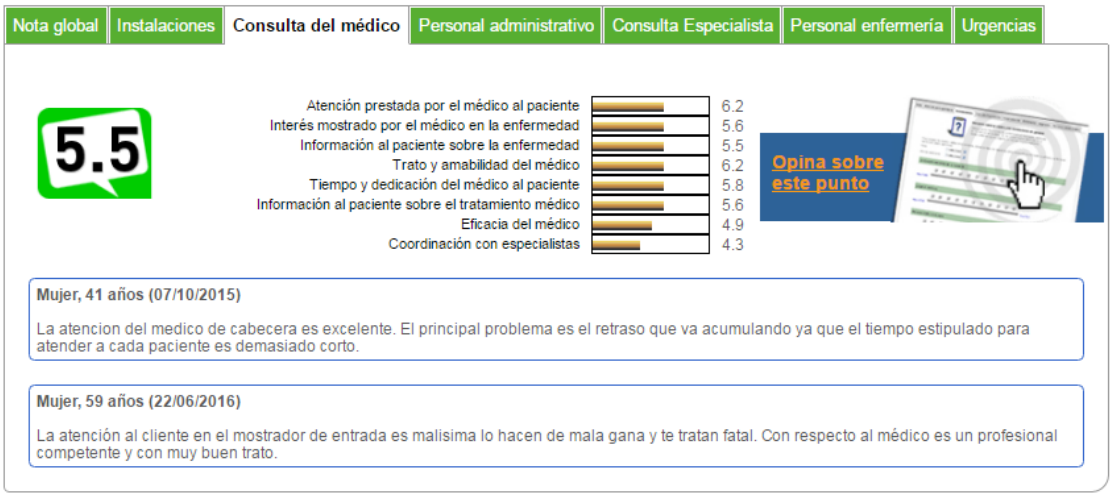


Ilustración 9: Captura de pantalla de una opinión negativa recogida en sanidadysalud.com

Si descomponemos la primera opinión de dicha ilustración en frases y sus polaridades obtendríamos:

- **"**La atencion del medico de cabecera es excelente.", con polaridad **no negativa**
- **"**El principal problema es el retraso que va acumulando ya que el tiempo estipulado para atender a cada paciente es demasiado corto.", con polaridad **negativa**

Según el criterio establecido anteriormente (esto es, si la opinión contiene al menos una frase con polaridad negativa se considerará negativa en su conjunto), **la opinión** completa de nuestro anterior ejemplo **tendría polaridad negativa,** ya que contiene una valoración negativa.

De esta forma, por tanto**, el sistema Senti-SyS** que se plantea **tendría la arquitectura que se muestra en la** Ilustración 10**.**

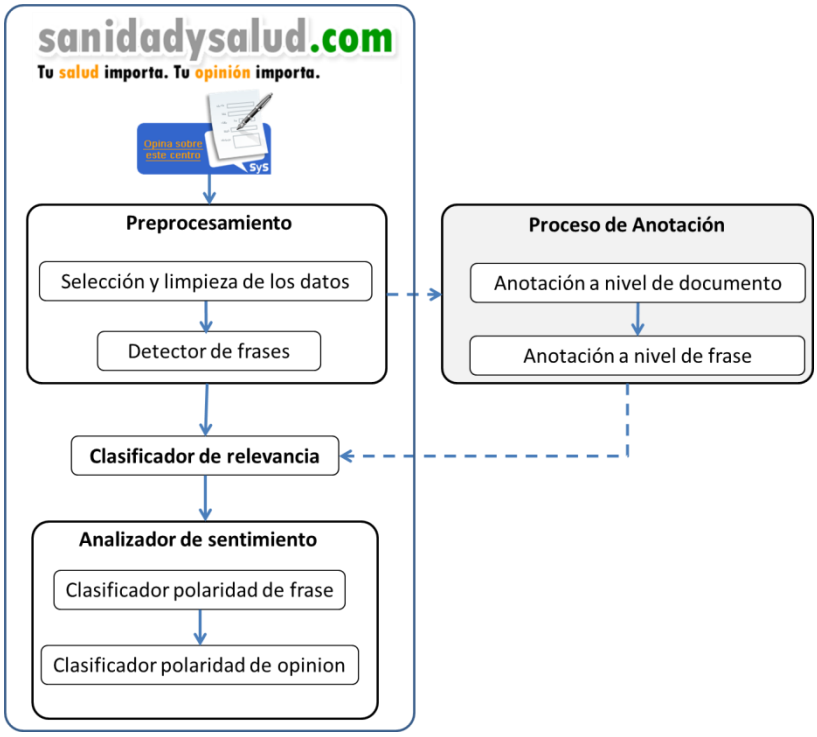


Ilustración 10: Arquitectura del sistema Senti-SyS incluyendo el Proceso de Anotación

4.2. Fases de desarrollo del proyecto según la metodología CRISP-DM

Para gestionar todo el proyecto de minería de textos se va a seguir la **metodología *CRISP-DM***. Dicha metodología parece ser, de forma sostenida, la más usada según los usuarios de Internet en los últimos años[5]. En nuestro caso, nos permite abordar un proyecto complejo con garantías de calidad, trazabilidad y escalabilidad alrededor de la perspectiva de negocio.

El proceso que sigue la metodología, junto con sus fases, se puede ver en la Ilustración 11.

A continuación, en los siguientes puntos se describe cada una de las fases, aplicada a nuestro proyecto.

4.2.1. Comprensión del negocio

La fase de comprensión del negocio se centra principalmente en establecer los objetivos y los criterios de éxito propios del mismo y transformarlos en objetivos y criterios de éxito del proceso de minería de datos. Además, en esta fase se detallan los recursos técnicos y humanos necesarios para el desarrollo del proyecto y se determina cuáles se encuentran ya disponibles. Por último, se aportará la lista de tareas clave del proyecto, así como un cronograma de implementación.

4.2.1.1. Objetivos del negocio

Los objetivos de negocio quedan recogidos en el punto 3.3 del presente documento.

[5] http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

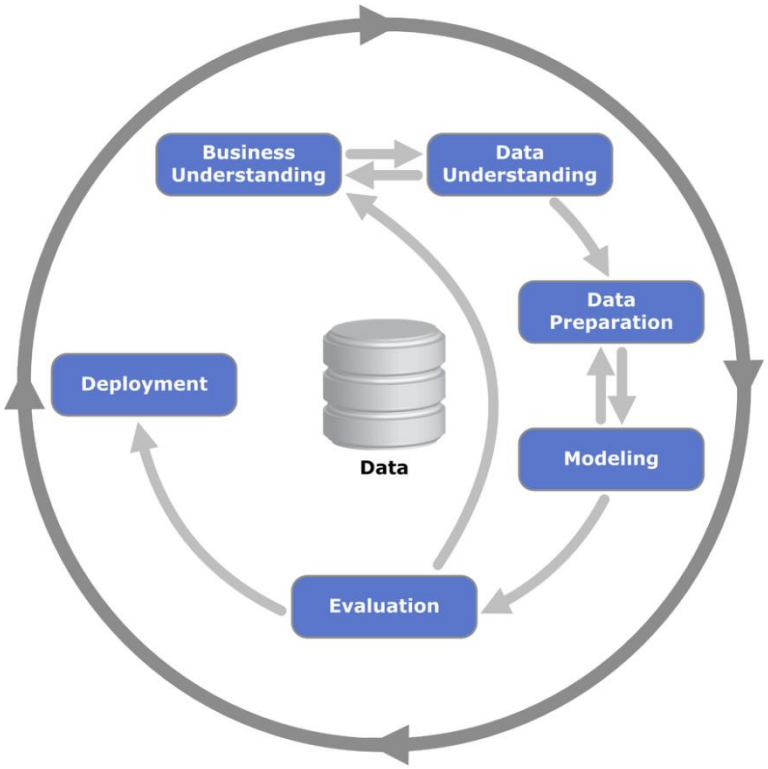


Ilustración 11: Proceso y fases de la metodología CRISP-DM[6]

4.2.1.2. Evaluación de la situación

Actualmente, el proceso de clasificación de las opiniones es manual, y éste será reemplazado por el sistema automático que se desarrollará.

Los **recursos necesarios** estimados para alcanzar el objetivo se pueden dividir en recursos humanos (detallados en la Tabla 1) y recursos técnicos (detallados en la Tabla 2).

Tabla 1: Recursos humanos estimados para la realización del proyecto

Recursos humanos		
Tipo	**Necesario**	**Disponible**
Anotador principal	Persona que aporta conocimiento experto suficiente para conseguir una anotación del conjunto de datos a analizar, alineada con las necesidades del proyecto: Identificar la relevancia de las opiniones, y clasificar la polaridad de las frases que conforman cada opinión.	Sí
Analista/Minero de datos	Persona que se encargará de la fase de minería de datos dentro del proyecto general.	Sí
Analista/Programador informático	Persona que se encargará de la puesta en producción del mejor modelo obtenido en la minería de datos.	Sí
Jefe de proyecto	Persona que establecerá el plan de proyecto y gestionará el desarrollo del mismo hasta su finalización y entrega.	Sí

[6] Fuente: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining (2016)

Tabla 2: Recursos técnicos estimados para la realización del proyecto

Recursos técnicos		
Tipo	**Necesario**	**Disponible**
Hardware	Servidor donde se ejecutará el sistema web.	- *Intel Xeon W3520 4 cores/ 8 threads 2.66 GHz+ 32 GB RAM 2 x 2 TB SATA 250 Mbps*
Hardware	Ordenador personal donde se efectuará la minería de datos.	- *Windows 7 Professional 64bits, Intel Core i3-2120 CPU @3.30 GHz,4GB.*
Software general	Sistema operativo del servidor.	- *Proxmox Virtual Environment,* - *CentOS release 6.7 (Final),* - *Cpanel & WHM 11.52.1.2*
Software general	Servidor de páginas web.	- *Apache/2.2.29 (Unix) mod_ssl/2.2.29*
Software general	Gestor de Base de Datos.	- *mysql Ver 14.14 Distrib 5.5.46, for Linux (i686) using readline 5.1*
Software general	Lenguaje de programación.	- *PHP 5.4.37 (cli) (built: Feb 4 2015 15:28:09)* - *Python 3.5 32-bit*
Software específico	Software para análisis de datos y minería de datos.	- *R version 3.2.2 (2015-08-14) x86_64-w64-mingw32/x64 (64-bit)* - *RStudio Version 0.99.484 – © 2009-2015 RStudio, Inc* - *Microsoft Excel 2010* - *WinPython 3.3.5.0 32bit/64bit*

Respecto a los **recursos técnicos disponibles,** hay que reseñar que el nuevo sistema estará integrado dentro del portal web ***sanidadysalud.com***; por tanto, parte de los recursos disponibles forman parte de la infraestructura IT del portal.

Para **especificar los requisitos del proyecto**, se ha usado el estándar *830-1998 - IEEE Recommended Practice for Software Requirements Specifications.* La especificación resultante se presenta como Anexo I.

4.2.1.3. Objetivos de la minería de datos

Para lograr los objetivos de la minería de datos, en primer lugar será necesario recolectar y preprocesar la colección de textos o corpus inicial. Este corpus inicial se usará después para crear los subsistemas que conformarán el sistema principal a partir del corpus enriquecido con el proceso de anotación, cuyo rendimiento objetivo viene determinado por los objetivos de negocio marcados:

- **Clasificador binario de relevancia** con un AUC de al menos 0.9.
- **Clasificador binario de polaridad** con un AUC de al menos 0.9.

4.2.1.4. Plan de proyecto

A continuación se presenta la lista de tareas principales identificadas a priori en el desarrollo del proyecto, junto con su calendario previsto de implementación (véase la Ilustración 12)

- Investigación previa: estado del arte;
- Identificación de los requisitos de negocio;
- Especificación de requisitos del software y plan del proyecto;
- Comienzo y preparación de la memoria del proyecto;

- Definición de los criterios de éxito del proceso de minería de datos;
- Recolección de los datos;
- Exploración inicial de los datos;
- Detección de frases;
- Proceso de anotación manual;
- Creación de los modelos de predicción y clasificación;
- Evaluación de los resultados obtenidos;
- Despliegue del software desarrollado;
- Conclusiones y cierre.

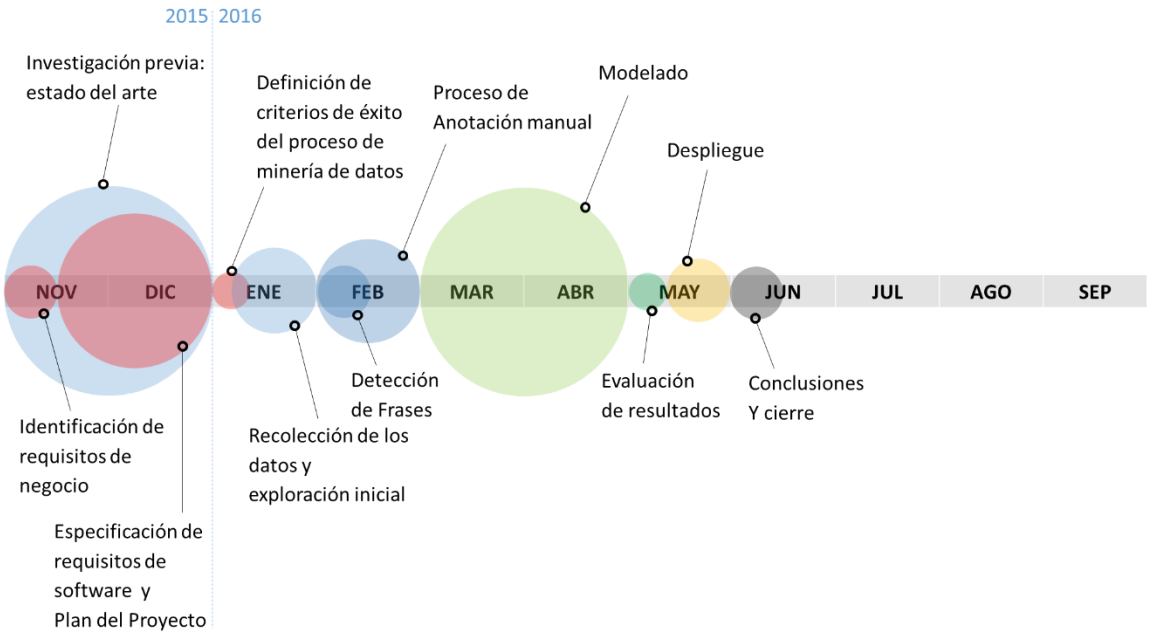


Ilustración 12: Calendario previsto de implementación de tareas identificadas

4.2.2. Comprensión de los datos

La fase de comprensión de los datos es la primera toma de contacto con los datos que hay que analizar. Incluye su recolección, su exploración y la comprobación de su calidad.

Al ser opiniones publicadas en Internet, los datos están alojados en una página web, en nuestro caso en el portal ***sanidadysalud.com***. Por tanto, la opción más general y habitual es descargar esos datos por medio de técnicas de ***web scraping.*** Sin embargo, para la realización del presente proyecto se ha permitido el acceso directo a la base de datos de ***sanidadysalud.com,*** debido a que el objetivo final del mismo es integrar el sistema de clasificación de sentimientos resultante en el propio portal.

Así pues, los datos son recolectados directamente de la base de datos.

4.2.2.1. Descarga de los datos y análisis exploratorio general

Para comenzar el análisis exploratorio inicial, se hizo recuento de las variables y observaciones que tiene nuestro conjunto de datos de partida. En total, consta de **3359 observaciones y 25 variables.**

A continuación, se extrajo una muestra aleatoria de 4 opiniones de estos datos en bruto. Al tener un número elevado de variables, se traspuso la matriz de muestra, para observar

adecuadamente tanto la estructura de los datos como los valores que suelen tomar los campos. El resultado se muestra en la Tabla 3.

Tabla 3: Muestra aleatoria de 4 opiniones para el análisis exploratorio

idobservación	1341	1635	875	2417
idencuesta	14940	5709	11128	13697
idcentro	102277	346	280133	387
idtipocentro	2	3	1	3
idusuario	8744	3308	6402	7972
idbloque	8	9	5	9
idservicio	0	0	0	0
comentarioencuesta	El tiempo que dedican a los pacientes no es suficiente, porque tienen muchos pacientes y no se suplen a los medicos que se dan de baja o faltan por algun motivo. Lo unico que hacen es sobrecargar al resto de medicos que normalmente se encuentran asfixiados con tanto sobrecupo[7]	Es una farmacia excelente, me dan una asistencia esplendida.	Plenamente insatisfecho por la atención médico-administrativa recibida	INCREIBLE. La mejor farmacia de benidorm sin duda alguna
estadocomencuesta	1	1	1	1
CODPROV	28	3	28	3
fecha	2016-01-21 06:21:27	2012-12-04 03:25:31	2014-04-29 10:36:49	2015-07-18 15:24:28
pais	ESP	ESP	ESP	ESP
region	29	60	29	60
city	Barcelona	Barcelona	Barcelona	Barcelona
geoipcodmu	281277	31339	280796	30664
geoipcodprov	28	3	28	3
estado	1	1	1	1
idbloque.1	8	9	5	9
descripcionbloque	Urgencias	Atención Farmacéutica	Hospitalización	Atención Farmacéutica
descripcionlarga	Opina sobre la atención posoperatoria	Opina sobre la atención posoperatoria	Opina sobre la atención posoperatoria	Opina sobre la atención posoperatoria
descripcionmedia	la atención posoperatoria	la atención posoperatoria	la atención posoperatoria	la atención posoperatoria
estado.1	1	1	1	1
antes	Encuesta sobre la atención	Encuesta sobre la	Encuesta sobre la atención	Encuesta sobre la atención

[7] Tanto en este ejemplo como en los anteriores, se observa que estos comentarios contienen faltas de ortografía y errores de puntuación léxicos y sintácticos. Esto puede ser una fuente de ruido en los modelos, por eso debe tenerse en cuenta en la fase de preprocesamiento.

idobservación	1341	1635	875	2417
	posoperatoria	atención posoperatoria	posoperatoria	posoperatoria
tieneservicio	N	N	N	N

De esta forma procedimos a la identificación de las variables relevantes para crear nuestro corpus a analizar:

- **comentarioencuesta:** opinión en texto libre del usuario que hace la encuesta o cuestionario;
- **fecha:** fecha en la que el usuario completó el cuestionario;
- **idencuesta:** identificador del cuestionario

El resto de variables constituyen **metainformación** recogida junto los cuestionarios. Como se comentó anteriormente, **su uso queda fuera del alcance del proyecto,** ya que éste estará centrado en análisis de las opiniones expresadas en los campos de texto, mediante técnicas de minería de textos.

La variable **fecha** es importante, debido a que **el año del cuestionario se usará para dividir la colección de textos u opiniones recolectadas en datos de entrenamiento y evaluación**. Por ello, se observó primeramente el desglose de opiniones por año para ver cómo se distribuyen (véase la Ilustración 13).

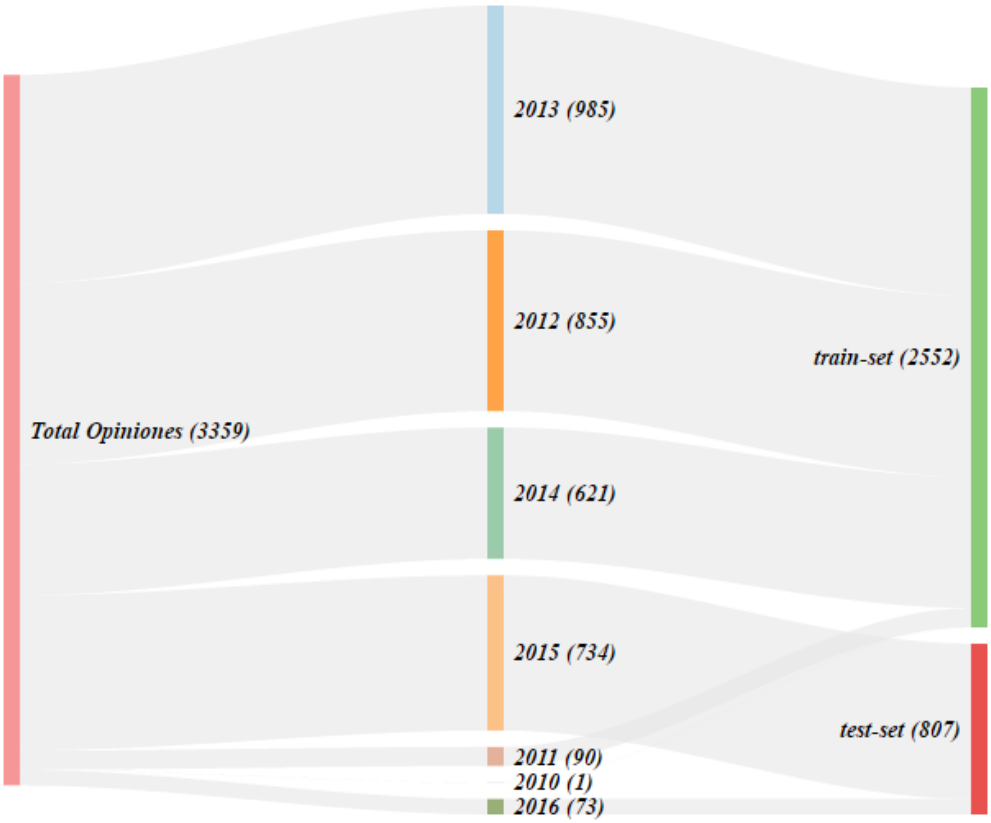


Ilustración 13: Distribución de las opiniones por año y conjunto de datos

4.2.3. Preparación de los datos

La fase de preparación de los datos se centra en seleccionar el conjunto de datos que finalmente se utilizará para el análisis. Sobre ese conjunto de datos se realizarán las transformaciones necesarias para que puedan ser aplicadas las técnicas de modelado seleccionadas. Estas transformaciones podrían aplicarse de forma cíclica junto con el modelado, hasta alcanzar el ajuste deseado de los algoritmos de modelado.

En nuestro caso particular, en esta fase se creará el **corpus** definitivo de opiniones que servirá de entrada al proceso de anotación.

Dicho proceso consiste en asignar una serie de etiquetas disponibles en un inventario predefinido a cada documento o frase del corpus. Para ello, se contará con un **anotador humano,** que decidirá y asignará etiquetas tanto a nivel general de la opinión como de las frases que constituyen dicha opinión. La **Guía de Anotación** creada para asistirle en dicho proceso se puede consultar en el Anexo II.

El corpus enriquecido por la doble anotación será la entrada a la siguiente fase (el modelado).

4.2.3.1. Preprocesado y creación del corpus inicial

Una vez filtada la metainformación, seleccionamos las columnas o variables comentadas para crear el corpus inicial:

- **comentarioencuesta:** opinión en texto libre del usuario que hace la encuesta o cuestionario;
- **fecha:** fecha en la que el usuario completó el cuestionario
- **idencuesta:** identificador del cuestionario

Además, se aplican una serie de **operaciones básicas de limpieza a los** textos para facilitar su posterior manipulación y procesamiento:

- Eliminación de saltos de línea;
- Eliminación de caracteres de escape (/);
- Sustitución de comillas dobles por simples.

4.2.3.2. Detección de frases

El problema de la detección de frases es actualmente un problema abierto en el campo del Procesamiento del Lenguaje Natural. En algunos casos, se trata como un problema resoluble mediante aprendizaje automático, en el que se entrena un clasificador para que determine si un signo de puntuación delimita una frase o no.

En el contexto del presente trabajo, a fin de reducir su complejidad, el problema se afrontará de una forma más simple: por medio de herramientas específicas de PLN. En particular, se construyó un algoritmo en ***Python*** que gira en torno a tres ideas:

- El uso de una librería denominada NLTK[8] para las tareas típicas del PLN.
- La identificación y aplicación de reglas específicas del dominio del problema como, por ejemplo, la detección de las abreviaturas típicas de doctor y doctora.
- La identificación y aplicación de reglas específicas para el corpus disponible, como por ejemplo la separación de palabras consecutivas unidas por un signo de puntuación.

[8] http://www.nltk.org/

A priori, es imposible saber el éxito del algoritmo; por tanto, se va construyendo de forma cíclica, por medio de una estrategia de ensayo y error sobre una muestra de las opiniones del corpus, hasta que separa razonablemente bien las opiniones de la muestra.

De esta forma, tal y como se ha comentado**, este algoritmo se aplicará sobre** el subconjunto del corpus inicial formado por la **colección de textos etiquetados como relevantes, y se comprobará su éxito definitivo con la segunda fase del proceso de anotación, a nivel de frase.**

4.2.3.3. Proceso de anotación

El proceso de anotación típico del análisis de sentimientos consiste en identificar y clasificar la polaridad de las opiniones expresadas a través de los documentos que conforman el corpus a analizar. Esta clasificación se manifiesta a través de la asignación de unas etiquetas disponibles de antemano a cada documento, a cada frase o a cada aspecto identificado, en función del nivel de profundidad requerido en el análisis. En particular, este último nivel basado en aspectos, el nivel más fino, requiere que el anotador también identifique en cada opinión la entidad y/o los atributos a los que se refiere, desde un inventario definido previamente. Por ello, éste también es una entrada al proceso de anotación.

Para dar soporte al anotador y crear el marco de referencia para el proceso de anotación, se ha creado una Guía de Anotación, que puede consultarse en el Anexo II.

Como se ha indicado anteriormente, en el presente trabajo, **el proceso de anotación es en realidad doble**: **anotación de comentarios y anotación de frases.**

En primer lugar, la anotación **de comentarios determina si la opinión es candidata a ser publicada, porque es relevante.**

Así, en este caso, el **proceso de anotación de comentarios** se define **como el proceso en el que, para cada texto o comentario tj, se trata de identificar P_l , donde la polaridad $P_l \in P_1$**, según se define P_1 en la Guía de Anotación[9].

A continuación, con la anotación **de frase**, lo **que se determinará será la polaridad** de las frases en las que se descompone cada opinión, marcada previamente como relevante, y con ello la polaridad general de la propia opinión. Como ya se ha dicho, el objetivo final es identificar las opiniones con valoraciones negativas.

Así, en el contexto del presente trabajo, el **proceso de anotación de frases** se define **como el proceso en el que, para cada frase $f_j \in t_j : P_l = 'relevante'$, se trata de identificar P_m , donde la polaridad $P_m \in P_2$,** según se define P_2 en la Guía de Anotación.[10]

Es por ello, que en esta segunda fase el corpus de opiniones hay que descomponerlo en las frases que constituyen dichas opiniones, y pasar ese nuevo corpus extendido como entrada a la segunda fase del proceso de anotación.

4.2.3.3.1. Fase I: Anotación de comentarios

Las entradas a esta fase serían:

[9] P_1 = *{relevante, no-relevante, otro-idioma, error}*

[10] P_2= *{negativa, no-negativa, error-frase}*

- El corpus en cuestión, formado por **3359 opiniones o documentos**.
- El **conjunto P_1** de etiquetas para el subsistema de análisis de relevancia.

Con ello, y siguiendo el procedimiento propuesto en la Guía de Anotación, se generan las salidas del proceso:

- Asignación a cada comentario de una de las etiquetas del inventario disponible P_1, con el objetivo de separar las opiniones relevantes respecto de las no relevantes. Esto servirá de entrada al subsistema de detección de relevancia de las opiniones.

Para hacer el proceso de anotación más operativo y minimizar errores en la manipulación de etiquetas, se creó una pequeña herramienta de anotación *online* integrada en *sanidadysalud.com*. La interfaz de esta herramienta se puede consultar en el Anexo III.

Los resultados de la distribución del etiquetado del punto 1 por etiqueta y año se han incluido en la Ilustración 14.

Frecuencia de etiquetas por años (%)	Años							
Etiquetas	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**Total general**
error	0.00%	0.00%	0.18%	0.30%	0.09%	0.09%	0.00%	0.65%
no-relevante	0.00%	0.03%	0.98%	1.22%	0.74%	0.68%	0.12%	3.78%
otro-idioma	0.00%	0.12%	0.21%	0.09%	0.09%	0.24%	0.00%	0.74%
relevante	0.03%	2.53%	24.08%	27.72%	17.56%	20.84%	2.05%	94.82%
Total general	**0.03%**	**2.68%**	**25.45%**	**29.32%**	**18.49%**	**21.85%**	**2.17%**	**100.00%**

Frecuencia de etiquetas por años (%)	Años							
Etiquetas	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**Total general**
error			6	10	3	3		22
no-relevante		1	33	41	25	23	4	127
otro-idioma		4	7	3	3	8		25
relevante	1	85	809	931	590	700	69	3185
Total general	**1**	**90**	**855**	**985**	**621**	**734**	**73**	**3359**

Ilustración 14: Distribución de las etiquetas del conjunto P_1 por años

Los datos de la mencionada ilustración muestran que las **opiniones relevantes son mayoría frente** a las **no relevantes** (94.82% frente a 3.78%). Esto confirma la hipótesis de que las clases objetivo están muy desequilibradas, lo cual puede representar un reto para nuestro clasificador.

Por último, se observa que **tanto el porcentaje de opiniones con error como el de opiniones en otro idioma distinto al castellano son muy bajos,** quedando ambos por debajo del 1%. Este tipo de opiniones queda fuera del alcance del presente trabajo, así que **se eliminaron del corpus en los posteriores análisis**.

Por su parte, el **año** es un dato clave, ya que, según los objetivos de negocio **servirá para dividir el corpus en los conjuntos de entrenamiento y validación por un lado, hasta el 2014 incluido; y evaluación, por otro, desde el 2015 incluido**. Así, pueden usarse para el **entrenamiento** y validación en esta fase un **75.97%** de los datos, y para **verificar la consecución del objetivo** o evaluar los modelos un **24.03%** de los datos (véase la Ilustración 14 y la Ilustración 15).

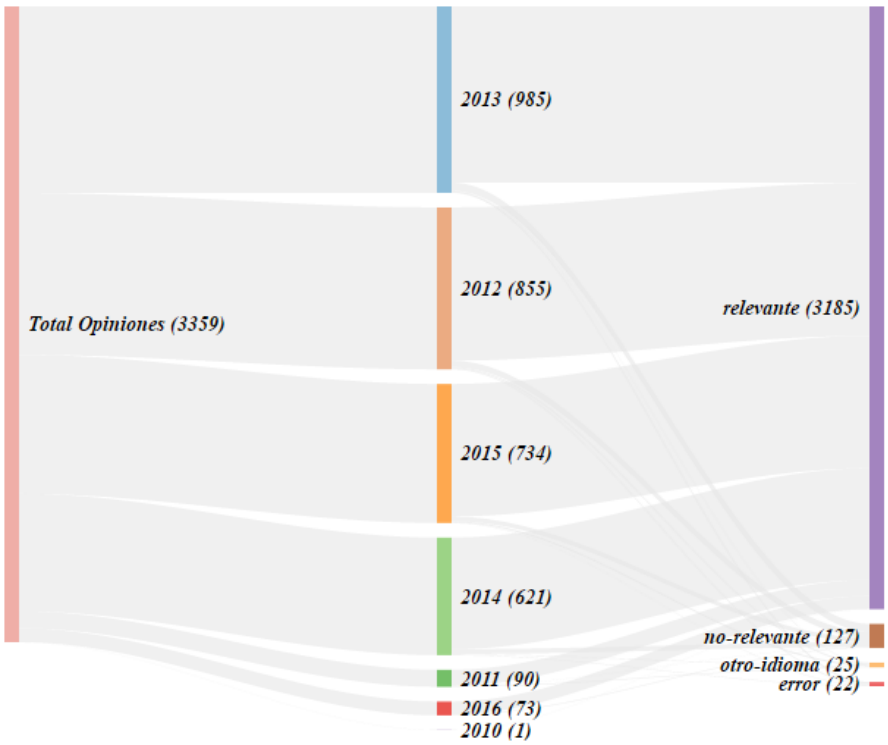


Ilustración 15: Distribución de opiniones por año y etiqueta de la fase I del proceso de anotación

4.2.3.3.2. Fase II: Anotación de frases

En esta fase se analiza la polaridad de las opiniones clasificadas como relevantes anteriormente. Para intentar extraer la máxima información posible de estas opiniones, primero se descompone el corpus en sus frases constituyentes. Es decir, al **corpus formado por las 3185 opiniones o comentarios** clasificados previamente como **relevantes**, se le **aplica un algoritmo de detección de frases**, que a su vez **crea un nuevo corpus extendido**.

Sin embargo, debido a que el problema de detección de frases es un problema complejo, y actualmente abierto en el campo de PLN, a priori es imposible saber si el resultado de la detección de frases efectuado será satisfactorio o no. Es por ello que el conjunto P_2 incluye una etiqueta **{error-frase}** para clasificar las frases incompletas o incomprensibles.

Descomposición de las opiniones en frases

El resultado del algoritmo de detección de frases aplicado fueron las **6793 frases contenidas en las 3185 opiniones clasificadas como relevantes**. Por tanto, en esta segunda fase las entradas serían:

- El corpus extendido formado por las **6793 frases o documentos**;
- El **conjunto P_2** de etiquetas para el sistema de análisis de polaridad.

Al igual que antes, el anotador humano puede contar con una pequeña herramienta de anotación online, integrada en *sanidadysalud.com*, para etiquetar las frases que conforman el corpus extendido.

Con ello, y siguiendo el procedimiento propuesto en la Guía de Anotación (véase el Anexo II), los resultados de la anotación por etiqueta y año quedan distribuidos como se muestra en la Ilustración 16.

Frecuencia de etiquetas por años	Años							
Etiquetas P2	2010	2011	2012	2013	2014	2015	2016	Grand Total
error-frase		7	60	61	44	34	1	207
negativa		90	784	1049	687	776	93	3479
no-negativa	2	86	828	926	542	660	63	3107
Grand Total	**2**	**183**	**1672**	**2036**	**1273**	**1470**	**157**	**6793**

Frecuencia de etiquetas por años (%)	Años							
Etiquetas P2	2010	2011	2012	2013	2014	2015	2016	Grand Total
error-frase	0,00%	0,10%	0,88%	0,90%	0,65%	0,50%	0,01%	3,05%
negativa	0,00%	1,32%	11,54%	15,44%	10,11%	11,42%	1,37%	51,21%
no-negativa	0,03%	1,27%	12,19%	13,63%	7,98%	9,72%	0,93%	45,74%
Grand Total	**0,03%**	**2,69%**	**24,61%**	**29,97%**	**18,74%**	**21,64%**	**2,31%**	**100,00%**

Ilustración 16: Distribución de las etiquetas del conjunto P_2 por años

Los datos incluidos en la mencionada ilustración muestran que la **etiqueta mayoritaria** es la **negativa (51.21%)** frente a la **no negativa (45.74%),** aunque ambas están bastante equilibradas.

También se observa que las frases con error son relativamente pocas **(3.05%)**. Por tanto, puede darse por válido el algoritmo de segmentación de las opiniones en frases. Estas frases con error quedaron finalmente fuera del corpus extendido.

Así, en resumen, se obtuvo un **corpus definitivo con 6586 frases**, negativas o no negativas.

Por lo que respecta al **año**, sobre este corpus definitivo se obtuvieron unos números muy parecidos a los casos anteriores: **el 75.83% de los datos** serían usados **para entrenamiento** y **validación,** frente al **24.27%** que serían usados **para la evaluación de los modelos** desarrollados (véase la Ilustración 16 y la Ilustración 17)

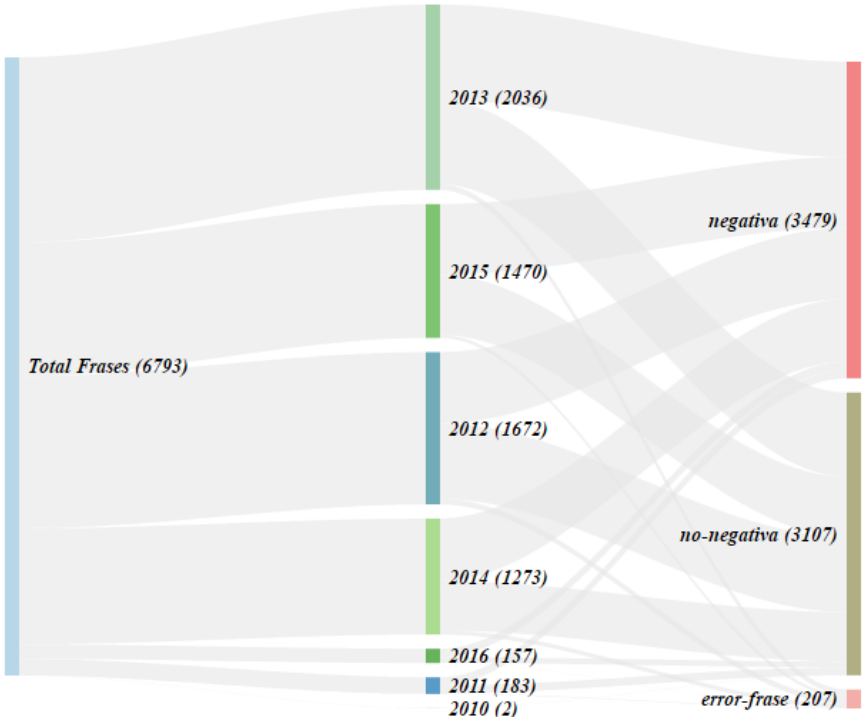


Ilustración 17: Distribución de frases por año y etiqueta de la fase II del proceso de anotación

A continuación trasladamos la información de polaridad desde el nivel de frase al nivel de opinión, quedando el resultado final de la clasificación de opiniones como se muestra en la Ilustración 18.

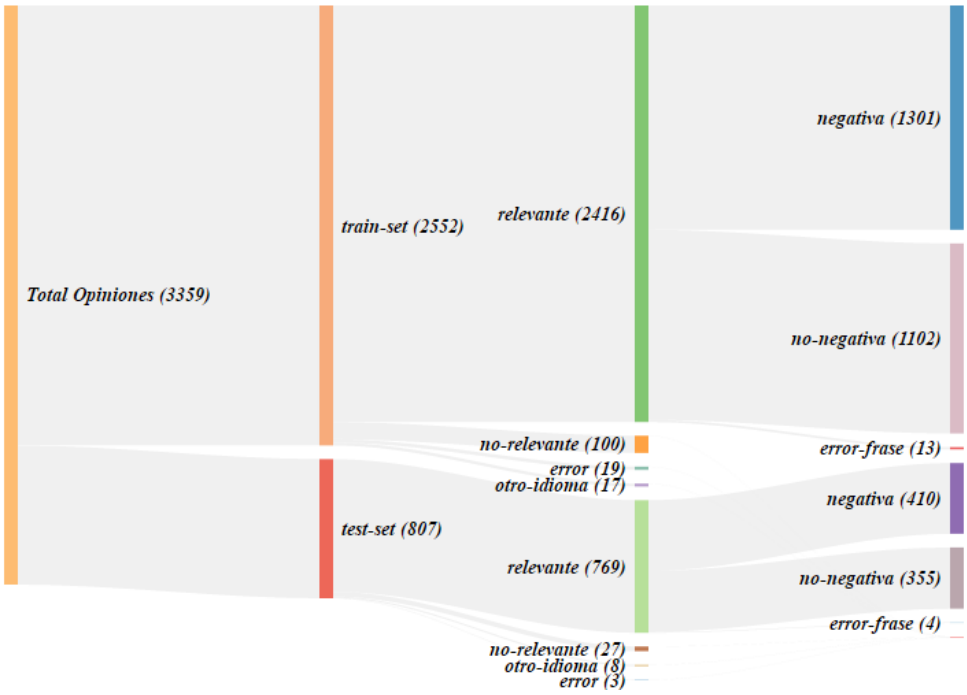


Ilustración 18: Distribución de la clasificación de opiniones según el proceso de anotación completo

4.2.4. Modelado

Durante la fase de modelado, como hemos comentado antes, se construyeron **los dos clasificadores** que serán los subsistemas principales del sistema **Senti-SyS:**

- Clasificador de Relevancia.
- Clasificador de Polaridad.

Para ello, en primer lugar había que **seleccionar una técnica de modelado que se adapte al problema de clasificación binaria**. De acuerdo con la literatura existente, las técnicas empleadas tradicionalmente son ***Naive Bayes***, ***Support Vector Machines (SVM) y KNN***. Sin embargo, podría emplearse cualquier técnica que se adapte a este tipo de problemas: árboles de decisión, regresión logística, redes neuronales, etc.

Una vez seleccionada la técnica de modelado, se procedió a la **creación del conjunto de variables independientes o características (*features*),** en función de las necesidades de cada técnica. De esta forma, la técnica ya podría ser ejecutada sobre el conjunto de datos de entrenamiento. Finalmente se fue **ajustando el modelo cíclicamente,** según los resultados que se fueron obteniendo; empleando la **validación cruzada repetida** para **monitorizar sus resultados frente a los objetivos** de minería de datos marcados inicialmente. Este proceso debía ser repetido por cada técnica de modelado seleccionada.

4.2.4.1. Clasificador de Relevancia

Como se ha visto el Clasificador de Relevancia tiene el objetivo de **clasificar las opiniones en relevantes y no relevantes** utilizando los métodos de aprendizaje automático supervisado seleccionados a partir del **corpus de opiniones anotado**. Por tanto, es un **problema de clasificación binaria.**

El principal reto de este clasificador es que **las clases objetivo están muy desequilibradas**: el 94.67% de las observaciones del corpus de entrenamiento son de la clase "relevante" y el 3.92% de la clase "no relevante".

En un principio, las **técnicas** que se seleccionaron para el **modelado** eran ***Naive Bayes***[11] ***y SVM***[12], debido a su simplicidad y rapidez de ejecución. Para ello se utilizaron las **librerías NLTK**[13] **y Scikit-learn**[14] de *Python*. Sin embargo, tras los primeros ajustes de los modelos se decidió incorporar nuevas técnicas para comparar resultados: ***Random Forest***[15] y ***Stochastic Gradient Descent***[16]

Los diferentes **parámetros de configuración de los algoritmos** se decidieron por medio de una **validación cruzada de 5 iteraciones y un *grid* exhaustivo de búsqueda**[17].

4.2.4.1.1. Extracción de características (feature extraction)

El primer paso del modelado consiste en extraer información de los textos para crear el conjunto de características con las que se alimentará la técnica de modelado seleccionada.

La forma típica de hacer esto es **transformar cada texto en un vector de palabras o términos, con los que se determinan las características del conjunto de datos.** Posteriormente, se **seleccionan las mejores** de éstas por métodos de selección de variables para incrementar el rendimiento de los modelos.

4.2.4.1.1.1. Vector de palabras

Uno de los métodos más conocidos para la creación de características a partir del texto original de la opinión es el **Vector de Palabras** (*Bag of Words, BOW*). Según dicha técnica, cada documento o texto *t* se define de la forma

$$t = (w_1, w_2, w_3, \dots, w_{|V|})$$

donde $|V|$ es el tamaño del vocabulario total que tiene el corpus y cada w_i toma valor [0,1] si el término o palabra aparece en el texto *t.*

Por tanto, el primer paso es separar cada texto, en este caso cada opinión, en su vector de palabras correspondiente (*word tokenization*).

De esta forma, **el corpus de entrada con tamaño M se transforma en una matriz de dimensión N x M,** donde **N será el número de características (*features)* y M el número de observaciones.**

Para caracterizar los textos, **el caso más sencillo** es aquel en el que el **vector de palabras** toma **valores binarios**, en función de si el término aparece o no en el texto.

Sin embargo, la anterior caracterización no tiene en cuenta que **la importancia de los términos es diferente** dentro de cada texto. La aproximación más simple de capturar esta importancia de cada término es calcular la **frecuencia de aparición de cada término *(Term Frecuency, TF).*** La Ilustración 19 incluye, a modo de ejemplo, el cálculo del vector de palabras para la frase del

[11] http://www.nltk.org/_modules/nltk/classify/naivebayes.html
[12] http://scikit-learn.org/stable/modules/svm.html
[13] http://www.nltk.org/
[14] http://scikit-learn.org/stable/
[15] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[16] http://scikit-learn.org/stable/modules/sgd.html
[17] http://scikit-learn.org/stable/modules/grid_search.html

ejemplo.

"los médicos no se preocupan de los pacientes, sólo se preocupan de los médicos"

Término	los	médicos	no	se	preocupan	de	pacientes	sólo
Frecuencia	3	2	1	2	2	2	1	1

Ilustración 19: Ejemplo de extracción de TF

Por otra parte, esta caracterización tiene el problema de **que las palabras comunes (*stop words*)**, como ocurre en nuestro ejemplo ("los", "de", "se"), **casi siempre son las que tienen mayor valor sin ser las más importantes**. Por tanto, esto es una fuente de ruido para la construcción de los modelos.

Para resolver ese problema, se han desarrollado otras medidas como ***TF-IDF,*** que **normaliza ese valor *TF* con la inversa de la frecuencia de aparición de ese término en todo el corpus (*IDF*).** Actualmente existen diferentes variantes de esta medida, que es una de las más usadas para medir la relevancia general de un término o característica dentro de una colección de textos o documentos.

La elección de la caracterización óptima se realiza en función del tipo de algoritmo que se vaya a utilizar para el modelado y el conjunto de datos a analizar. En este caso, al ser textos relativamente cortos (longitud máxima 500 caracteres) y el corpus relativamente pequeño (3359 textos), **se utilizó simplemente la ocurrencia de aparición de los términos**.

4.2.4.1.1.2. N-gramas

La técnica del vector de palabras es una técnica sencilla y que ofrece buenos resultados. Sin embargo, puede mejorarse con la generalización a ***N-gramas,*** que trata también de capturar información relativa al orden de las palabras, cosa que no ocurre con el vector de palabras.

De esta forma, si **N = 2 se construyen *bigramas,*** agrupando los términos consecutivos de dos en dos. La Ilustración 20 muestra los bigramas correspondientes a la misma frase del ejemplo anterior.

"médicos no" "se preocupan" "de los" "pacientes sólo" "~~se preocupan~~" "~~de los~~"

"los médicos no se preocupan de los pacientes, sólo se preocupan de los médicos"

"los médicos" "no se" "preocupan de" "los pacientes" "sólo se" "~~preocupan de~~" "~~los médicos~~"

Ilustración 20: Proceso de extracción de bigramas

De la misma forma, **si N = 3, se construyen *trigramas*,** y **si N= 1 se obtienen unigramas,** lo que es equivalente al vector de palabras.

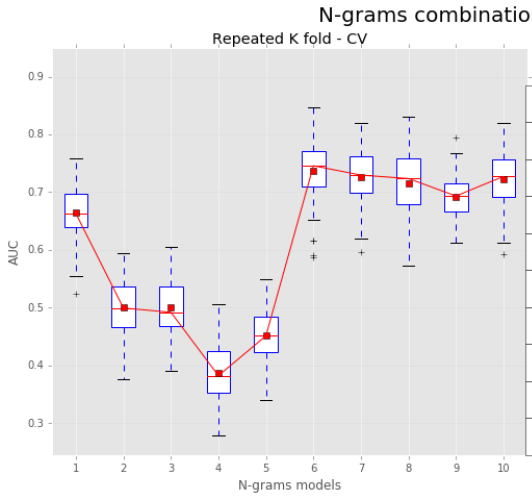
4.2.4.1.1.3. Pruebas

Debido a que el conjunto de datos de entrenamiento no era muy extenso, se utilizó una **validación cruzada de 4 iteraciones, repetida 20 veces,** para **evitar** en la medida de lo posible el **sobreajuste** y **facilitar la comparación de los modelos obtenidos**. Además, para minimizar el problema del gran desequilibrio de las clases objetivo, **se utilizó estratificación** en el diseño de dicha validación cruzada**.**

Con los resultados se construyeron gráficos del tipo **diagrama de cajas** (*boxplot*), ya que son ideales para ver el balance sesgo-varianza de los modelos conseguidos. En todos ellos se

representa la mediana de los valores como una línea horizontal roja y la media como un cuadrado rojo.

Para comenzar la serie de pruebas, primero se averiguó qué forma de extraer las características era más apropiada para este conjunto de datos: unigramas, bigramas y/o trigramas.



Models

	Median	Mean	Std.
1. NaiveBayesNLTK_unigrams	0.6629	0.6651	0.0468
2. NaiveBayesNLTK_bigrams_unigrams	0.4994	0.5009	0.0477
3. NaiveBayesNLTK_trigrams_unigrams	0.492	0.4998	0.0472
4. NaiveBayesNLTK_trigrams_bigrams	0.3823	0.3873	0.0488
5. NaiveBayesNLTK_trigrams_bigrams_unigrams	0.4512	0.4519	0.0442
6. SVC_unigrams	0.7458	0.7364	0.0523
7. SVC_bigrams_unigrams	0.7296	0.7265	0.0505
8. SVC_trigrams_unigrams	0.7237	0.7147	0.0573
9. SVC_trigrams_bigrams	0.6934	0.6918	0.0358
10. SVC_trigrams_bigrams_unigrams	0.7273	0.723	0.051

Ilustración 21: Comparativa de extracción de características para NaiveBayesNLTK y SVC

Como puede observarse en la Ilustración 21, en ambas técnicas, el **uso únicamente de unigramas** era la mejor opción. Además, en este primer test, SVC fue superior a NaiveBayesNLTK en sesgo (mediana de 0.74 frente a 0.66). En varianza son similares.

De esta forma, se escogieron **los unigramas**, para intentar optimizar el modelo en las siguientes subfases.

4.2.4.1.1.4. Palabras comunes (stop words) y uso de raíces (stemming)

En este punto se probaron dos de las **técnicas** más utilizadas para **eliminar ruido del espacio de características:**

- **Quitar las palabras comunes (*stop words*, SW)** de nuestra lista de características**:** las palabras comunes son aquellas que ocurren más frecuentemente en un idioma pero no poseen una carga léxica relevante y, por ello, aportan muy poca información a los modelos**.** Esta lista puede definirse *ad hoc* para cada problema y dominio determinados para refinar los modelos. En nuestro caso, por simplificar**, se utilizó por defecto la lista de palabras comunes que incluye la librería NLTK.**

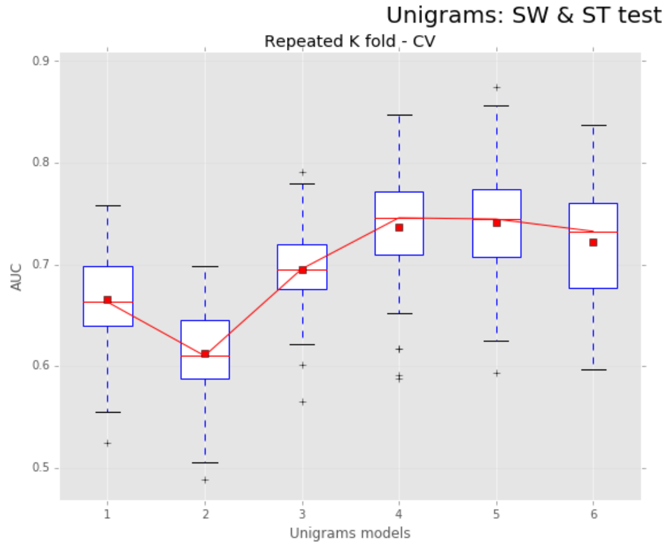
- **Reducir las características o términos a su raíz (*stemming*, ST)**: Con ello se pretende agrupar términos que pueden ser derivaciones de la misma palabra o palabras muy relacionadas semánticamente**. De esta forma, se reduce el ruido del conjunto de datos.** Para ello se utilizó la librería ***Snowball***[18] a través de NLTK.

En ambos casos se usaron las versiones para el idioma castellano.

[18] http://www.nltk.org/_modules/nltk/stem/snowball.html

4.2.4.1.1.5. Pruebas

En primer lugar, se probó a introducir gradualmente las técnicas de *SW* y *ST* para ver su efecto con los diferentes algoritmos. Se incluye una comparativa de su comportamiento en la Ilustración 22.



Models

	Median	Mean	Std.
1. NaiveBayesNLTK unigrams	0.6629	0.6651	0.0468
2. NaiveBayesNLTK unigrams sw	0.61	0.6126	0.0448
3. NaiveBayesNLTK unigrams sw st	0.6955	0.6954	0.0391
4. SVC unigrams	0.7458	0.7364	0.0523
5. SVC unigrams sw	0.7446	0.7418	0.0476
6. SVC unigrams sw st	0.7326	0.7223	0.053

Ilustración 22: Comparación de técnicas SW y ST para modelos NaiveBayesNLTK y SVC

Podemos observar en la anterior ilustración que el uso de SW y ST produce resultados diferentes. En los modelos NaiveBayesNLTK, el uso de SW y ST combinado mejora claramente al modelo base. Sin embargo con los modelos SVC la mejora no está tan clara; parece que el uso de SW sí mejora ligeramente el modelo base, pero combinándolo con ST lo empeora, aunque también muy ligeramente. Se esperaba que ambas técnicas combinadas redujeran el sobreajuste del modelo a los datos de entrenamiento. No obstante, en un conjunto de datos más bien reducido como el que tenemos, puede ocurrir que no esté claro si su uso mejora o empeora el modelo, como ocurre en este caso.

Podemos observar en la Ilustración 23 las palabras comunes extraídas del corpus con una nube de palabras, donde el tamaño de la fuente representa la frecuencia de aparición de la palabra en el corpus.

Ilustración 23: Nube de palabras comunes extraídas del corpus de opiniones

4.2.4.1.2. Selección de variables (Feature Selection)

En la minería de textos habitualmente nos enfrentamos al **problema de la alta dimensionalidad**. Por tanto, también **es habitual utilizar tanto métodos de reducción de la dimensionalidad como de selección de variables importantes,** de cara a mejorar el resultado de los clasificadores mediante la eliminación de ruido del espacio de características utilizadas en el entrenamiento, así como el tiempo de cálculo.

En particular, el **proceso de selección de variables** (Manning et al., 2008) consiste en seleccionar un subconjunto de características del espacio de aquellas disponibles, de forma que se use exclusivamente este subconjunto para el entrenamiento y clasificación.

En la Ilustración 24 se puede ver el crecimiento en número de características, según vamos aumentando el tamaño de los n-gramas.

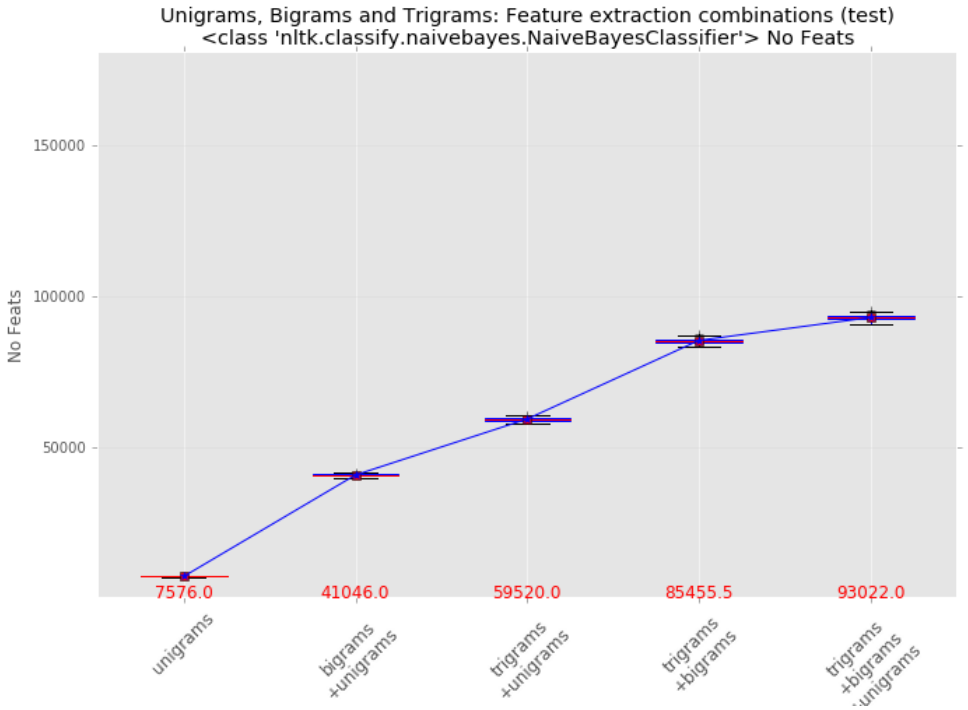


Ilustración 24: Comparación del número de características conseguido con los diferentes n-gramas

Por tanto, las siguientes estrategias van enfocadas a reducir ese número de términos o características, para mejorar el poder de clasificación del modelo eliminando posibles fuentes de ruido.

El algoritmo para seleccionar las *k* mejores características se incluye en la Ilustración 25 (Manning et al., 2008):

SELECTFEATURES($\mathbb{D}, c, k$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $L \leftarrow []$
3 **for each** $t \in V$
4 **do** $A(t,c) \leftarrow$ COMPUTEFEATUREUTILITY($\mathbb{D}, t, c$)
5 APPEND($L, \langle A(t,c), t \rangle$)
6 **return** FEATURESWITHLARGESTVALUES(L, k)

Ilustración 25: Algoritmo de selección de las k mejores características

4.2.4.1.2.1. Información Mutua (MI)

La **medida *PMI*** *(Pointwise Mutual Information)* indica cuánta información aporta una palabra sobre otra o el nivel de coocurrencia de ambas. Por ello, en minería de textos este método puede usarse tanto para calcular el ***PMI*** entre un par de características dado, como para calcular el ***PMI*** de cada característica con cada clase objetivo.

De esta forma, en el primer caso podremos **averiguar los n-gramas con asociación más fuerte;** y en el segundo, hacer establecer **un listado de las características más importantes** en cuanto a la información que aportan.

Formalmente, se define como:

$$\boldsymbol{PMI}(\boldsymbol{x},\boldsymbol{y}) = \log\frac{\Pr(x,y)}{\Pr(x)\Pr(y)}$$

donde *x* e *y* son dos eventos pertenecientes a las variables discretas *X* e *Y,* que se asume que son independientes.

Por ello, es una medida simétrica, que puede tener valores positivos y negativos; y adopta el valor 0 cuando *X* e *Y* son independientes.

Una debilidad de esta medida (Yang y Pedersen, 1997) es que no es comparable entre términos con una frecuencia de aparición muy distinta.

Para evitar errores de cálculo en coma flotante al multiplicar números pequeños, la fórmula se puede reescribir de la siguiente manera:

$$\boldsymbol{PMI}(\boldsymbol{x},\boldsymbol{y}) = \log \boldsymbol{Pr}(\boldsymbol{x},\boldsymbol{y}) - \log \boldsymbol{Pr}(\boldsymbol{x}) - \log \boldsymbol{Pr}(\boldsymbol{y})$$

Si en vez de fijarnos en los eventos individuales de *X* e *Y*, tenemos en cuenta el promedio de todos los eventos posibles, entonces estamos hablando de la **Información Mutua (MI)**. Por tanto, **MI se define como el valor esperado de PMI sobre todos los posibles eventos de *X* e *Y***:

$$\boldsymbol{MI}(\boldsymbol{x},\boldsymbol{y}) = \sum_{y\in Y}\sum_{x\in X}\boldsymbol{Pr}(\boldsymbol{x},\boldsymbol{y})\log\frac{\boldsymbol{Pr}(\boldsymbol{x},\boldsymbol{y})}{\boldsymbol{Pr}(\boldsymbol{x})\,\boldsymbol{Pr}(\boldsymbol{y})}$$

De esta forma, MI adopta el valor 0 si *X* e *Y* son completamente independientes entre sí; y toma el valor máximo cuando toda la información proporcionada por *X* es compartida por *Y*, y viceversa. De esta forma, conocer *X* determina el valor de *Y*, y viceversa.

Cabe resaltar también que esta medida toma siempre valores no negativos.

4.2.4.1.2.2. Test de la χ^2 (CHI)

El **test de la $\boldsymbol{\chi^2}$** calcula un estadístico para **determinar si una clase *i* y la característica *w* son independientes**. Este test hace la suposición inicial de independencia y, si el estadístico tiene un valor elevado, se rechaza esa hipótesis inicial (hipótesis nula), concluyendo que pueden ser dependientes. Si por el contrario el valor es 0, son claramente independientes.

Por tanto, se puede **crear un listado de características en función de ese estadístico,** como medida de la importancia de esta característica en cuanto a su dependencia con la clase objetivo.

La expresión de cálculo de este estadístico (Yang y Pedersen, 1997) es la siguiente:

$$\boldsymbol{\chi^2}(\boldsymbol{t},\boldsymbol{c}) = \frac{\boldsymbol{N}*(\boldsymbol{AD}-\boldsymbol{CB})^2}{(\boldsymbol{A}+\boldsymbol{C})*(\boldsymbol{B}+\boldsymbol{D})*(\boldsymbol{A}+\boldsymbol{B})*(\boldsymbol{C}+\boldsymbol{D})}$$

donde *N* es el número de documentos, *A* es el número de documentos de clase *c* que contienen el término *t*, *B* es el número de documentos de otras clases diferentes de *c* que contienen *t*, *C* es el número de documentos de clase *c* que no contienen *t* y *D* es el número de documentos de otras clases que no contienen *t.*

A partir de ahí se computa el estadístico χ^2 entre cada clase objetivo *c* y cada término *t* del corpus, combinándose después el estadístico para cada término *t* mediante el promedio y máximo para calcular la importancia global:

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} Pr(c_i) * \chi^2(t, c_i)$$

$$\chi^2_{max}(t) = \max_i \{\chi^2(t, c_i)\}$$

La principal diferencia entre el test de la χ^2 y la medida ***MI*** es que el estadístico χ^2 calculado está normalizado y, por ello, es un valor comparable para todos los términos con una misma clase. Sin embargo, por esto mismo **tiene el inconveniente de que para términos poco frecuentes no es significativo** (Yang y Pedersen, 1997).

4.2.4.1.2.3. Pruebas

Además de eliminación directa de las palabras comunes, podemos suponer que en el corpus existen también otros **términos con una frecuencia de ocurrencia muy baja,** que también están aportando más ruido que información al modelo. En particular, como se ha mencionado, **pueden alterar el resultado de los métodos estadísticos de selección de características**, como ocurre con el test de la χ^2.

Por ello, antes de seleccionar las mejores, **se eliminaron del espacio de características los términos que sólo aparecen una vez** en todo el corpus.

Una vez hecho esto, se intentó mejorar los mejores modelos conseguidos hasta ahora seleccionando las ***k* mejores características usando las mencionadas técnicas estadísticas el test de la χ^2 y MI.** La Ilustración 26 incluye una comparativa de los resultados obtenidos con cada método de selección, si se selecciona un porcentaje determinado de las mejores características para NaiveBayesNLTK.

En la mencionada ilustración puede verse que **para NaiveBayesNLTK la selección óptima del porcentaje de características es el 50% por el test de la χ^2.**

Además, se comprueba que ambos métodos de selección de características se comportan de forma diferente: **el test de la χ^2** empieza por seleccionar términos poco frecuentes pero muy correlacionados con las clases objetivo; por eso el poder de clasificación con pocas características es muy limitado, pero va creciendo según se van incorporando indicadores de calidad al espacio de características hasta, de hecho, alcanzar **el mejor modelo**.

Por otra parte, la selección por **MI se comporta de forma muy estable desde los porcentajes más bajos**, al seleccionar los mejores indicadores desde el comienzo. Los modelos son un poco peores que con el test de la χ^2, pero mejoran el modelo base con el 100% de características.

La Ilustración 27, por su parte, muestra la comparativa equivalente a la anterior para los **modelos SVC**. En ella se observa que ocurre lo mismo que en el caso anterior: el pico de AUC se alcanza con **el 50% de las mejores características seleccionadas por** el test de la χ^2**.**

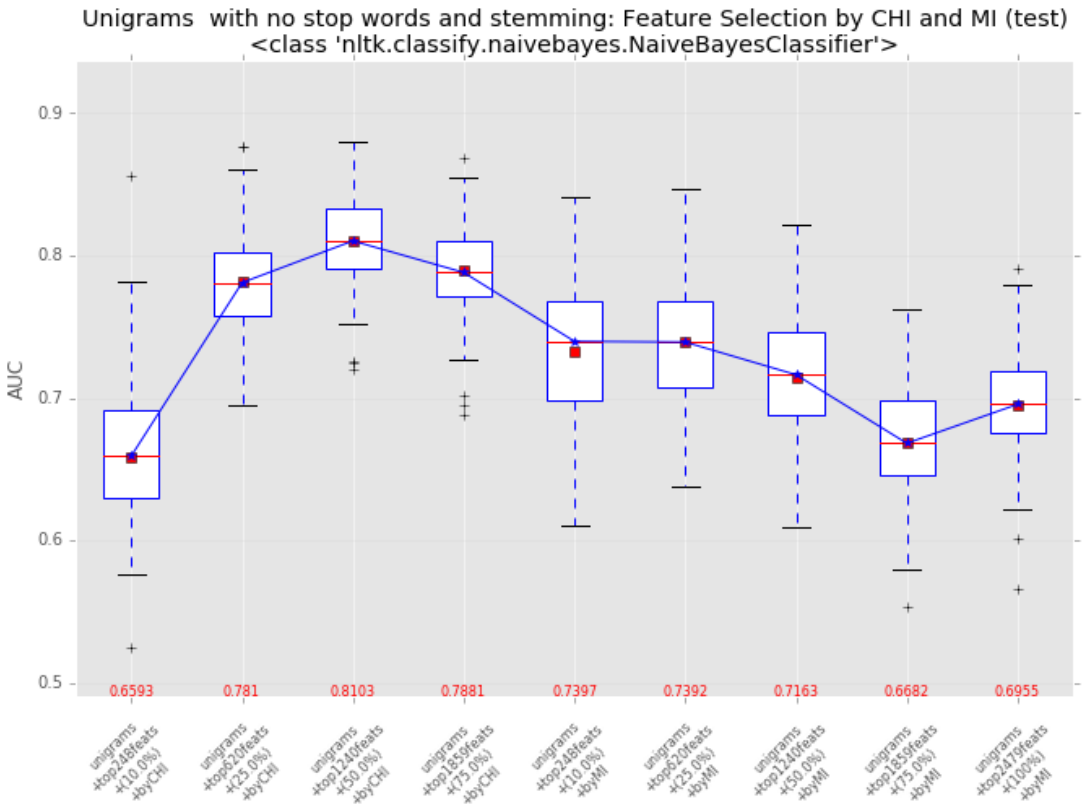


Ilustración 26: Comparación de la selección de características para NaiveBayesNLTK

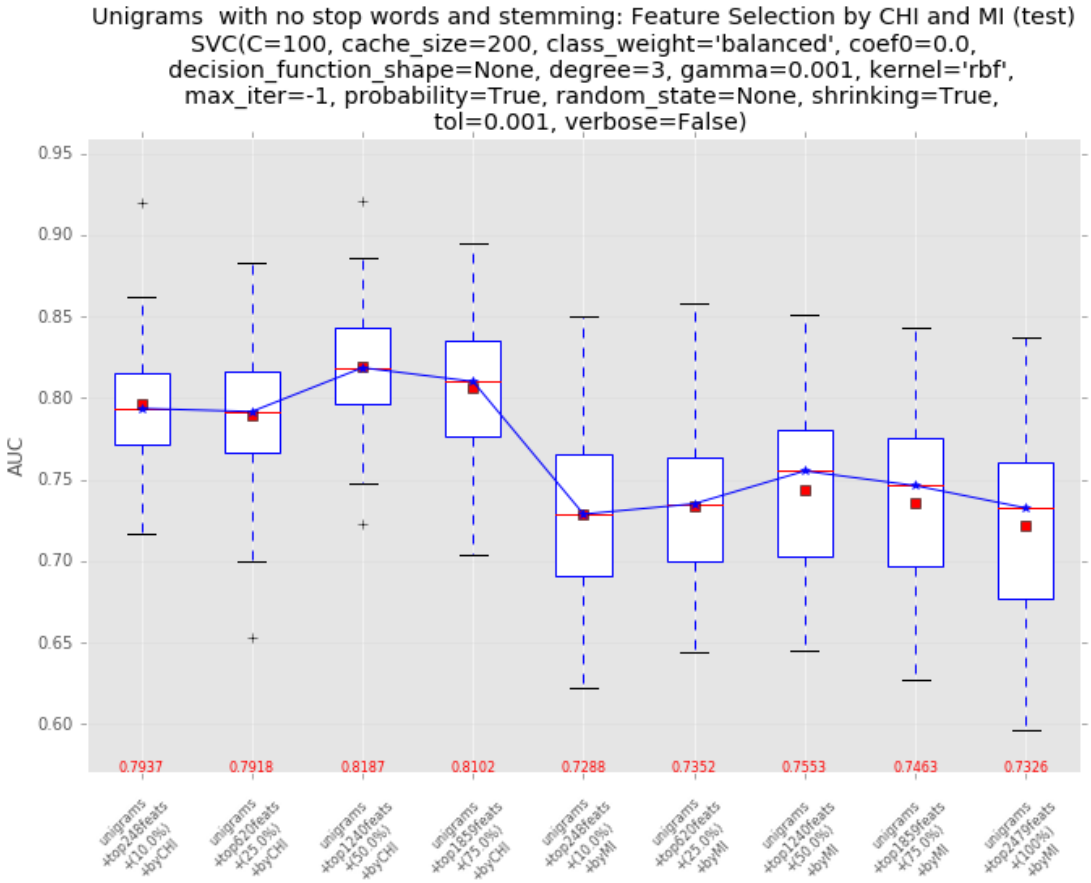


Ilustración 27: Comparación de selección de características para SVC

Si se comparan ambos modelos, puede verse que las medianas **de ambos están en torno al 0.80 de AUC,** si bien es verdad que los mejores modelos NaiveBayesNLTK tienen una varianza ligeramente inferior. Por ello, es complicado decir qué modelo es mejor, así que en este punto se introdujeron dos nuevos algoritmos para tener nuevas referencias:

- **Stochastic Gradient Descent (SGD)[19]:** Es un algoritmo basado en el método de optimización descenso del gradiente que se utiliza en minería de textos debido a su alta eficiencia para grandes conjuntos de datos.
- **RandomForest (RF)[20]:** Es un algoritmo perteneciente a la familia de los ensambladores de árboles de decisión. Suele dar bastante buen resultado para problemas diversos, debido a que controla muy bien el sobreajuste y proporciona una gran exactitud. Sin embargo, una alta dimensionalidad de los conjuntos de datos le afecta significativamente, debido a que los subespacios de características que crea son aleatorios para cada árbol, por lo que pueden incluir características poco informativas. En este caso, se incluyó para comprobar si el efecto de la selección de variables con los métodos estadísticos mencionados limita esa sensibilidad a la alta dimensionalidad sin disminuir su rendimiento predictivo.

De nuevo, ambos clasificadores se ajustaron por validación cruzada de 5 iteraciones sobre el conjunto de datos de entrenamiento.

A partir de ahí, se reprodujeron las comparativas anteriores para los nuevos algoritmos. Como puede verse en la Ilustración 28 y en la Ilustración 29.

De esta forma, puede observarse en la Ilustración 28, que en el caso de **SGD** ocurre lo mismo que en los casos anteriores: se alcanza el máximo AUC seleccionando el **50% de las mejores características por medio del** test de la χ^2**.** Sin embargo, en este caso el AUC está por encima de los anteriores, en torno a 0.85.

Respecto a los **modelos conseguidos con RF, se puede ver en la** Ilustración 29 que lo ideal sería **seleccionar el 75% de las características por medio del** test de la χ^2. Además, se observa que es el caso en el que la selección por el test de la χ^2 o por MI tiene menos impacto, probablemente debido a la naturaleza aleatoria del algoritmo.

[19] http://scikit-learn.org/stable/modules/sgd.html
[20] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

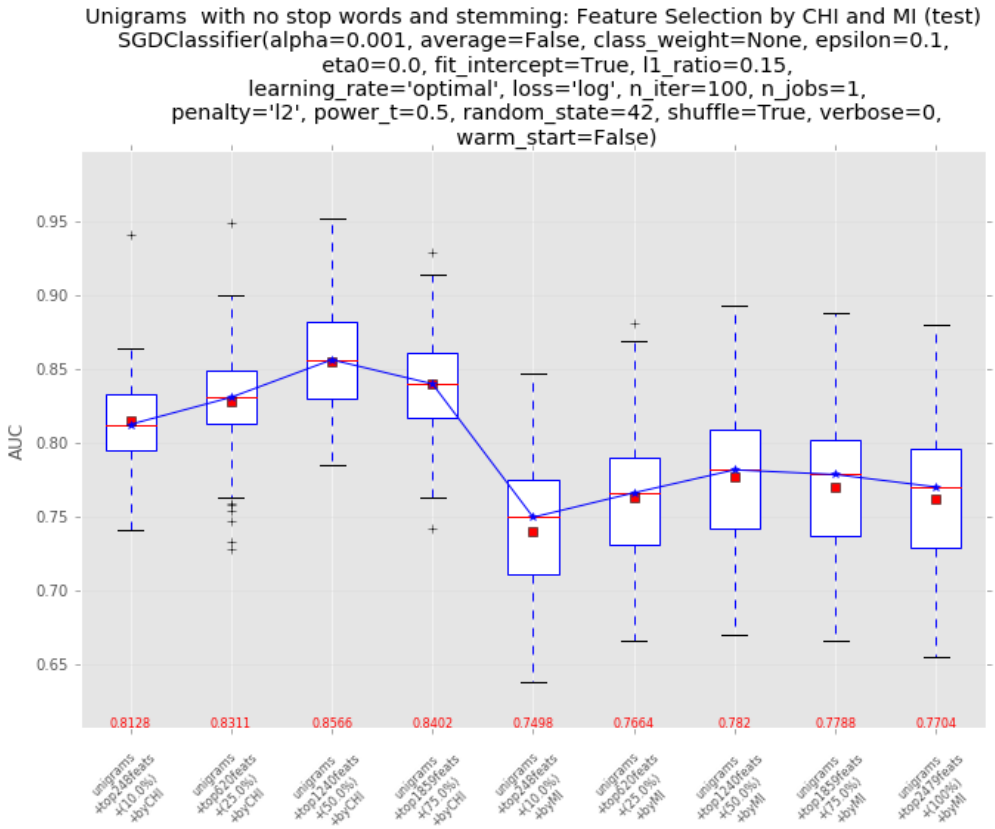


Ilustración 28: Comparación de selección de características para SGD

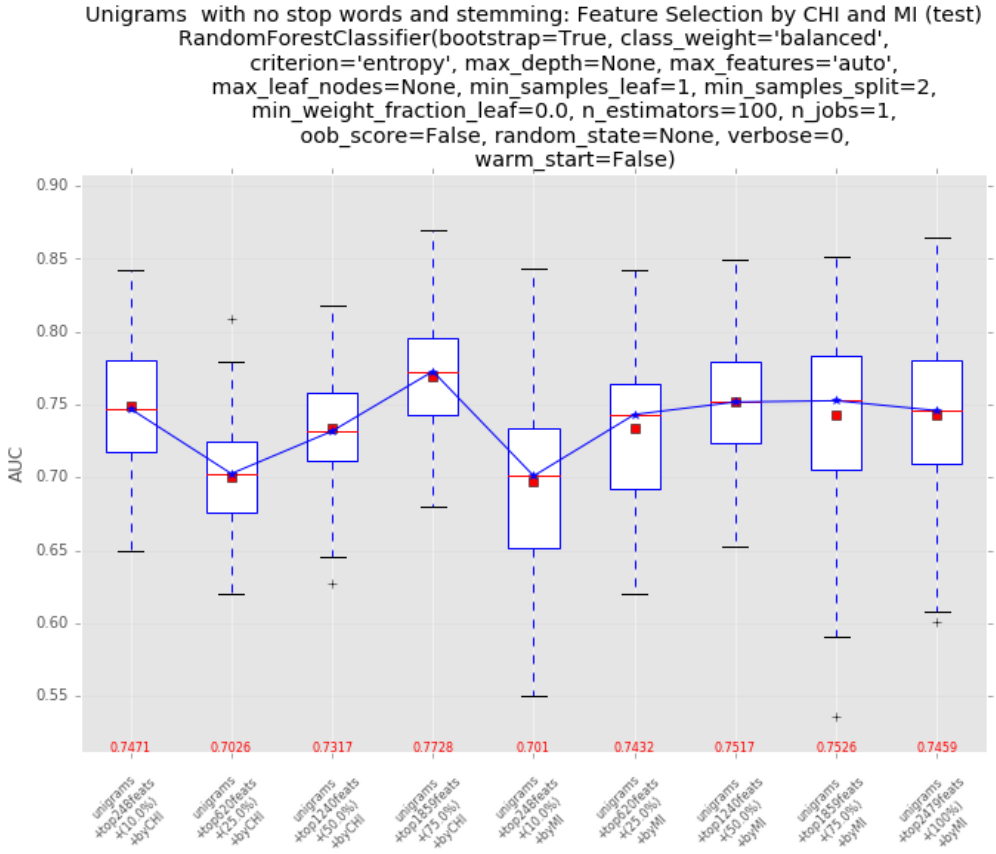


Ilustración 29: Comparación de selección de características para RF

Por tanto, con respecto a todos los modelos conseguidos, la selección por **el test de la χ^2 es la mejor opción** para la selección de las mejores características. Así, se comparan **las 100 mejores características obtenidas con ambas** por medio de las nubes de palabras o, en este caso, características, en las que el tamaño de la fuente está relacionado con su importancia según los test estadísticos. Puede verse **que las características son bastante diferentes, lo que explica la diferencia de comportamiento de los algoritmos.** Incluso puede confirmarse, como se menciona en la literatura (Manning et al., 2008), que **los términos más importantes para** el test de la χ^2 **parecen términos más específicos e infrecuentes,** aunque altamente indicativos, mientras que **los términos seleccionados por MI parecen ser más generales.**

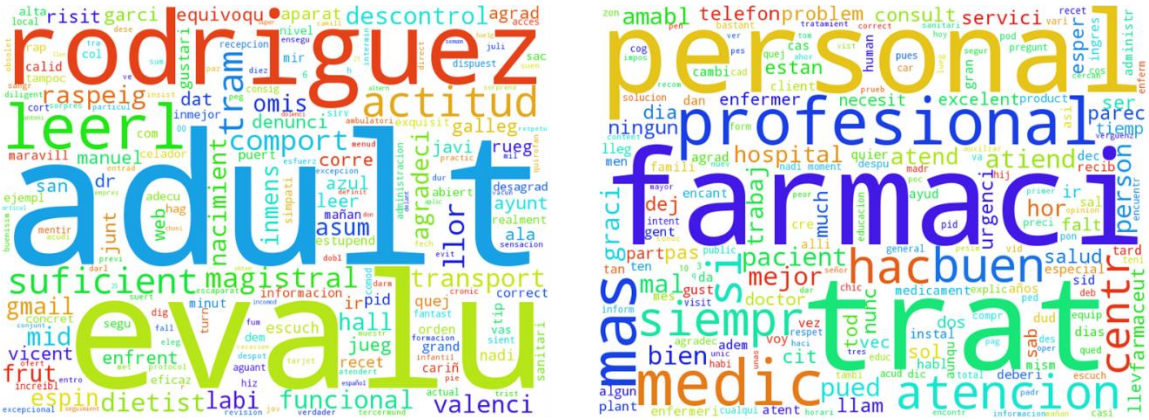


Ilustración 30: Comparativa de las 100 mejores características según el test CHI y la medida MI respectivamente

Por último comparamos los mejores modelos conseguidos según la validación cruzada repetida.

En la Ilustración 31 se ha incluido la comparación de los mejores modelos conseguidos hasta entonces junto con las distintas técnicas de selección de variables y eliminación de ruido. En ella puede observarse por un lado, que el **mejor modelo** conseguido es el **modelo SGD construido con el 50% de las mejores características,** obtenidas **a partir de los unigramas**; y por el otro, que **los modelos RF parece que son los que peor se comportan**.

No obstante, parece que **ninguno de los modelos conseguidos alcanza en promedio el 0.9 de AUC que marca el objetivo**. Por tanto, se siguió intentando mejorar los modelos ya conseguidos.

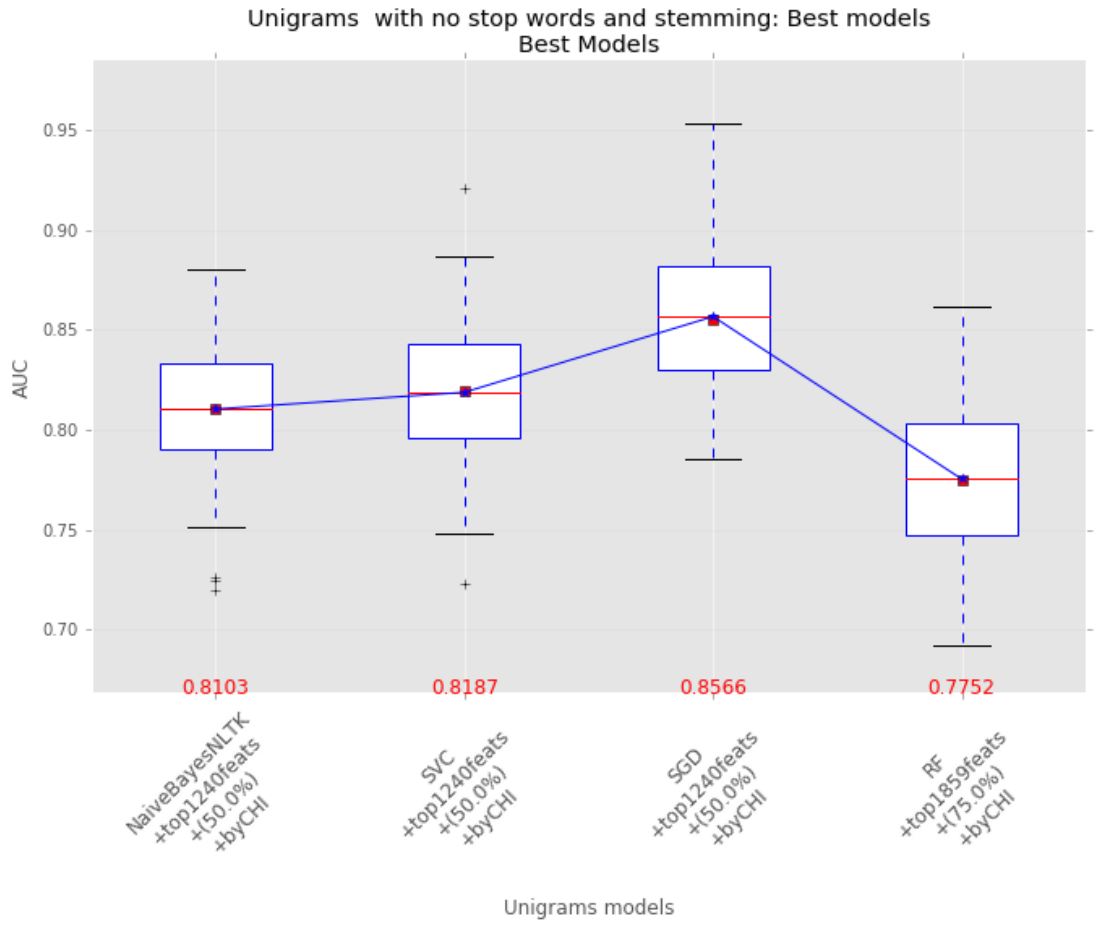


Ilustración 31: Mejores modelos conseguidos con los diferentes algoritmos

4.2.4.1.3. Aproximación con submuestreo

Debido a que las **clases objetivo** del problema están **muy desequilibradas (96% vs. 4%),** se intentó una aproximación al problema que utiliza **submuestreo de la clase mayoritaria** (relevante) para intentar corregir este sesgo inicial y mejorar los modelos.

Sin embargo, el submuestreo directo de la clase mayoritaria llevaría a perder información valiosa. Para solucionarlo se propuso **dividir el conjunto de datos inicial *C* compuesto por C_1 observaciones de la clase minoritaria (no relevante) y C_0 observaciones de la clase mayoritaria (relevante) en *K* conjuntos de datos donde cada uno contiene C_1 observaciones de la clase minoritaria y C_{0K} observaciones de la clase mayoritaria, y donde:**

$$C_{0K} = \frac{C_1\,(100-p)}{p} \quad \text{y} \quad K = \frac{C_0}{C_{0k}}$$

Una vez obtenidos **los *k* nuevos conjuntos de datos,** donde **la clase minoritaria original representa un porcentaje *p* de las observaciones**, entrenamos **un clasificador independiente para cada uno de ellos**, de forma que **resultaron *k* predicciones**. Finalmente, se propuso un **ensamblado** simple de estas predicciones **promediándolas para devolver una única predicción.** La Ilustración 32 muestra la arquitectura de la solución propuesta.

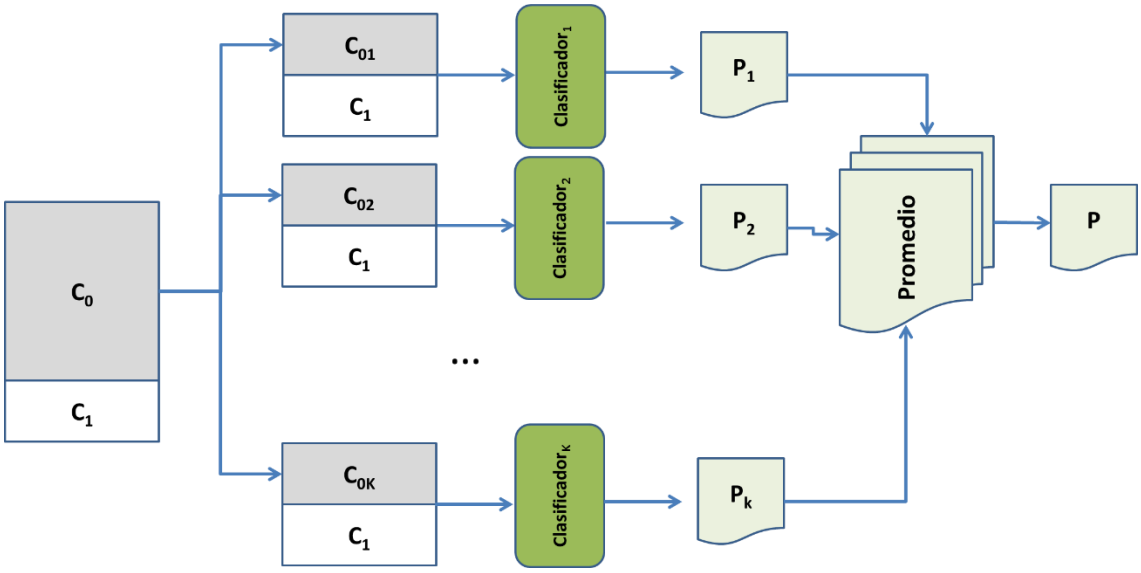


Ilustración 32: Arquitectura de la solución propuesta para reequilibrar las clases objetivo

4.2.4.1.3.1. Pruebas

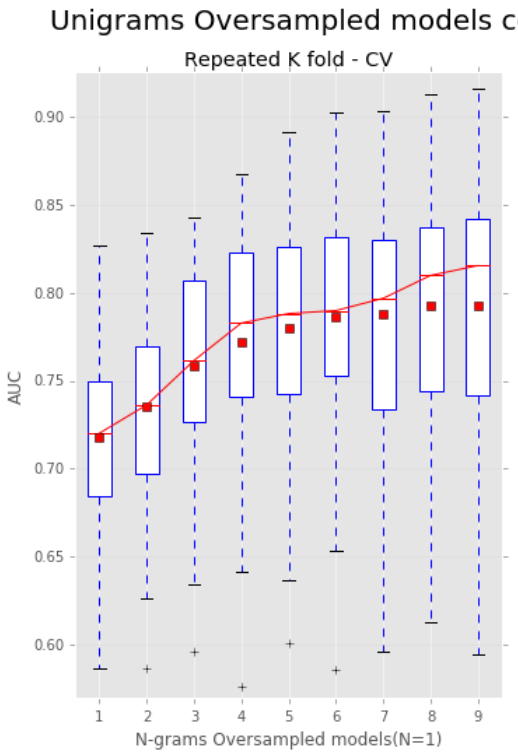
En primer lugar, tratamos de **ajustar el porcentaje *p* de la clase minoritaria** que contiene cada conjunto de datos submuestreado, en este caso cambiando a una **validación cruzada repetida de 10 iteraciones y 4 repeticiones, para intentar no perder demasiada información en cada entrenamiento y/o iteración.** Todo ello se hizo sin seleccionar un subconjunto de las características, es decir, usando el 100% de ellas (con una frecuencia de aparición mayor que 1).

En la Ilustración 34 puede verse el resultado de la comparación de los distintos porcentajes *p* para el algoritmo NaiveBayesNLTK.

Como puede verse en la mencionada ilustración, esta técnica de submuestreo **mejora el sesgo de los modelos NaiveBayesNLTK pero añade varianza**. Esto es debido a que los clasificadores individuales son más pequeños, según va creciendo el porcentaje de observaciones de la clase minoritaria desde 10 hasta 50. En particular, el problema se acentúa, pues el conjunto de datos inicial contiene un número relativamente reducido de observaciones.

En este caso, probablemente **el modelo con mejor equilibrio sesgo-varianza es el que se alcanza cuando p=30 o p=40**, es decir, cuando la clase minoritaria está representada con un 30% o un 40% de las observaciones.

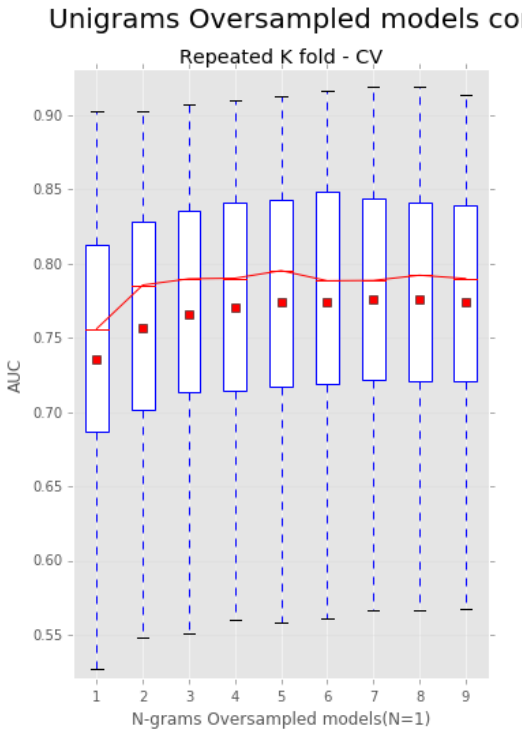
Repetimos el estudio con el resto de algoritmos, para ver cómo se comportan con esta técnica de submuestreo. En la Ilustración 34 vemos los resultados para los **modelos SVC**.



Models

	Median	Mean	Std.
1. NaiveBayesNLTK oversamp(no) 2479feats(100%)	0.7201	0.7177	0.0544
2. NaiveBayesNLTK oversamp(10%) 2479feats(100%)	0.7362	0.7354	0.0547
3. NaiveBayesNLTK oversamp(15%) 2479feats(100%)	0.7615	0.7585	0.0576
4. NaiveBayesNLTK oversamp(20%) 2479feats(100%)	0.7828	0.7719	0.063
5. NaiveBayesNLTK oversamp(25%) 2479feats(100%)	0.7882	0.78	0.0675
6. NaiveBayesNLTK oversamp(30%) 2479feats(100%)	0.7899	0.7865	0.0678
7. NaiveBayesNLTK oversamp(35%) 2479feats(100%)	0.7969	0.7881	0.0731
8. NaiveBayesNLTK oversamp(40%) 2479feats(100%)	0.8099	0.7926	0.0759
9. NaiveBayesNLTK oversamp(50%) 2479feats(100%)	0.8155	0.793	0.0795

Ilustración 33: Comparación de porcentajes *p* para el submuestreo utilizando el modelo NaiveBayesNLTK



Models

	Median	Mean	Std.
1. SVC oversamp(no) 2479feats(100%)	0.7561	0.7357	0.0982
2. SVC oversamp(10%) 2479feats(100%)	0.7855	0.7565	0.0943
3. SVC oversamp(15%) 2479feats(100%)	0.7897	0.7656	0.0914
4. SVC oversamp(20%) 2479feats(100%)	0.7901	0.7708	0.0891
5. SVC oversamp(25%) 2479feats(100%)	0.7952	0.7737	0.0897
6. SVC oversamp(30%) 2479feats(100%)	0.7884	0.7745	0.0892
7. SVC oversamp(35%) 2479feats(100%)	0.7886	0.7759	0.0883
8. SVC oversamp(40%) 2479feats(100%)	0.7921	0.7755	0.0873
9. SVC oversamp(50%) 2479feats(100%)	0.7899	0.7739	0.0862

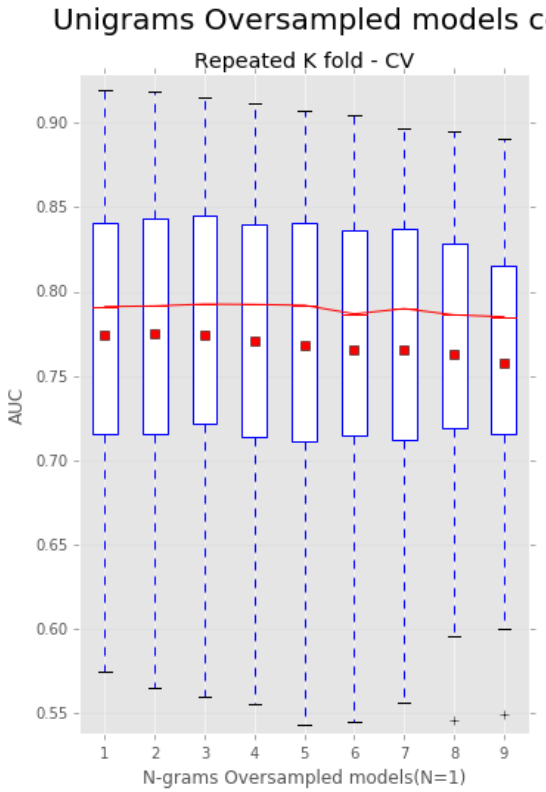
Ilustración 34: Comparación de porcentajes *p* para el submuestreo utilizando el modelo SVC

En el caso de **los modelos con SVC, parece que la técnica del submuestreo también mejora los resultados originales**, aunque en menor medida. Además, todos los modelos se comportan de forma similar con independencia del porcentaje *p* elegido. Por ello, es más difícil seleccionar el mejor modelo sin más información. Así que se siguieron ajustando los modelos

mediante la selección de las mejores características según el test de la χ2 y la medida MI, como se hizo anteriormente.

En el caso de **SGD** (véase la Ilustración 35) **se observa que la técnica de submuestreo no mejora el modelo base**, pues la **relación sesgo-varianza es claramente peor.** Probablemente el motivo es que, al ser un algoritmo tan dependiente del ajuste de parámetros, éstos deban ser reajustados a las nuevas dimensiones de los conjuntos de datos. Sin embargo, es una operación muy costosa computacionalmente; por ello, se deja como mejora optimizar el algoritmo para cada porcentaje *p.*

Por último, **en el caso de RF** (véase la Ilustración 36**) tampoco se observó una clara mejora del modelo sin submuestreo.** El motivo podría ser que el propio algoritmo RandomForest ya tiene incorporado este proceso de submuestreo internamente y, por tanto, aunque aleatoriamente, éste ya podría llevar incorporada la posible mejora que puede suponer el submuestreo.



Models

	Median	Mean	Std.
1. SGD oversamp(no) 2479feats(100%)	0.7911	0.7741	0.0903
2. SGD oversamp(10%) 2479feats(100%)	0.7915	0.7755	0.0905
3. SGD oversamp(15%) 2479feats(100%)	0.7926	0.7747	0.0904
4. SGD oversamp(20%) 2479feats(100%)	0.7924	0.7708	0.0886
5. SGD oversamp(25%) 2479feats(100%)	0.7919	0.7685	0.0899
6. SGD oversamp(30%) 2479feats(100%)	0.7868	0.7657	0.0894
7. SGD oversamp(35%) 2479feats(100%)	0.7899	0.7653	0.0871
8. SGD oversamp(40%) 2479feats(100%)	0.7862	0.7632	0.0863
9. SGD oversamp(50%) 2479feats(100%)	0.7851	0.7577	0.0825

Ilustración 35: Comparación de porcentajes *p* para el submuestreo utilizando el modelo SGD

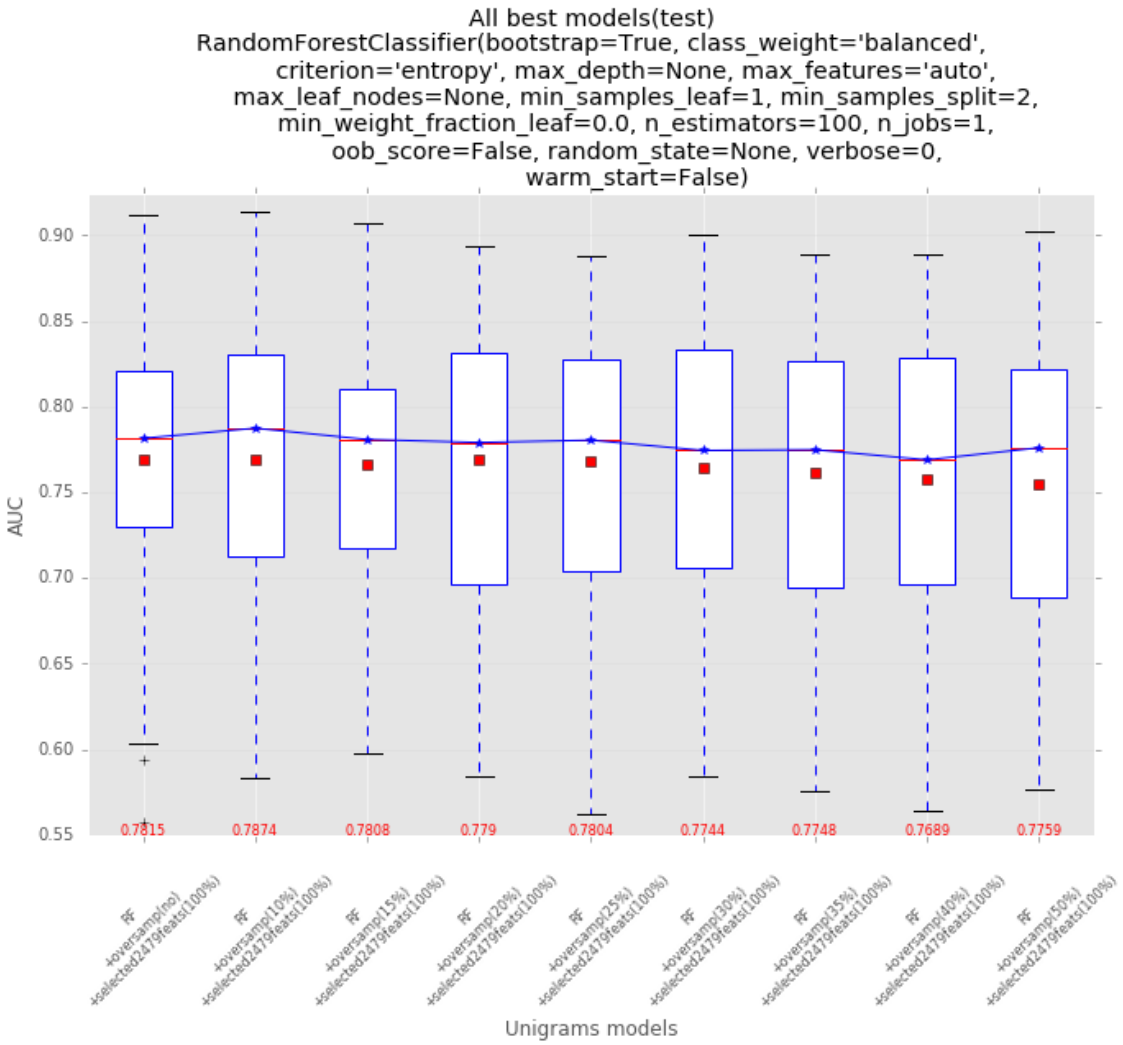
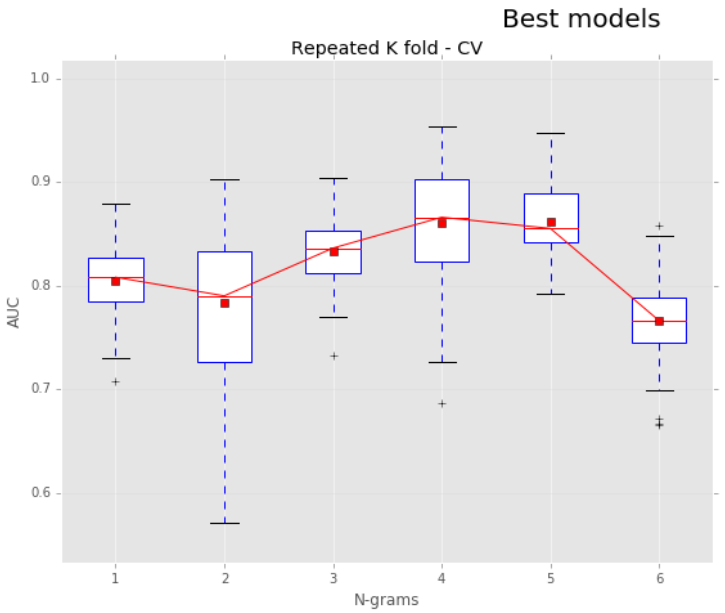


Ilustración 36: Comparación de porcentajes *p* para el submuestreo utilizando el modelo RF

Como se puede observar en los gráficos anteriores, **la varianza de los modelos** mostrada con la validación cruzada repetida con submuestreo **no es aceptable**, sobre todo comparada con la varianza de los modelos sin submuestreo. Además, excepto en el caso de NaiveBayesNLTK, **se aprecia una clara separación entre la mediana y la media** en el resto de modelos, lo que es indicativo de que la validación cruzada produce iteraciones en las que la predicción es muy mala.

Esto podría ser debido a que **el conjunto de datos disponibles para entrenamiento puede resultar insuficiente según se van equilibrando las observaciones de las clases objetivo, dando como resultado modelos entrenados con muy pocas observaciones (y por tanto, con un vocabulario muy limitado), que no generalizan muy bien.** Además, esto se acentúa en función del algoritmo elegido para el modelado.

No obstante, incorporamos los mejores modelos conseguidos con submuestreo a los anteriores, para comparar sus resultados (véase la Ilustración 37)



Models

	Median	Mean	Std.
1. NaiveBayesNLTK top1240feats (50.0%) byCHI	0.8079	0.8041	0.0347
2. NaiveBayesNLTK oversamp(30%) 2479feats(100%)	0.7899	0.7828	0.0719
3. SVC top1240feats (50.0%) byCHI	0.836	0.8335	0.0304
4. SVC oversamp(50%) 1240feats(50.0%)	0.8658	0.8601	0.0548
5. SGD top1240feats (50.0%) byCHI	0.855	0.8621	0.0321
6. RF top1859feats (75.0%) byCHI	0.766	0.7661	0.0408

Ilustración 37: Mejores modelos obtenidos con submuestreo o sin submuestreo para el Clasificador de Relevancia

Como se observa en la mencionada ilustración, **el mejor modelo en cuanto a sesgo es el obtenido con SVC y un *p* = 50, donde además se ha seleccionado sólo el 50% de las características con el test de la χ2.**

En cuanto a **la relación sesgo-varianza, el mejor sería el conseguido con SDG seleccionando el 50% de las características mediante el test de la χ2.**

Por otra parte, también se ve que **los dos modelos con submuestreo (2 y 4) tienen una varianza mucho mayor.**

4.2.4.2. Clasificador de Polaridad

El Clasificador de la Polaridad de Senti-Sys también resuelve un problema de **clasificación binaria, en este caso con** el objetivo **de clasificar las frases en negativas y no negativas** utilizando los métodos de aprendizaje automático supervisado seleccionados a partir del **corpus de frases anotado**.

Ahora bien de cara a la construcción de este segundo clasificador, la distribución de las clases está bastante equilibrada en el corpus de entrenamiento: el 52.26% de las observaciones tiene clase "negativa" y el 47.74% clase "no-negativa".

En este caso, el algoritmo de partida para la fase de **modelado** fue ***Naive Bayes*, nuevamente** por medio de la librería NLTK.

Para poder **comparar los resultados** de las diferentes pruebas y minimizar el sobreajuste de los modelos, se decidió emplear nuevamente **validación cruzada repetida con 4 iteraciones y 20 repeticiones**. En este caso sin estratificar.

4.2.4.2.1. Extracción de características (feature extraction)

Como en los anteriores casos, el primer paso era **transformar cada opinión en un vector de palabras,** de tal forma que se pudiera aplicar los algoritmos sobre esta nueva matriz.

En particular, en este caso, también **se usaron n-gramas** como generalización de las palabras. Por tanto, los *tests* de este apartado se enfocaron en averiguar la combinación óptima de estos n-gramas para crear las características del conjunto de datos.

4.2.4.2.1.1. Pruebas

En el primer test se pretendía observar el AUC alcanzado con modelos cuyas características se construyen combinando los diferentes n-gramas: unigramas, bigramas y/o trigramas. El resumen gráfico de los resultados obtenidos se muestra en la Ilustración 38.

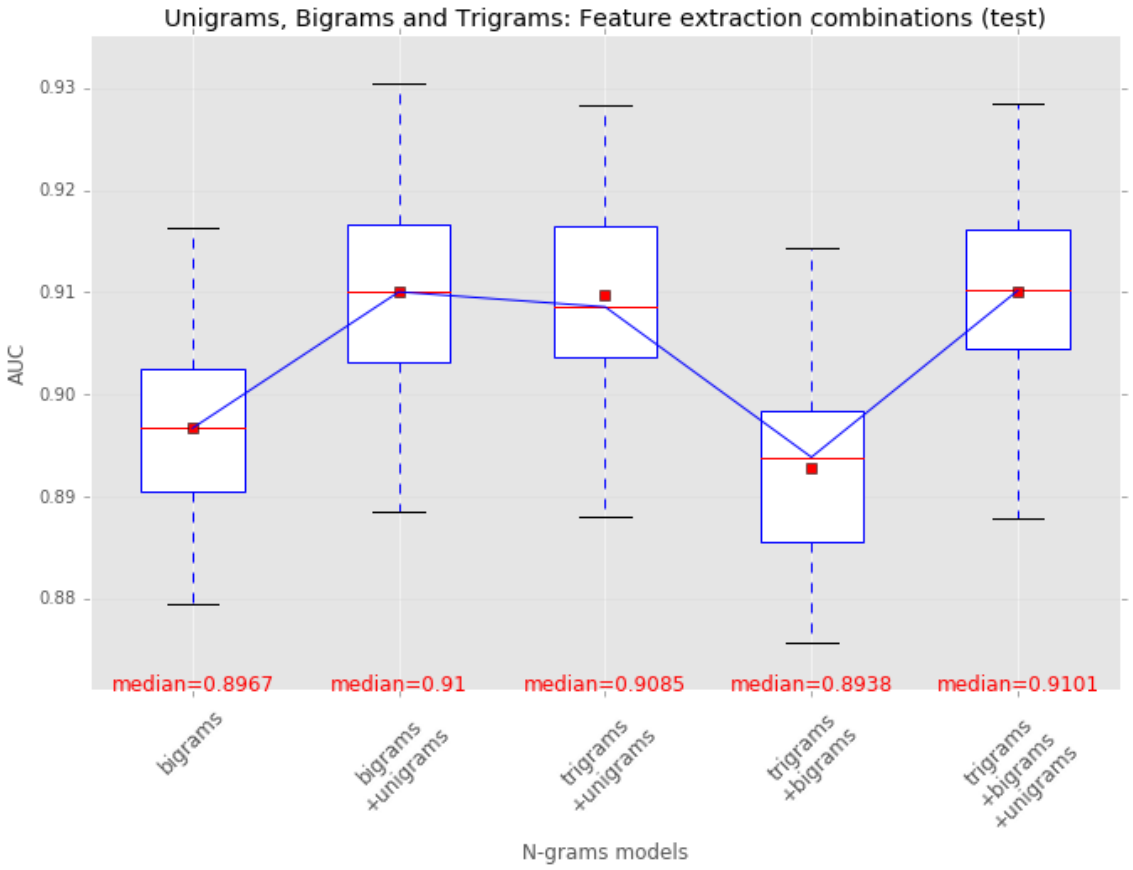


Ilustración 38: Comparativa de unigramas, bigramas y trigramas combinados

De esta forma, se vio que **combinando los n-gramas se mejoraba el modelo base.** En particular, el modelo conseguido con **unigramas y bigramas** se comporta ligeramente mejor que el resto.

Por tanto, se procedió a intentar optimizar los dos mejores modelos: **bigramas+unigramas y unigramas+bigramas+trigramas**

A continuación, sobre éstos, se hizo uso tanto de la técnica de **eliminación de palabras comunes** (***stop words, SW***) como de **reducción a la raíz** (***stemming, ST***). Sin embargo, ambas se pueden aplicar tanto antes de crear los bigramas y trigramas como después. Por ello, se probaron las distintas combinaciones (véase la Ilustración 39)

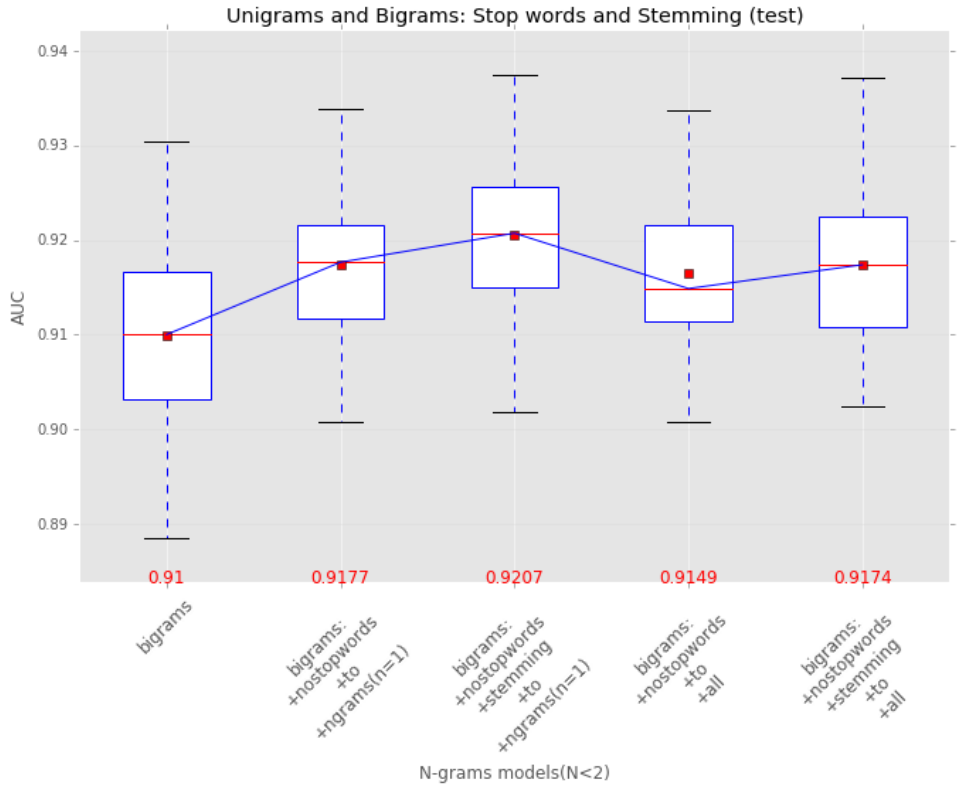


Ilustración 39: Comparación de uso de SW y ST aplicados antes o después de crear los bigramas

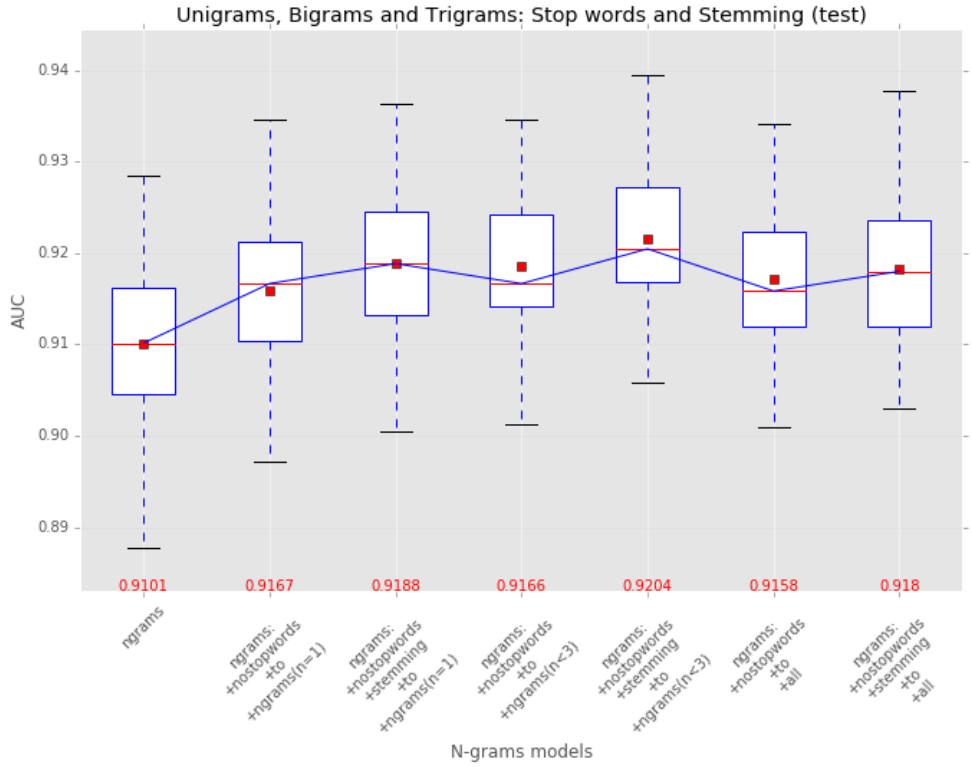


Ilustración 40: Comparación de uso de SW y ST con los diferentes n-gramas

En primer lugar, se observa **que tanto la eliminación de palabras comunes como la reducción a la raíz rebajan el ruido en todos los casos**, ya que se mejoran los modelo base (unigramas+bigramas y unigramas+bigramas+ trigramas) en cuanto a reducción de varianza.

En segundo lugar, puede apreciarse que **los mejores modelos** se consiguen **aplicando las mencionadas técnicas tanto a unigramas como a bigramas,** modelo ngrams+stopwords+stemming+to+ngrams (n<3)**, dejando los trigramas completos**, aunque la diferencia es mínima (AUC + 0.002)

Por tanto, quedó demostrado demostrado que ambas técnicas mejoran los modelos, sin embargo, es difícil escoger cual modelo es el mejor, dadas las escasas diferencias, así que **se probaron las técnicas de selección de características sobre los tres modelos que usan tanto *SW* como *ST*.**

4.2.4.2.2. Selección de variables (Feature Selection)

Debido a la técnica seleccionada para crear características (vector de palabras) nuestro conjunto de datos tiene **una dimensionalidad muy alta,** dependiende del vocabulario del corpus, como se pueden observar en Ilustración 41.

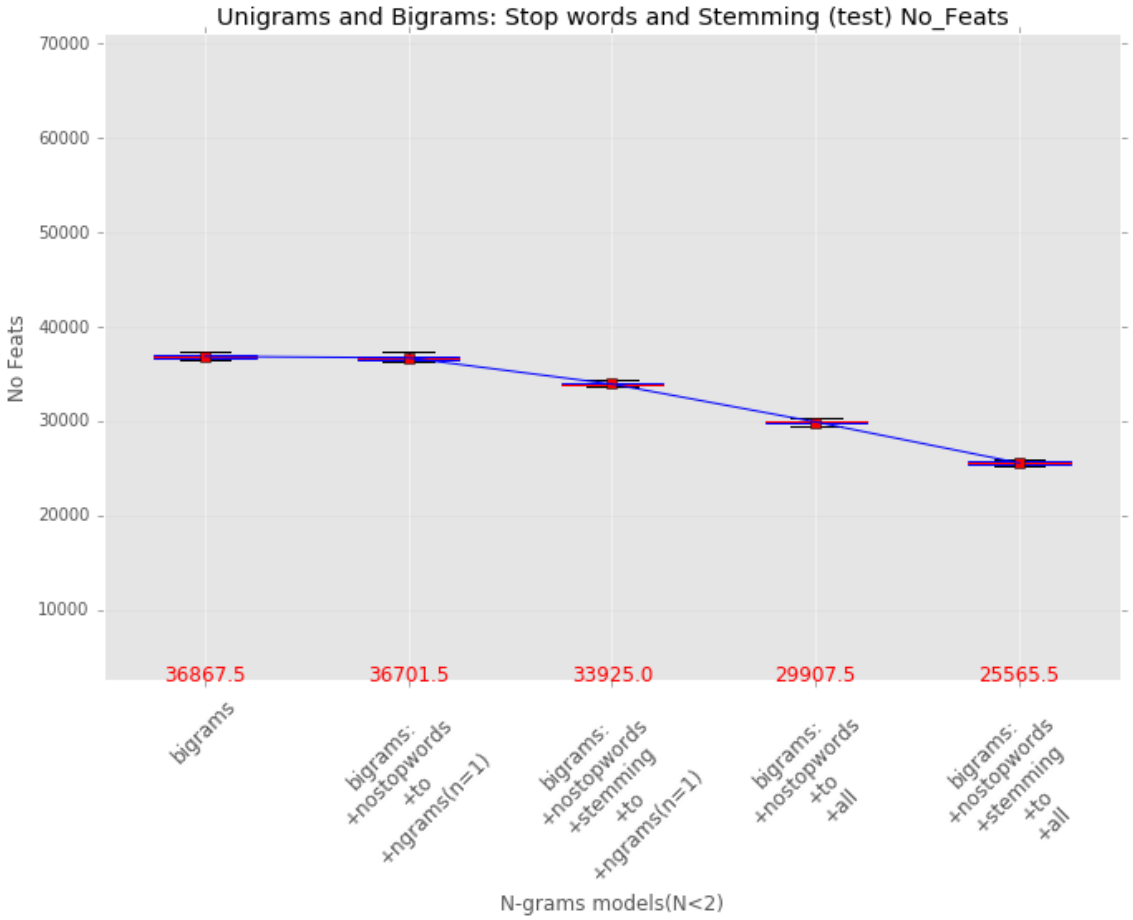


Ilustración 41: Comparación de la dimensionalidad con los diferentes n-gramas

Una **alta dimensionalidad no es un problema en sí mismo**, más allá del aumento de recursos técnicos necesarios para procesar los modelos en función del algoritmo empleado, sin embargo, sí puede ser un problema si implica que el conjunto de características creado incorpora ruido al modelo.

Por ello, en primer lugar, eliminamos como características los términos que sólo aparecen una vez en todo el corpus. De esta forma, el número de características bajó drásticamente, sin afectar al poder de predicción, como se aprecia en la Ilustración 42.

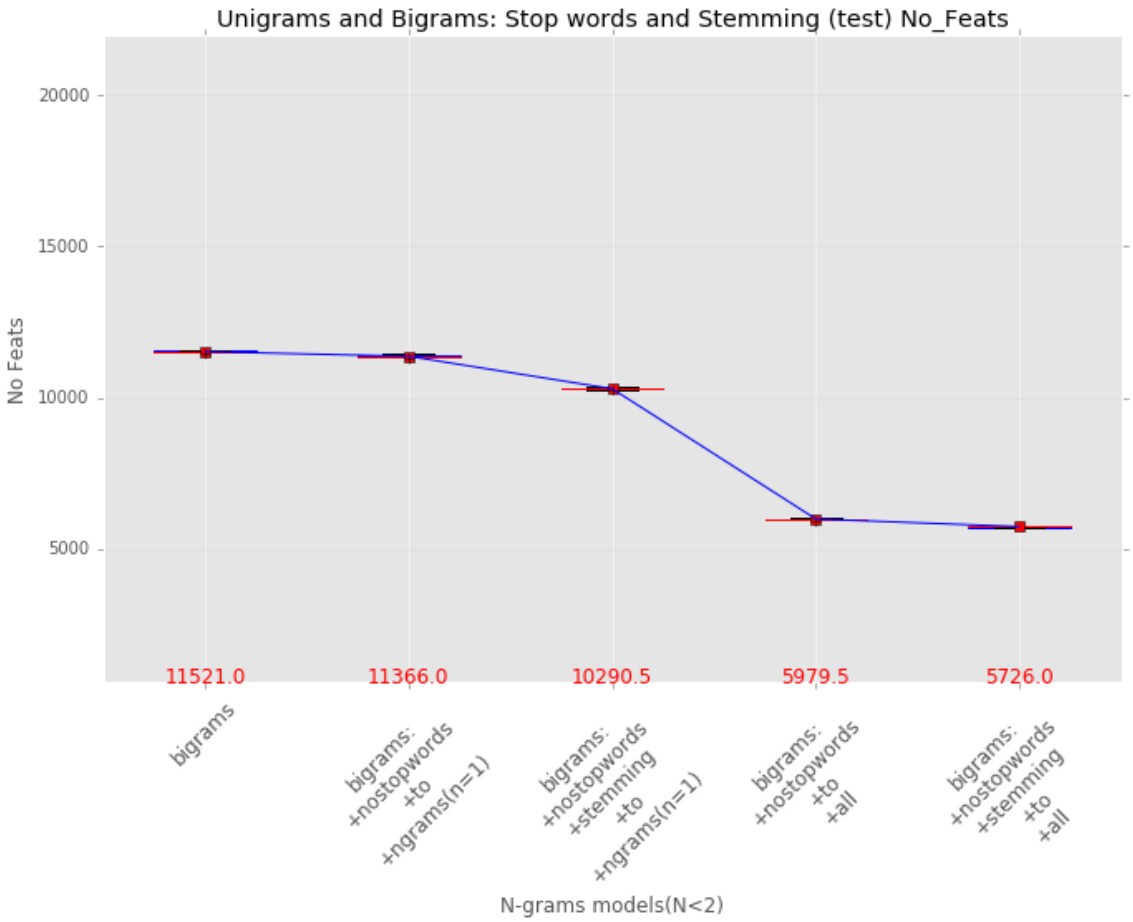


Ilustración 42: Comparación de la dimensionalidad poniendo un límite inferior de frecuencia a 2 apariciones

Una vez hecho esto, al igual que con el Clasificador de Relevancia, se probaron los métodos ya descritos de selección de variables mediante técnicas estadísticas, para comprobar si reduciendo la dimensionalidad se mejoraba el rendimiento del mejor modelo.

Además, se intentó **combinar ambas técnicas de selección de variables, test de la χ^2 y MI**, creando una **nueva función de selección**, que llamaremos **MIX** en el contexto del presente trabajo, para comprobar si se mejoraban los resultados conseguidos. Esta función fue comparada con las anteriores mediante las pruebas de validación cruzada repetida.

En particular, **los pasos para calcular la medida MIX** con la que ordenar las características de mejor a peor serían:

1. Se calcula el estadístico χ^2 para todas las características.
2. Se reordenan las características por el estadístico χ^2 de menor a mayor.
3. Se calcula la medida MI para todas las características.
4. Se reordenan las características por la medida MI de menor a mayor.
5. Para cada característica, se promedia la posición que ocupa según las anteriores medidas.
6. Se reordenan las características de mayor a menor, en función del anterior promedio.

Una vez calculada la medida, se utiliza el algoritmo de selección de las *k* mejores variables visto con anterioridad.

4.2.4.4.2.2.1. Pruebas

Para **seleccionar las *k* mejores características**, se empleó también validación cruzada repetida utilizando el **test de la χ^2** y las medidas ***MI*** y ***MIX***.

En primer lugar se estudió la selección de características sobre modelo donde se hace ***stop words*** (SW) y ***stemming*** (ST) sobre los unigramas únicamente, dejando íntegros los bigramas y trigramas. La comparativa de los resultados obtenidos puede verse en Ilustración 43.

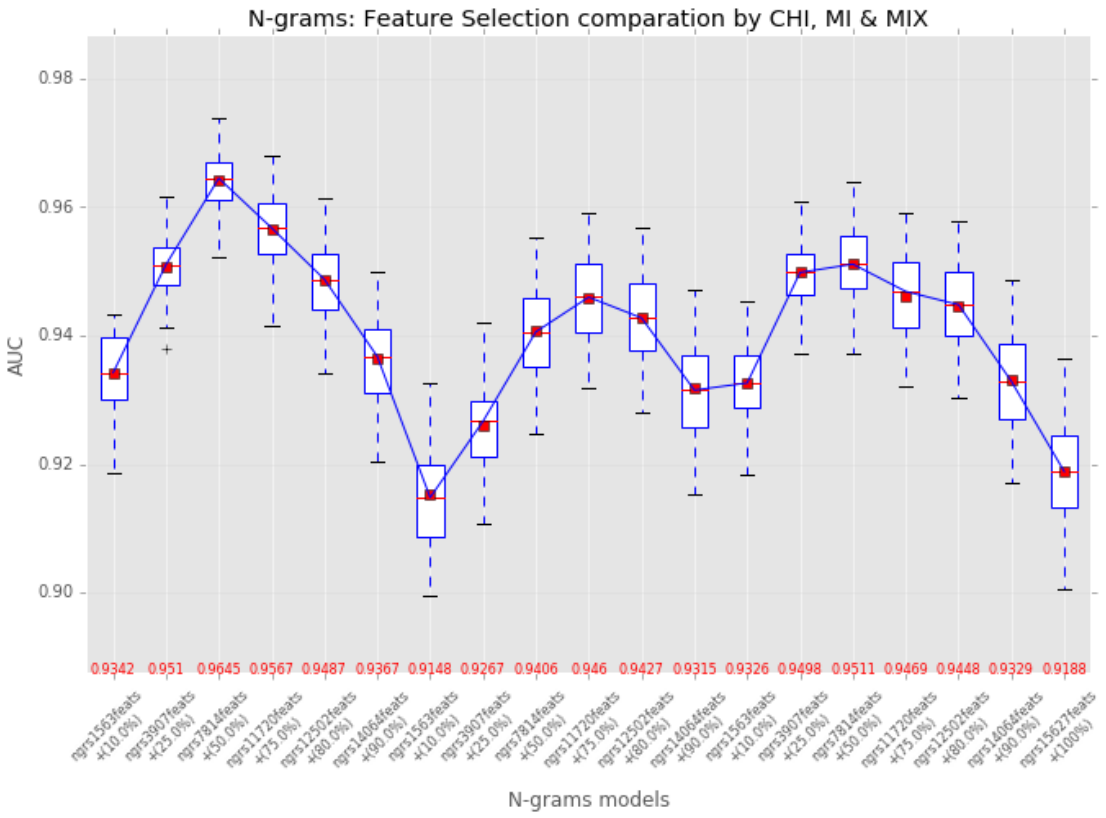


Ilustración 43: Comparativa de selección de k% mejores características con SW y ST aplicados cuando N = 1

De esta forma, en la anterior ilustración, puede observarse que, en ambos casos, **el mejor modelo se consigue cuando se seleccionan el 50% de las mejores características utilizando el test de la χ^2 para localizarlas.**

A continuación, se repitió el mismo estudio con el modelo donde se aplican ***SW*** y ***ST*** tanto a unigramas como a bigramas, dejando los trigramas completos (véase la Ilustración 44).

En este caso puede apreciarse que el patrón de comportamiento es exactamente el mismo que en el test anterior.

Por último, repetimos el estudio aplicando las técnicas SW y ST a todos los n-gramas, obteniéndose unos resultados similares, aunque ligeramente inferiores, que pueden observarse en la Ilustración 45.

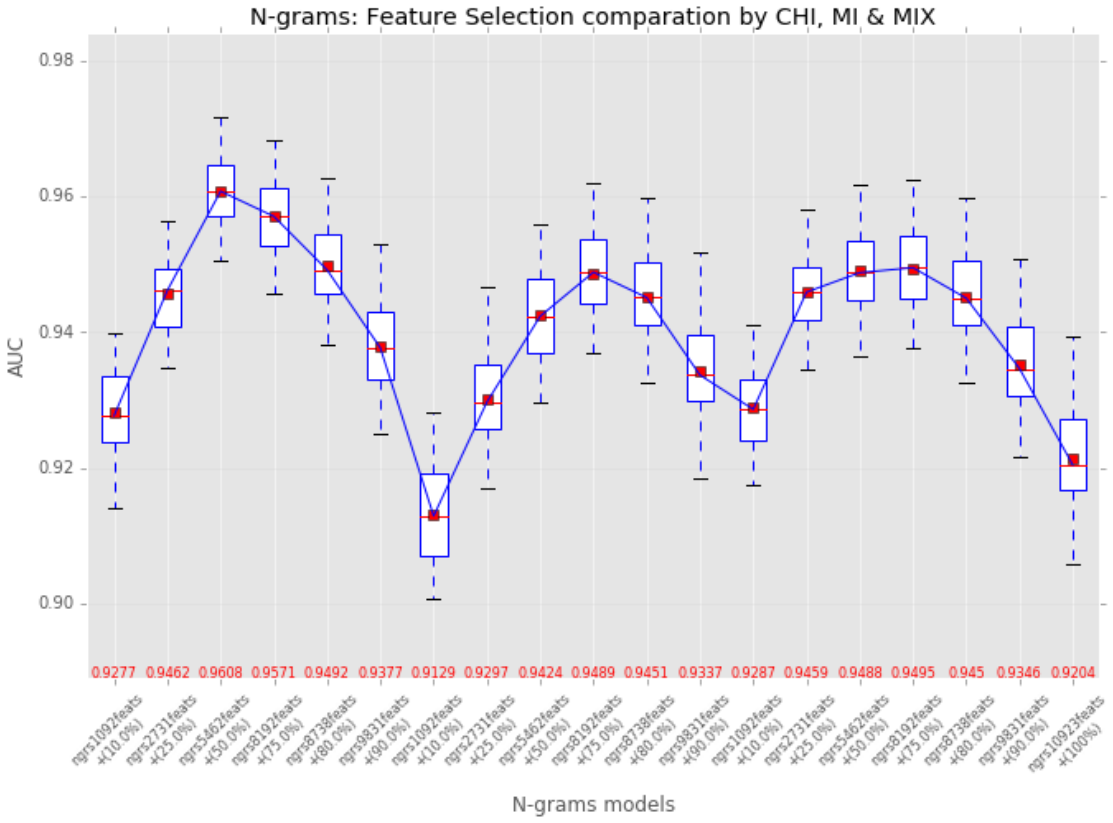


Ilustración 44: Comparativa de selección de k% mejores características con SW y ST aplicados cuando N <3

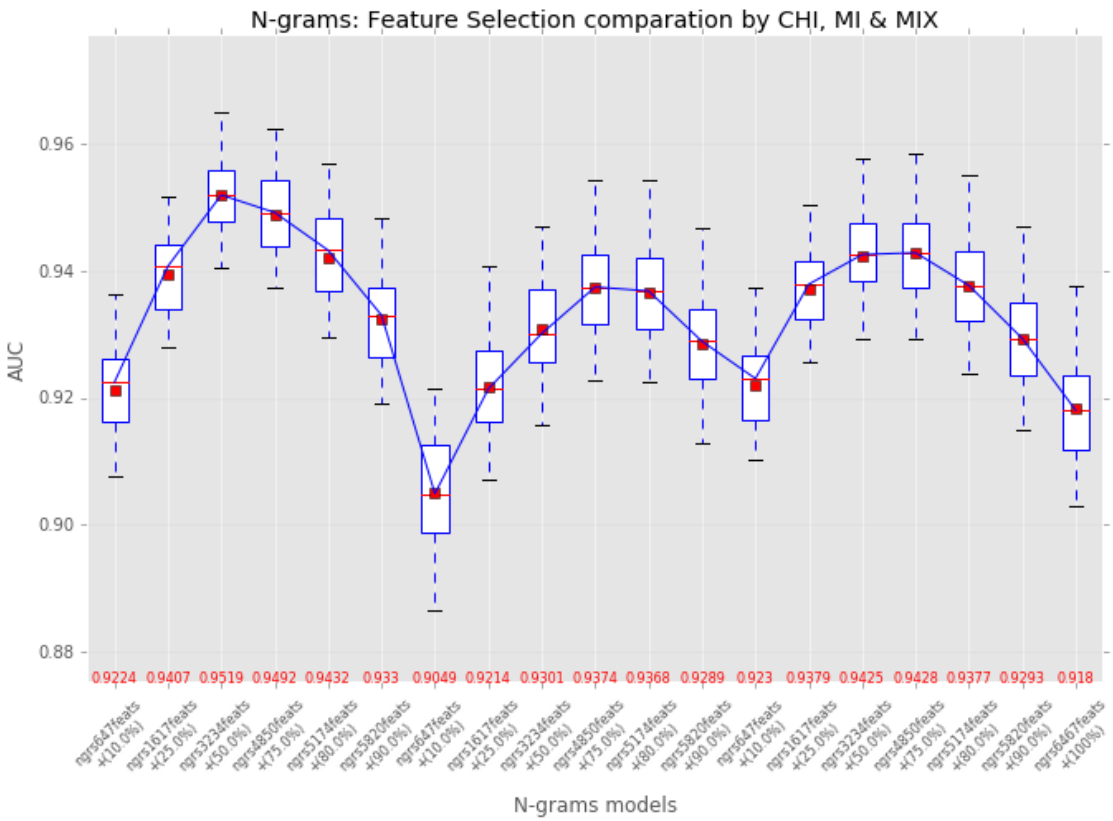


Ilustración 45: Comparativa de selección de k% mejores características con SW y ST sobre todos los n-gramas

Por otra parte, se observó que en todos los casos anteriores la **selección por el test de la χ^2 es claramente la mejor opción**, seguida de la medida MIX y la función MI (por ese orden). Por tanto, en este caso **la combinación del test de la χ^2 y MI no aporta más información que la mejor técnica, el test de la χ^2**.

Finalmente, se repitió el estudio de selección de características con los modelos construidos usando únicamente unigramas y bigramas para crear características, y se compararon todos los mejores modelos obtenidos (véase la Ilustración 46).

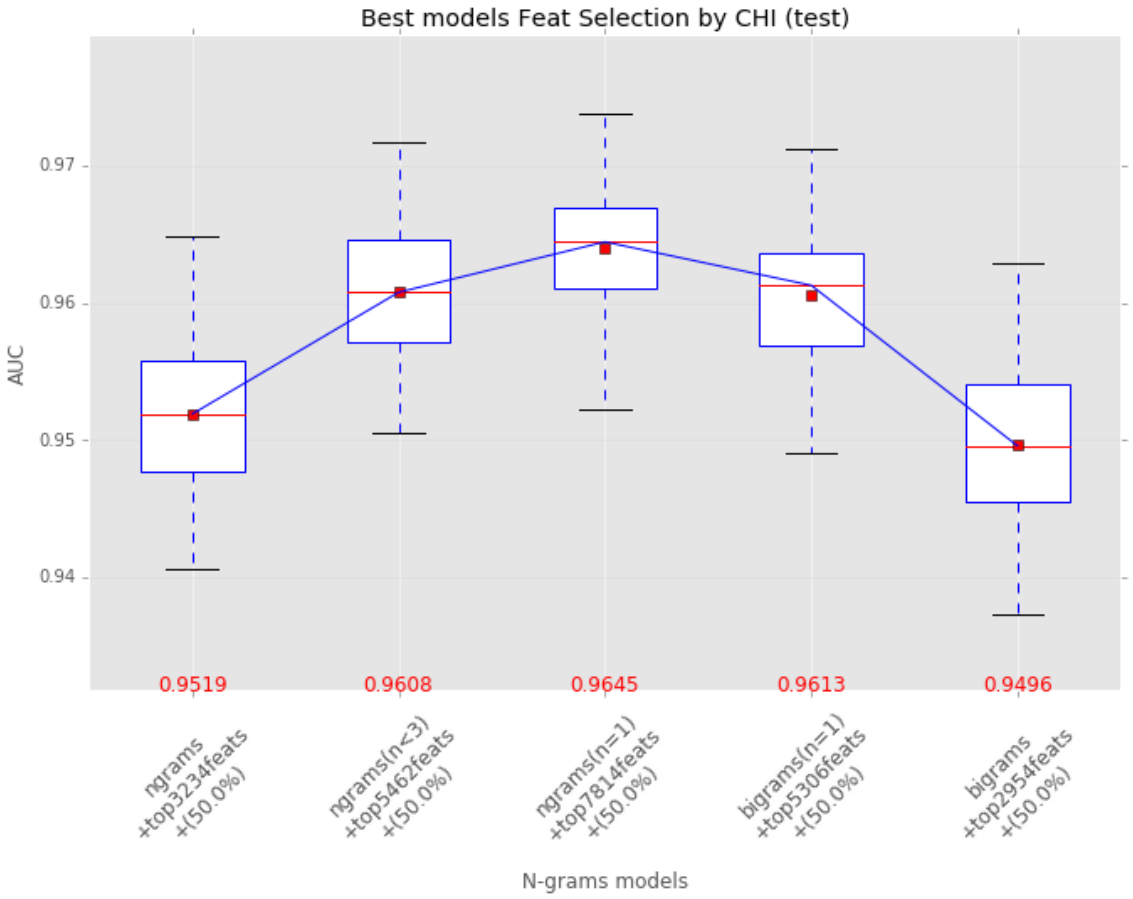


Ilustración 46: Comparación de los mejores modelos conseguidos con n-gramas y bigramas

De esta forma, comprobamos que el **mejor modelo** de todos, según la validación cruzada repetida, es el que se consigue mediante **n-gramas aplicando SW y ST solamente a los unigramas y seleccionando el 50% de las mejores características mediante el test de la χ^2.**

Además, observando las 100 mejores características del mejor modelo con una nube de palabras (incluida en la Ilustración 47), se aprecia que la mayoría son términos, (o combinaciones de ellos) habitualmente cargados de polaridad (excelente, gracias, amable, encantador, muy bien, mal). Por otra parte, también aparecen términos que en un contexto general son neutros, pero en nuestro corpus están más relacionados con opiniones positivas (farmacia) o con opiniones negativas (teléfono).

Ilustración 47: Top 100 características del mejor modelo según el test de la χ^2

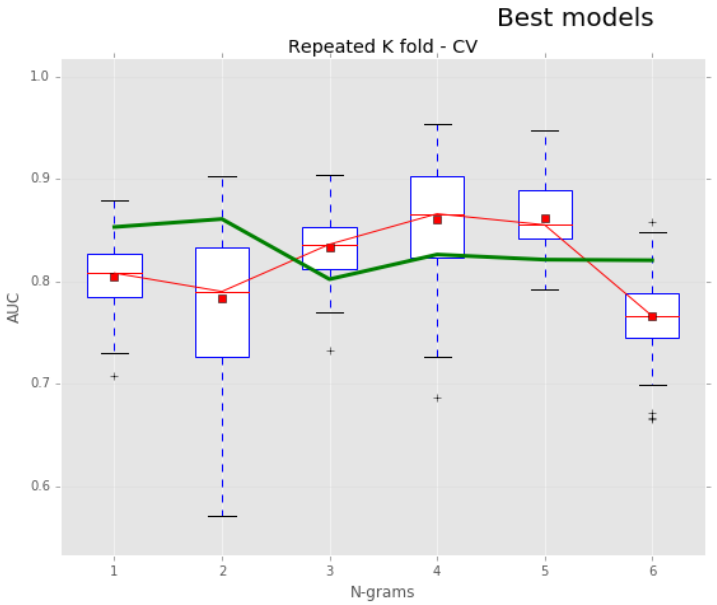
4.2.5. Evaluación de resultados

En la fase de evaluación de resultados se comprueba el grado en que los modelos definitivos responden a los objetivos de negocio marcados según el grado de cumplimiento de los criterios de éxito marcados.

Para ello, para los mejores modelos de cada subsistema se procederá a verificar sus rendimientos contra los datos del conjunto de evaluación

4.2.5.1. Clasificador de Relevancia

Los resultados de la evaluación de los mejores modelos obtenidos para el Clasificador de Relevancia con o sin submuestreo se han incluido en la Ilustración 48.



Models

	Median	Mean	AUC
1. NaiveBayesNLTK top1240feats (50.0%) byCHI	0.8079	0.8041	0.8529
2. NaiveBayesNLTK oversamp(30%) 2479feats(100%)	0.7899	0.7828	0.8607
3. SVC top1240feats (50.0%) byCHI	0.836	0.8335	0.8019
4. SVC oversamp(50%) 1240feats(50.0%)	0.8658	0.8601	0.826
5. SGD top1240feats (50.0%) byCHI	0.855	0.8621	0.821
6. RF top1859feats (75.0%) byCHI	0.766	0.7661	0.8204

Ilustración 48: Evaluación de los mejores modelos obtenidos con submuestreo o sin submuestreo

Lo primero que se observa es que, como se podía intuir por los resultados de las validaciones cruzadas, **ningún modelo alcanza el AUC de 0.9** que se ha marcado como objetivo de negocio.

Además, **el mejor modelo final (modelo 2) no coincide con el mejor modelo de la validación cruzada repetida,** probablemente como consecuencia de no disponer de más observaciones en el conjunto de datos de entrenamiento. De ello se deduce que **las validaciones cruzadas repetidas están o ligeramente sobreajustadas o infraajustadas.**

Los dos mejores modelos se consiguen por medio de NaiveBayesNLTK, que sorprendentemente da muy buenos resultados, a pesar de ser el algoritmo más simple.

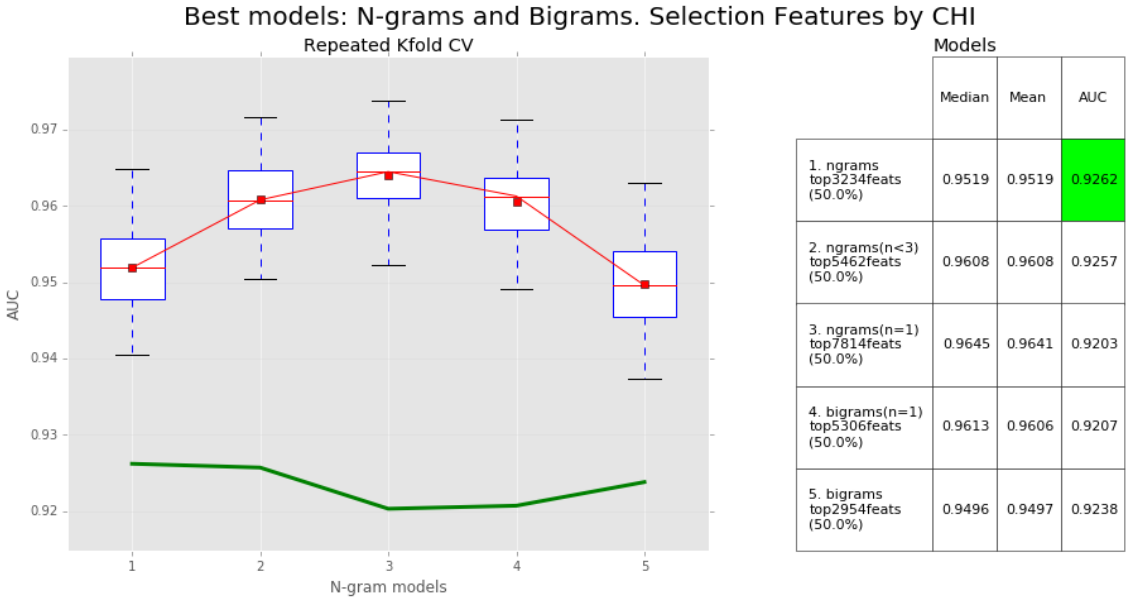
Respecto a la técnica de submuestreo, ésta añade más varianza pero también mejora el sesgo en ambos casos, en especial para los modelos SVC. En cualquier caso, parece **una técnica que requiere tener mucho cuidado, para evitar sobreajuste o infraajuste en la validación cruzada repetida, y la ganancia no siempre parece asegurada debido a su varianza.**

4.2.5.2. Clasificador de Polaridad

Para este clasificador, **los mejores modelos obtenidos se consiguieron eliminando las palabras comunes (SW), reduciendo los términos a su raíz (ST) y seleccionando las mejores características con el test de la χ2 (todo ello con NaiveBayesNLTK).**

En particular, el **mejor modelo** se conseguía claramente cuando se extraían las características **por medio de unigramas, bigramas y trigramas** y se aplicaban **SW y ST a los unigramas únicamente.**

Por tanto, se procedió a evaluar dichos modelos con los datos de evaluación (*test-set*) para corroborar los resultados de las validaciones cruzadas y confirmar que el AUC superaba el 0.9 que marcaban los objetivos del negocio, como se infería de la validación cruzada repetida (véase la Ilustración 49).



Models

	Median	Mean	AUC
1. ngrams top3234feats (50.0%)	0.9519	0.9519	0.9262
2. ngrams(n<3) top5462feats (50.0%)	0.9608	0.9608	0.9257
3. ngrams(n=1) top7814feats (50.0%)	0.9645	0.9641	0.9203
4. bigrams(n=1) top5306feats (50.0%)	0.9613	0.9606	0.9207
5. bigrams top2954feats (50.0%)	0.9496	0.9497	0.9238

Ilustración 49: Evaluación de los mejores modelos para el Clasificador de Polaridad

En primer lugar, se observó que **el objetivo del negocio marcado se había cumplido,** ya que el mejor **AUC es 0.9262**; y, aunque no coincide el mejor modelo de la validación cruzada con el

mejor modelo para los datos de evaluación, la diferencia es tan escasa que probablemente sea un problema menor de ruido.

Por otra parte, puede apreciarse que el resto de **los mejores modelos obtenidos según la validación cruzada repetida también superan el objetivo del 0.9**. Por tanto, dada la naturaleza del problema planteado (análisis de sentimientos), se considera que **los modelos alcanzados son muy robustos**.

4.2.6. Despliegue

Durante la fase de despliegue se aborda el plan para la puesta en producción del modelo definitivo. En nuestro caso, la puesta en marcha consistirá en integrarlo en el portal web ***sanidadysalud.com.*** Además, se determinará la estrategia que había que seguir para el mantenimiento y supervisión del modelo una vez implantado.

Así, finalmente, para poner Senti-SyS en producción es necesario **incorporar dos subsistemas adicionales**: la **conexión con la BBDD** para extraer y guardar los datos; y la **optimización de los puntos de corte**, para convertir las probabilidades del modelo en predicciones. La arquitectura definitiva del sistema resultante se muestra en la Ilustración 50.

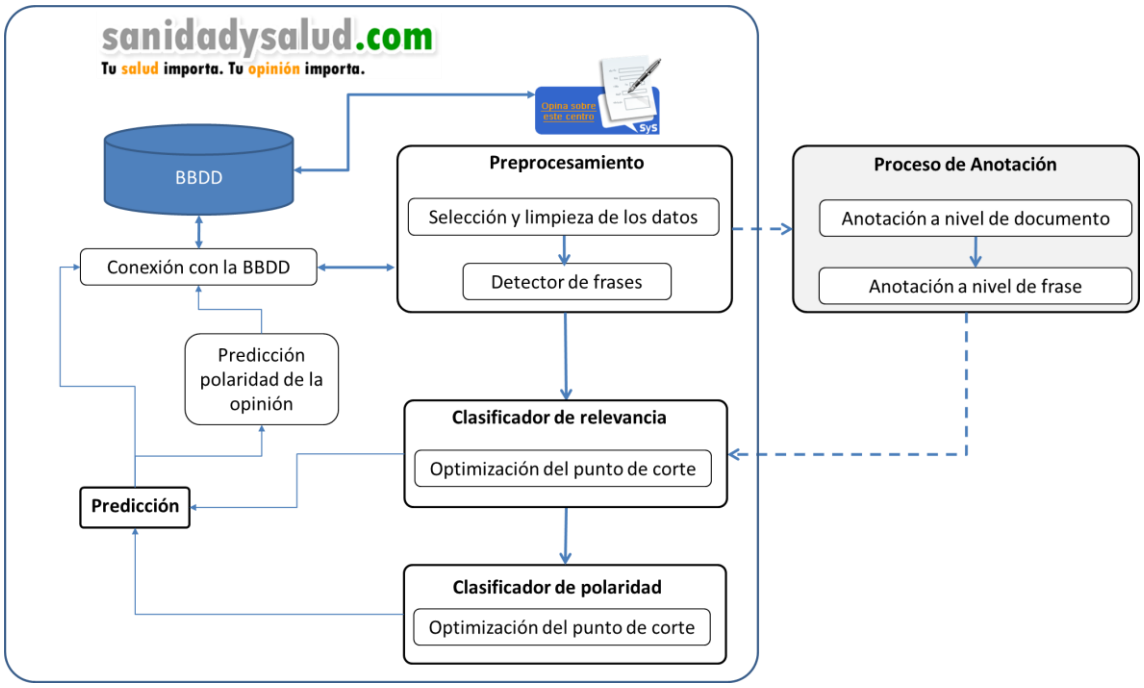
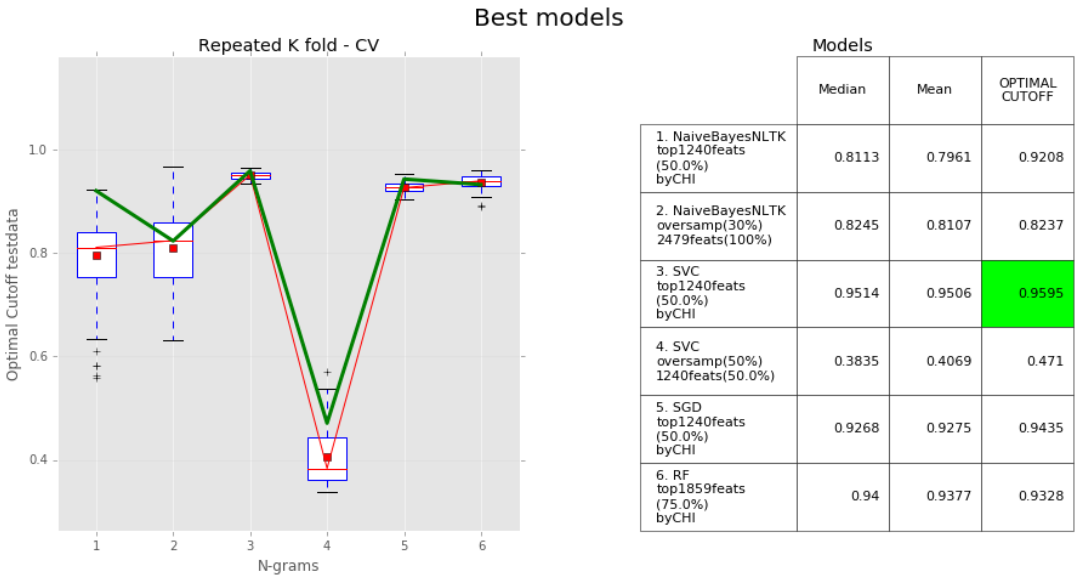


Ilustración 50: Ampliación de la arquitectura del sistema Senti-SyS para el despliegue

El **punto de corte de probabilidad para cada clasificador**, relevancia y polaridad, quedará definido a partir de **la comparación entre los puntos de corte óptimos conseguidos por medio de la validación cruzada repetida y los puntos de corte óptimos para las predicciones del conjunto de evaluación.**



Models

	Median	Mean	OPTIMAL CUTOFF
1. NaiveBayesNLTK top1240feats (50.0%) byCHI	0.8113	0.7961	0.9208
2. NaiveBayesNLTK oversamp(30%) 2479feats(100%)	0.8245	0.8107	0.8237
3. SVC top1240feats (50.0%) byCHI	0.9514	0.9506	0.9595
4. SVC oversamp(50%) 1240feats(50.0%)	0.3835	0.4069	0.471
5. SGD top1240feats (50.0%) byCHI	0.9268	0.9275	0.9435
6. RF top1859feats (75.0%) byCHI	0.94	0.9377	0.9328

Ilustración 51: Punto de corte óptimo para los datos de evaluación para los modelos del clasificador de relevancia

Una vez **definidos los puntos de corte** para cada clasificador, ya **se pueden transformar las probabilidades conseguidas con los modelos a predicciones de clase**. Dichas predicciones se guardarán en la BBDD para cada opinión.

Por último, quedaría agregar la nueva lógica al controlador que gestiona en **sanidadysalud.com** la publicación las de las opiniones actualmente. Como quedó definido en el propósito del presente trabajo, dicha lógica permitirá la **publicación automática de las opiniones clasificadas como relevantes y no negativas,** y **propondrá la revisión manual de las clasificadas como relevantes y negativas.**

Respecto al mantenimiento del sistema clasificador, se propone un **reentrenamiento cada 6 meses**. Para ello, se repetirán los distintos procesos del presente trabajo, actualizando los conjuntos de datos de entrenamiento y evaluación.

Se propone dejar los **últimos seis meses** anteriores a la fecha de reentrenamiento **como conjunto de datos de evaluación**, utilizándose el resto para el reentrenamiento de los modelos.

5. Conclusiones y trabajos futuros

En esta memoria se ha presentado un trabajo cuya finalidad es **ofrecer una solución a un problema del mundo real desde un punto de vista propio de minería de datos**, aprovechando además el conocimiento experto del dominio del problema disponible.

Dicho problema es un caso particular de análisis de sentimientos, y entra dentro de la categoría de la Clasificación de Textos. Como se ha podido observar en la literatura científica, dicho problema es aún un **problema abierto dentro del área de investigación del Procesamiento del Lenguaje Natural**.

En particular, se trata de identificar las opiniones negativas de usuarios del portal de opiniones sanitarias **sanidadysalud.com** para su revisión manual; mientras el resto, las opiniones no negativas, se publicarían automáticamente sin intervención humana. Esto permitirá ahorrar recursos destinados al mantenimiento del portal.

Para ello, como se ha visto anteriormente, **el problema principal puede descomponerse en dos subproblemas distintos** para su solución: (i) el análisis de la polaridad de la opinión (negativa o no negativa); y (ii) la detección de las opiniones irrelevantes (*spam*).

De esta manera, la solución que se ofrece en el presente proyecto hace frente a dicha situación y **propone una clasificación en dos niveles**:

- **clasificación de la relevancia** del texto, para determinar si una opinión es irrelevante o no.
- **clasificación de la polaridad** del texto, para determinar si la opinión es negativa o no.

En particular, el problema de la clasificación de la relevancia ha requerido sortear el obstáculo que suponía el gran desequilibrio de las clases objetivo del conjunto de datos, lo cual es típico de la clasificación de opiniones irrelevantes (*spam)*. Para ello, **se ha propuesto una solución basada en remuestreo y ensamblaje**. Sus resultados han sido esperanzadores, pero se necesita refinar más la solución intentando reducir la varianza de los modelos conseguidos.

También **se han comparado diversos métodos de aprendizaje automático** supervisado sobre el conjunto de datos disponible. Sobre todos ellos se ha aplicado **la validación cruzada repetida,** para comparar dichos métodos y garantizar en la medida de lo posible una buena generalización de los modelos conseguidos, a pesar de su coste computacional.

En este sentido, se han comparado dos de los métodos base empleados en la literatura relacionada: *Naive Bayes* y *SVM.* Ambos han arrojado resultados razonablemente buenos, a pesar de ser métodos no demasiado modernos. De hecho de entre todos los métodos empleados, el que mejores resultados ha producido en ambos subproblemas es *Naive Bayes.*

Para crear el espacio de características específico y seleccionar las más importantes para dichos modelos se han empleado diferentes técnicas estadísticas. Además, se ha probado a combinarlas, con la esperanza de que los puntos débiles de cada una se matizaran, aunque sin resultados reseñables.

Por otra parte, para desarrollar la solución informática a este problema **se han empleado los dos lenguajes principales** dentro del mundo de la ciencia y **minería de datos en la actualidad**: ***R* y *Python***.

Asimismo, dentro del proyecto, con el fin de disponer de un ***gold standard*** con el que poder contrastar los resultados, **se ha anotado manualmente la colección de textos completa disponible,** con la ayuda de un experto humano en el dominio.

No obstante, **durante el desarrollo del proyecto se han producido numerosos retrasos**, fundamentalmente en el proceso de anotación, más costoso de lo esperado, y en la detección de frases; y también durante el modelado, por el inconveniente de tener que ir aprendiendo el lenguaje de programación (*Python*) a la vez que se modelaba. En la Ilustración 52 se puede observar la diferencia entre el calendario previsto y el real.

Por último, en cuanto a los **resultados del proyecto,** han sido **razonablemente buenos,** dado que el conjunto de datos no era muy extenso y no había referencias de resultados anteriores más allá de la literatura relacionada, debido a que era un problema real sin resolver.

Ahora bien, en el caso del clasificador de la relevancia, no se ha llegado al objetivo marcado y por ello, quedan muchas ideas por explorar para mejorar sus resultados como **trabajo futuro**.

Para ello, en orden de importancia se propone:

1. **Utilizar más datos para el entrenamiento**, pues el conjunto de datos actual se queda corto para algunas técnicas de aprendizaje.
2. **Incorporar al espacio de características** la **metainformación** recogida junto con los cuestionarios de los usuarios.
3. **Optimizar la selección de las mejores características** empleando o combinando las técnicas estadísticas con técnicas de selección de variables a partir de otros métodos de modelado (*randomForest*, *Gradient Boosting, XtremeGradientBoosting, Stepwise, Lasso*, etc)
4. **Optimizar el ensamblado** de los "mini" clasificadores con optimización numérica (maximizando el AUC) o utilizando técnicas de metaaprendizaje (*stacking, boosting, bagging, etc*)
5. **Probar enfoques más modernos de creación de características** (*word2vec*) **o modelado** (*deep learning*) que actualmente están dando buenos resultados.
6. **Probar enfoques de análisis semánticos y sintácticos en vez de, o además de, un enfoque léxico.**

No obstante, a pesar del espacio de mejora existente, con los resultados obtenidos, **la puesta en producción del sistema desarrollado permitirá ahorrar recursos en el mantenimiento del portal web *sanidadysalud.com***.

Por otra parte, **más allá de la mejora en el poder de predicción**, la evolución del sistema puede ir dirigida a **ampliar el servicio fuera del ámbito de *sanidadysalud.com***, de forma que también se puedan clasificar opiniones sanitarias recogidas en redes sociales, blogs sanitarios o páginas web del ámbito sanitario.

Además, con la tecnología desarrollada en el presente trabajo, **también se podría ampliar el nivel de servicio** para que el análisis profundice hasta los aspectos de los centros sanitarios, de forma que la polaridad de cada frase se construya a partir de las polaridades de los aspectos identificados previamente en ella.

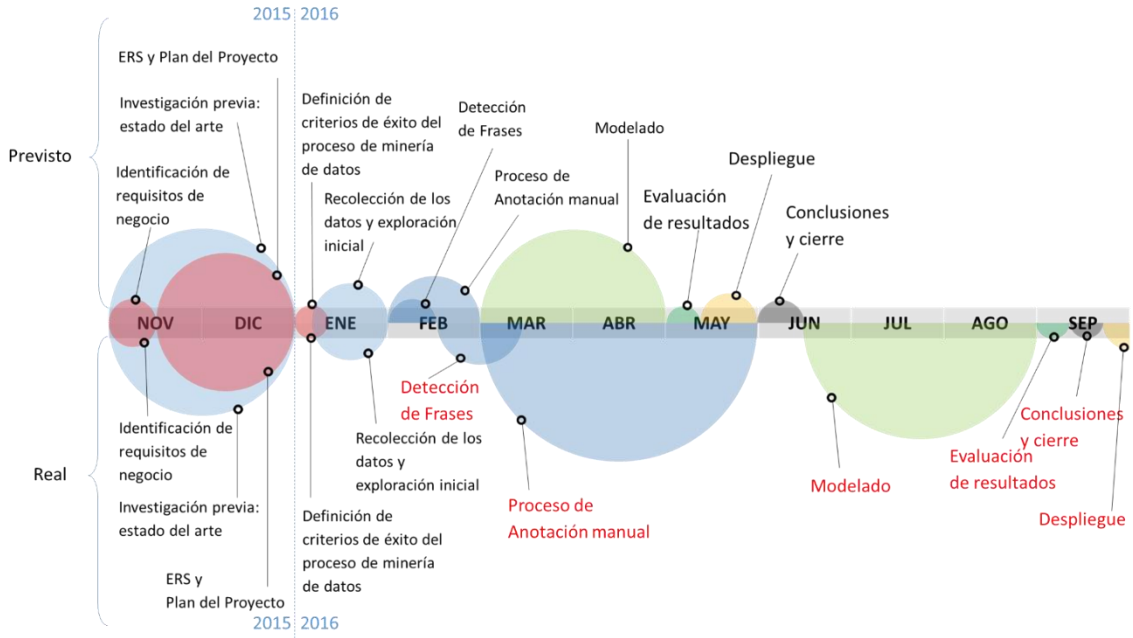


Ilustración 52: Cronograma del proyecto previsto frente al real

6. Referencias

- Real Academia Española. 2014. *Sentimiento.* En Diccionario de la lengua española *(23.a ed.).* Recuperado el 16 de diciembre de 2015 de http://dle.rae.es/?id=XbTu91V
- Bird, Steven, Loper, Edward y Klein, Ewan. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.
- Flach, Peter. 2012. *Machine Learning: The Art and Science of Algorithms that make Sense of Data.* Cambridge
- Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining* [en línea]. Morgan & Claypool Publishers: [Consulta: 02 de enero de 2016]. Texto en PDF. Disponible en: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
- Manning, Christopher D., Raghavan, Prabhakar y Schütze, Hinrich. 2008. *Introduction to Information Retrieval, Cambridge University Press.*
- Pang, Bo, Lee, Lillian, y Vaithyanathan, Shivakumar, 2002. *Thumbs up? Sentiment classification using machine learning techniques*. En*: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, Pennsylvania, USA, 2002, pp. 79-86.
- Piatetsky-Shapiro, Gregory. 2014. *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. En: KDnuggets.com [Sitio Web]. octubre, 2014. [Consulta: 10 Enero 2016]. Disponible en: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html
- *SemEval 2015, 2015. Proceedings of the 9th International Workshop on Semantic Evaluation*. Denver, Colorado, June 4-5, 2015. Disponible en: http://www.aclweb.org/anthology/S15-2045
- Sewell, Martin, 2007. *Ensemble Methods*.
- Sokolova, M., y Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*, p. 427-437.
- Turney, Peter. 2002. *Thumbs Up* or *Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.* En: *Proceeding of ACL 2002*, 2002, pp. 417-424.
- Villena-Román, Julio. 2015. *Introducción al análisis de sentimientos (minería de opiniones).* En: Meaning Cloud Blog. [Blog]. octubre, 2015. [Consulta: 02 Enero 2016]. Disponible en: http://www.meaningcloud.com/es/blog/introduccion-al-analisis-de-sentimientos-mineria-de-opinion/
- Villena-Román, Julio, Janine García-Morera, Miguel A., García Cumbreras, Eugenio Martínez Cámara, M.Teresa Martín Valdivia, and L. Alfonso Ureña López, (eds). 2015. *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*. CEUR WS Vol 1397. Disponible en: http://ceur-ws.org/Vol-1397/
- V. S. Jagtap, Karishma Pawar, 2013. *Analysis of different approaches to Sentence-Level Sentiment Classification*. En: *International Journal of Scientific Engineering and Technology.* April 2013 (ISSN: 2277-1581) Volume 2 Issue 3, PP : 164-170
- Wiebe, J.; Bruce, R.; and O'Hara, T., 1999. *Development and use of a gold standard data set for subjectivity classifications*. En: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (ACL-99), 246–253.

- Yang, Yiming and Jan O. Pedersen, *1997. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the 14th International Conference on Machine Learning (ICML), pages 412-420.*
- *IEEE 830-1998 — IEEE Recommended Practice for Software Requirements Specifications. 1998.* doi:10.1109/IEEESTD.1998.88286. ISBN 0-7381-0332-2.
- IBM Software Group. 2011. *IBM SPSS Modeler CRISP-DM Guide*. [Consulta: 10 Enero 2016]. Disponible en: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

Wikipedia

- *Natural Language Processing* (s.f.). En *Wikipedia*. Recuperado el 16 de diciembre de 2015 de https://en.wikipedia.org/wiki/Natural_language_processing
- *Text Mining* (s.f.). En *Wikipedia*. Recuperado el 16 de diciembre de 2015 de https://en.wikipedia.org/wiki/Text_mining
- *Web Mining* (s.f.). En *Wikipedia*. Recuperado el 16 de diciembre de 2015 de https://en.wikipedia.org/wiki/Web_mining
- *Document Classification* (s.f.). En *Wikipedia*. Recuperado el 16 de diciembre de 2015 de https://en.wikipedia.org/wiki/Document_classification
- *Sentiment Analysis* (s.f.). En *Wikipedia*. Recuperado el 16 de diciembre de 2015 de https://en.wikipedia.org/wiki/Sentiment_analysis
- *Automatic Summarization* (s.f.). En *Wikipedia*. Recuperado el 16 de diciembre de 2015 de https://en.wikipedia.org/wiki/Automatic_summarization
- *Part Of Speech Tagging* (s.f.). En *Wikipedia*. Recuperado el 02 de enero de 2016 de https://en.wikipedia.org/wiki/Part-of-speech_tagging
- *Web Scraping* (s.f.). En *Wikipedia*. Recuperado el 02 de enero de 2016 de https://es.wikipedia.org/wiki/Web_scraping
- *TF-IDF* (s.f.). En *Wikipedia*. Recuperado el 02 de enero de 2016 de https://es.wikipedia.org/wiki/Tf-idf
- *Named Entity Recognition* (s.f.). En *Wikipedia*. Recuperado el 02 de enero de 2016 de https://en.wikipedia.org/wiki/Named-entity_recognition
- *Cross Industry Standard Process for Data Mining* (s.f.). En *Wikipedia*. Recuperado el 02 de enero de 2016 de https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
- *Bag of Words model* (s.f.). En *Wikipedia*. Recuperado el 31 de Julio de 2016 de https://en.wikipedia.org/wiki/Bag-of-words_model

Python packages

- John D. Hunter , 2007. *Matplotlib: A 2D Graphics Environment*, Computing in Science & Engineering, 9, 90-95 (2007),DOI:10.1109/MCSE.2007.55

- Wes McKinney, 2010. *Data Structures for Statistical Computing in Python*, Proceedings of the 9th Python in Science Conference, 51-56 (2010)
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, 2011. *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, 12, 2825-2830 (2011)
- Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux, 2011. *The NumPy Array: A Structure for Efficient Numerical Computation*, Computing in Science & Engineering, 13, 22-30 (2011), DOI:10.1109/MCSE.2011.37

R packages

- Baptiste Auguie, 2015. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.0.0. https://CRAN.R-project.org/package=gridExtra
- Rinker, T. W., 2013. *qdap: Quantitative Discourse Analysis Package*. 2.2.4. University at Buffalo. Buffalo, New York. http://github.com/trinker/qdap
- Markus Gesmann and Diego de Castillo. *Using the Google Visualisation API with R*. The R Journal, 3(2):40-44, December 2011.
- Hadley Wickham, 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Hadley Wickham and Romain Francois, 2015. *dplyr: A Grammar of Data Manipulation*. R package version 0.4.3. https://CRAN.R-project.org/package=dplyr
- Hadley Wickham, 2016. *tidyr: Easily Tidy Data with `spread()` and `gather()` Functions*. R package version 0.5.1. https://CRAN.R-project.org/package=tidyr
- Yihui Xie, 2015. *DT: A Wrapper of the JavaScript Library 'DataTables'*. R package version 0.1. https://CRAN.R-project.org/package=DT

Anexo I: Especificación de requisitos software

Especificación de requisitos de software

Proyecto: Clasificador de textos automático en sanidadysalud.com: Senti-SyS
Revisión 2.0

Ficha del documento

Fecha	Revisión	Autor	Verificado dep. calidad.
noviembre de 2015	1.0	Sergio Rincón	[Firma o sello]
Julio de 2016	2.0	Sergio Rincón	

Documento validado por las partes en fecha: 29 / 7 / 2016

Por el cliente	Por la empresa suministradora
sanidadysalud.com Tu salud importa. Tu opinión importa.	
Fdo. D. Sergio Rincón García	Fdo. D. Sergio Rincón García

Contenido

Introducción

El presente documento presentará de forma organizada los requisitos que son indispensables para desarrollar un sistema automático de clasificación de textos recogidos en el portal web ***sanidadysalud.com***. Dichos textos se corresponden con las opiniones de los usuarios del portal sobre diferentes aspectos de los centros sanitarios españoles: farmacias, centros de atención primaria y hospitales.

Este documento está estructurado en base al estándar IEEE *Recommended Practice for Software Requirements Specification ANSI/IEEE 830 1998*.

Propósito

El propósito general de este documento es definir de manera clara y precisa todas las funcionalidades y restricciones del sistema clasificador que se desea construir. El documento servirá de base al equipo de desarrollo del sistema.

Alcance

Este sistema clasificador será denominado **Senti-SyS** y estará restringido a clasificar las opiniones expresadas en castellano recogidas en el portal web ***sanidadysalud.com*** con el objetivo de identificar las que tienen polaridad negativa.

Para la clasificación de los textos se emplearán **técnicas de minería de datos y aprendizaje automático supervisado** sobre el conjunto de opiniones seleccionado.

Personal involucrado

Nombre	Sergio Rincón García
Rol	Desarrollo del sistema
Categoría profesional	Ingeniero Superior en Informática
Responsabilidades	Gestión, análisis y desarrollo del proyecto
Información de contacto	serincon@ucm.es
Aprobación	Antonio Pareja Lora

Definiciones, acrónimos y abreviaturas

- **Usuario:** Persona que accede al portal web ***sanidadysalud.com***
- **Escala *Likert*[21]:** Escala psicométrica comúnmente utilizada en cuestionarios. Al responder a una pregunta de un cuestionario elaborado con la técnica de Likert, se especifica el nivel de acuerdo o desacuerdo con una declaración (elemento, ítem o reactivo o pregunta).

[21] Fuente Wikipedia: https://es.wikipedia.org/wiki/Escala_Likert

- **Encuesta:** Cuestionario de satisfacción que se ofrece a los usuarios, con una lista de preguntas de valoración en escala *Likert* y un campo libre de texto para que expresen su opinión.
- **Opinión:** En el contexto del presente documento, texto asociado a una encuesta sobre un aspecto concreto de un centro sanitario expresado libremente por un usuario.
- **Polaridad**[22]**:** Rasgo semántico de un elemento gramatical que exige un contexto afirmativo o negativo.
- **Aprendizaje automático**[23]: Rama de la inteligencia artificial cuyo objetivo es crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos.
- **Aprendizaje automático supervisado:** Tipo de aprendizaje automático en el que una parte, o todos, los ejemplos proporcionados están asociados a una respuesta posible, que en el caso particular de la tarea de clasificación se corresponde con una categoría o clase objetivo.
- **Proceso de Anotación**: Proceso por el cual se asigna una etiqueta de las existentes en un inventario predefinido de antemano a cada texto de la colección de textos a analizar.
- **Corpus**: Colección de textos u opiniones para clasificar.
- **Corpus de entranamiento**: Colección de textos etiquetados con la etiqueta real para entrenar al clasificador.
- **Clasificador**[24]**:** Algoritmo utilizado para asignar un elemento entrante no etiquetado en una categoría concreta conocida.
- **Clasificador de textos:** Caso particular de Clasificador, donde el elemento entrante es un texto. En el contexto del presente documento, los textos serán las opiniones de los usuarios y las categorías serán “negativo” y “no negativo”.
- **Sentimiento**[25]**:** Estado afectivo del ánimo.
- **IT:** Acrónimo inglés para referirse a Tecnología de la Información o Tecnología Informática.
- **Infraestructura IT**: Conjunto de hardware y software sobre la que se asientan los diferentes servicios tecnológicos.
- **ERS**: El presente documento: Documento de Especificación de Requisitos Software.
- **BBDD**: Acrónimo para referirse a Base de Datos
- **Característica:** Propiedad medible de un fenómeno observado. Se corresponde con una variable independiente Y exploratoria de una variable dependiente X
- **Cobertura:** En el caso de un problema de clasificación, es una medida que se corresponde con la tasa de verdaderos positivos. Es decir, el porcentaje de positivos capturados.
- **Precisión:** En el caso de un problema de clasificación, es una medida que se corresponde con el porcentaje que representan los verdaderos positivos sobre el total.

[22] Fuente RAE: http://dle.rae.es/?id=TVFFNyI
[23] Fuente Wikipedia: https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico
[24] Fuente Wikipedia: https://es.wikipedia.org/wiki/Clasificador_(matem%C3%A1ticas)
[25] Fuente RAE: http://dle.rae.es/?id=XbTu91V

- **Medida-F:** En el caso de un problema de clasificación, es una medida que se corresponde con la media harmónica de la cobertura y la precisión.
- **Punto de corte:** Es un umbral de probabilidad que se establece a partir del cual, todas las predicciones mayores que son clasificadas como positivas y las menores como negativas.
- **Punto de corte óptimo:** Es el punto de corte que minimiza la tasa de falsos positivos mientras maximiza la tasa de verdaderos positivos.
- **Espacio ROC:** Es un espacio definido por la tasa de verdaderos positivos y la tasa de falsos positivos.
- **Curva ROC:** Es una curva construida en el espacio ROC a partir de diferentes puntos de corte para el conjunto de predicciones disponibles. Con cada punto de corte se define un punto en el espacio ROC.
- **AUC:** Es una medida que representa el área bajo la curva ROC. Se corresponde con la probabilidad de que el clasificador otorgue una puntuación superior a una observación positiva seleccionada aleatoriamente que a una observación negativa seleccionada aleatoriamente.

Referencias

- *830-1998 - IEEE Recommended Practice for Software Requirements Specifications*

Resumen

Este documento consta de tres secciones. La presente sección es la introducción y proporciona una visión general del ERS. En la siguiente sección se da una descripción general del sistema, con el fin de conocer las principales funciones que debe realizar, los datos asociados y los factores, restricciones, supuestos y dependencias que afectan al desarrollo. En la última sección se definen detalladamente los requisitos que debe satisfacer el sistema.

Descripción general

Perspectiva del producto

El sistema clasificador formará parte del portal web ***sanidadysalud.com*** como parte de su módulo de procesamiento de las encuestas de satisfacción.

Dichas encuestas se procesan diariamente a las 00:00h. Por tanto, a esa hora el sistema clasificador leerá de la BBDD las encuestas no procesadas, para después clasificarlas y actualizar nuevamente la BBDD.

Funcionalidad del producto

El sistema permitirá **clasificar las opiniones** expresadas libremente por los usuarios sobre diferentes aspectos de los centros sanitarios españoles a través el portal web sanidadysalud.com, en **negativas** y **no negativas.**

Estas opiniones estarán asociadas a un aspecto de un centro sanitario, a través de un cuestionario con campos codificados en escala Likert y un campo de escritura de texto libre adicional para recoger dichas opiniones.

Por ello, primero, el sistema extraerá la opinión del cuestionario asociado. A continuación, detectará su idioma (si es castellano u otro). Después, detectará si la opinión cumple con el propósito general de sanidadysalud.com, esto es si es apta o relevante para su publicación y la clasificará en "relevante" o "no relevante."

Para cada opinión que haya sido clasificada como "relevante", ésta será dividida en las frases que la constituyen de forma que cada frase será clasificada individualmente para detectar la polaridad. En caso contrario se grabará en la BBDD la opinión como "no relevante".

Una vez se tengan las polaridades de todas las frases de una opinión relevante, ésta será clasificada como "negativa" si contiene al menos una frase clasificada como "negativa", y "no negativa" en caso contrario. A continuación se grabará en la BBDD su clasificación.

Por tanto, el sistema principal se divide funcionalmente en **dos subsistemas clasificadores principales**:

- **Clasificador de relevancia**
- **Clasificador de polaridad**

Características de los usuarios del software

El sistema clasificador se comunicará automáticamente con el portal web sanidadysalud.com sin necesidad de interacción humana, por ello el usuario directo puede considerarse el propio portal web; y los usuarios indirectos serían los administradores del portal sanidadysalud.com que son las personas impactadas por la función y salida del sistema clasificador.

Tipo de usuario	Directo: sanidadysalud.com
Formación	-
Habilidades	-
Actividades	-

Tipo de usuario	Indirecto: Administradores de sanidadysalud.com
Formación	Formación técnica universitaria
Habilidades	Experiencia técnica experta
Actividades	Servicios IT

Restricciones

Restricciones técnicas

El sistema clasificador estará integrado en el portal web sanidadysalud.com; por tanto, hará uso de su infraestructura IT y estará restringido a las limitaciones de ésta en cuanto a sistema operativo, lenguajes de programación, hardware y software específicos.

En particular el sistema clasificador estará desarrollado en Python 3, para facilitar la integración con la infraestructura existente.

Restricciones funcionales

El tamaño máximo del texto de cada opinión será de 500 caracteres, incluyendo espacios en blanco y signos de puntuación.

El idioma de la opinión será el castellano.

Suposiciones y dependencias

Dependencias técnicas

El sistema clasificador estará desarrollado en *Python* 3 utilizando las librerías: *scikit-learn* (aprendizaje automático), *numpy* (estructuras de datos), *pandas* (estructuras de datos), *nltk* (procesamiento de lenguaje natural) y *re* (expresiones regulares).

Dependencias funcionales

El sistema clasificador estará compuesto por uno o varios algoritmos de aprendizaje automático supervisado.

Por ello, para entrenar a los distintos algoritmos se requiere disponer de un listado de opiniones clasificadas previamente de forma manual por una o varias personas, que será utilizado para entrenar a dichos algoritmos.

Este listado requiere una doble anotación para los dos subsistemas que forman el sistema principal:

- Clasificación de la colección de textos a nivel de opinión con las etiquetas: “relevante” y ”no relevante”
- Clasificación de la colección de textos a nivel de frase sólo para las opiniones clasificadas como “relevante” previamente con las etiquetas: “negativa” y “no negativa”

Evolución previsible del sistema

La evolución natural del sistema clasificador es poder ofrecer el servicio fuera del ámbito de sanidadysalud.com, de forma que también se puedan clasificar opiniones sanitarias recogidas en redes sociales, blogs sanitarios o páginas web del ámbito sanitario.

Además, se podría ampliar el servicio del sistema para que se pueda profundizar en el nivel de análisis en cuanto a los diferentes aspectos de los centros sanitarios, de forma que la polaridad de cada frase se construya a partir de las polaridades de los aspectos detectados en ella.

Requisitos específicos

Requisitos comunes de los interfaces

Interfaces de usuario

No se define ninguna interfaz de usuario

Interfaces de hardware

No se define ninguna interfaz de usuario

Interfaces de software

No se define ninguna interfaz de usuario

Interfaces de comunicación

La comunicación con la BBDD se hará a través del puerto por defecto para MySQL (3306) a través del conector Python – MySQL para el sistema operativo de sanidadysalud.com (centOS 6)

Requisitos funcionales

En este apartado se presentan los requisitos funcionales que deberán ser satisfechos por el sistema. Todos los requisitos aquí expuestos son esenciales, es decir, no sería aceptable un sistema que no satisfaga alguno de los requisitos expuestos. Los requisitos se han especificado de manera que sea fácil comprobar si el sistema los ofrece o no y si los ofrece de manera adecuada (criterio de testabilidad)

Número de requisito	**RF01**
Nombre de requisito	Leer encuestas BBDD
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	El sistema clasificador se conectará diariamente en la hora prefijada a la BBDD de sanidadysalud.com para leer las encuestas con estado de no procesadas. La salida será una lista con todas las encuestas para procesar

Número de requisito	**RF02**
Nombre de requisito	Procesar encuesta
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	La entrada será una lista con todas las encuestas a procesar. Cada encuesta de esa lista será procesada individualmente. El procesamiento consiste en extraer la información relevante de cada encuesta: - Identificador de la encuesta - Identificador del centro - Identificador del tipo de centro - Identificador de usuario - Identificador del bloque de preguntas - Fecha de grabación - Texto de opinión - Estado de la encuesta Se devuelve una tabla con la anterior información por cada encuesta recibida en la entrada.

Número de requisito	**RF03**
Nombre de requisito	Procesar opinión
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Recibe una lista de opiniones para ser procesadas. Cada opinión recibida debe ser transformada a un espacio de características. Para ello las acciones previstas son: - Normalizar texto: pasar a minúsculas y quitar caracteres especiales. - Separar palabras - Eliminar palabras comunes - Reducir a la raíz de la palabra - Crear una lista con las raíces, que serán las características para el modelado Se devuelve la tabla de opiniones recibida de entrada con las nuevas columnas añadidas: - Lista de palabras de la opinión - Lista de características extraídas

Número de requisito	**RF04**
Nombre de requisito	Detectar frases
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Recibe una lista de opiniones para ser procesadas. Cada opinión recibida como entrada debe ser separada en sus frases constituyentes. Devuelve una tabla con la siguiente información: - Identificador de la encuesta - Identificador de la frase - Identificador del centro - Identificador del tipo de centro - Identificador de usuario - Identificador del bloque de preguntas - Fecha de grabación - Texto de opinión - Frase de la opinión - Estado de la encuesta - Etiqueta (puede ser nula) Cada identificador de frase se generará a partir del identificador de la encuesta a la que pertenece. Por ejemplo: 1234#1, 1234#2, 1234#3,….

Número de requisito	**RF05**
Nombre de requisito	Procesar frase
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Recibe una lista de frases para ser procesadas. Cada frase recibida debe ser transformada a un espacio de características. Para ello las acciones previstas son: - Normalizar texto: pasar a minúsculas y quitar caracteres especiales. - Separar palabras - Eliminar palabras comunes - Reducir a la raíz de la palabra - Crear bigramas y trigramas - Crear una lista con los unigramas (raíces), bigramas y trigramas extraídos, que serán las características para el modelado. Se devuelve la tabla de opiniones recibida de entrada con las nuevas columnas añadidas: - Lista de palabras de la frase - Lista de características extraídas

Número de requisito	**RF06**
Nombre de requisito	Leer corpus
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	El sistema clasificador leerá desde un fichero con extensión *csv* el corpus de entrenamiento, que consiste en una lista de opiniones con la etiqueta real que le corresponde. Devuelve una tabla con la información leída.

Número de requisito	**RF07**
Nombre de requisito	Entrenar clasificador relevancia
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Recibe de entrada una lista de opiniones procesadas. Para cada opinión el campo Etiqueta no debe ser nulo. El sistema actuará de la siguiente forma: - Extrae el espacio de características total del corpus - Puntúa las características con el test de la χ^2 (CHI) y las ordena de mayor a menor. - Selecciona el primer 50% de ellas. - Crea una validación cruzada repetida con las características seleccionadas. - Utiliza el AUC como métrica de evaluación en la validación

	cruzada. Devuelve los resultados (probabilidades) agregados por iteración de todas las iteraciones de la validación cruzada y el clasificador obtenido. Los resultados por cada iteración incluyen: - Número de iteración - Número de repetición - *AUC* - *Recall* por cada clase objetivo - *Precision* por cada clase objetivo - *F-score* por cada clase objetivo - Punto de corte óptimo - Número de características - Identificador del modelo clasificador

Número de requisito	**RF08**
Nombre de requisito	Entrenar clasificador polaridad
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Recibe de entrada una lista de frases procesadas. Para cada frase el campo Etiqueta no debe ser nulo. El sistema actuará de la siguiente forma: - Extrae el espacio de características total del corpus - Puntúa las características con el test de la χ^2 (CHI) y las ordena de mayor a menor. - Selecciona el primer 50% de ellas. - Crea una validación cruzada repetida con las características seleccionadas. - Utiliza el AUC como métrica de evaluación en la validación cruzada. Devuelve los resultados (probabilidades) agregados por iteración de todas las iteraciones de la validación cruzada y el clasificador obtenido. Los resultados por cada iteración incluyen: - Número de iteración - Número de repetición - *AUC* - *Recall* por cada clase objetivo - *Precision* por cada clase objetivo - *F-score* por cada clase objetivo - Punto de corte óptimo - Número de características - Identificador del modelo clasificador

Número de requisito	**RF09**
Nombre de requisito	Clasificar relevancia
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Recibe de entrada una lista de opiniones procesadas y un clasificador entrenado. El sistema actuará de la siguiente forma: - Extrae el espacio de características total del corpus - Puntúa las características con el test de la χ^2 (CHI) y las ordena de mayor a menor. - Selecciona el primer 50% de ellas. - Crea una predicción de probabilidad para cada opinión de la lista. Devuelve la lista de entrada con la nueva columna: - Probabilidad

Número de requisito	**RF10**
Nombre de requisito	Clasificar polaridad
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Recibe de entrada una lista de frases procesadas y un clasificador entrenado. El sistema actuará de la siguiente forma: - Extrae el espacio de características total del corpus - Puntúa las características con el test de la χ^2 (CHI) y las ordena de mayor a menor. - Selecciona el primer 50% de ellas. - Crea una predicción de probabilidad para cada frase de la lista. Devuelve la lista de entrada con la nueva columna: - Probabilidad

Número de requisito	**RF11**
Nombre de requisito	Predictor
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Determina si las probabilidades obtenidas superan el punto de corte o no, para después asignarles una clase objetivo. Recibe de entrada los resultados de las iteraciones de la validación cruzada repetida, una lista de textos (opiniones o frases) procesadas y clasificadas y una lista de etiquetas para asignar

	Halla el punto de corte definitivo como el promedio de los puntos de corte de cada iteración de la validación cruzada repetida. Para cada probabilidad recibida calcula si supera el punto de corte definitivo o no. Si supera el punto de corte le asigna la etiqueta 2, si no lo supera le asigna la etiqueta 1 Devuelve la tabla de entrada con la nueva columna: - Predicción

Número de requisito	**RF12**
Nombre de requisito	Calcular polaridad de opinión
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Recibe de entrada la tabla de frases procesas, clasificadas y con predicción. El sistema agrupa la tabla por el identificador de encuesta y añade una nueva columna: - Predicción polaridad La columna predicción toma el valor “negativa” si al menos una frase de la opinión tiene predicción “negativa” en cualquier otra caso toma el valor “no negativa”.

Número de requisito	**RF13**
Nombre de requisito	Guardar predicción BBDD
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Recibe de entrada la lista de opiniones procesas, clasificadas y con su predicción distinta de nulo. El sistema clasificador se conectará a la BBDD de sanidadysalud.com para actualizar los registros de las opiniones recibidas, cambiando su estado a “procesada” y actualizando los campos de clasificación a “no relevante”, “negativa” o “no negativa”

Número de requisito	**RF14**
Nombre de requisito	Guardar clasificador
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Recibe de entrada un clasificador entrenado y lo guarda en disco.

Número de requisito	**RF15**
Nombre de requisito	Leer clasificador
Tipo	Requisito
Fuente del requisito	Negocio
Prioridad del requisito	Alta/Esencial
Descripción	Lee un clasificador entrenado desde un fichero. Devuelve el clasificador.

Requisitos no funcionales

Requisitos de rendimiento

El sistema deberá clasificar todas las opiniones no procesadas diariamente en menos de 5 minutos.

Seguridad

La seguridad de la información de forma general estará garantizada por la política general de seguridad del portal web sanidadysalud.com.

La comunicación entre el sistema clasificador y el portal web serán comunicaciones internas que no salen del servidor web.

Fiabilidad

El sistema admite una tolerancia a fallos de 48 horas que se estima como tiempo máximo para que una encuesta relevante sea publicada.

Disponibilidad

Se estima necesario que el sistema debe estar disponible durante la ventana de ejecución establecida el 95% de las veces.

Mantenibilidad

En general, los periodos de mantenimiento serán incluidos dentro de las ventanas de mantenimiento del portal web pero fuera de la ventana de ejecución estimada del sistema clasificador (Diariamente de 00:00h a 00:05h)

Por otra parte, desde el punto de vista funcional el sistema clasificador debe ser reentrenado con nuevos datos agregados cada seis meses.

Portabilidad

El sistema clasificador debe estar diseñado de tal forma que pueda funcionar en una infraestructura IT equivalente a la de sanidadysalud.com pero fuera del entorno de ésta.

Otros requisitos

No se especifican

Apéndices

No se especifican

Anexo II: Guía de Anotación

Introducción

La presente guía de anotación tiene el objetivo de fijar el marco de actuación para el proceso de anotación del corpus formado por las opiniones recogidas en el portal *sanidadysalud.com*

Anotación de comentarios

Para determinar las etiquetas necesarias, se necesita cierta familiaridad con el corpus. Por ello, en primer lugar, se procedió a hacer un muestreo simple de las opiniones recogidas en el mismo. Así vemos que existen **diferentes tipos de opiniones,** siendo las principales las siguientes:

- **Valoraciones totalmente negativas:**
  - "*Funciona como un sanatorio de los años 20, todos juntos y revueltos.....tuve a un familiar ahí y salió horrorizado.....*"
  - *"Muchas consultas carecen de sala de espera(hay que esperar en un pasillo) y en ocasiones hay que hacerlo de pie al no tener ni una silla libre. Olor a cañeria."*
- **Valoraciones totalmente positivas:**
  - "*Quiero felicitar a TODOS los empleados por la atencion que ha recibido mi madre como paciente de S. SOCIAL PUBLICA*."
  - "*Solo decir que desde que entre para que me operara hasta que me dieron el alta, totalmente satisfecho y orgulloso desde el primero hasta el último.Muchas gracias por todo. Habitacion 501."*
- **Valoraciones tanto con referencias positivas como referencias negativas:**
  - "*EL PERSONAL SANITARIO A NIVEL DE AUXILIARES Y ENFERMERIA SON MUY ATENTOS Y SOLICITOS, EL PERSONAL MEDICO DEJA BASTANTE QUE DESEAR SEGUN MI EXPERIENCIA.*"
  - "*En conjunto no está mal. Adolece de calidad alimentaria. Sin tener problemas de aparato digestivo, no llegué a ver ni una sóla pieza de fruta en mis menús durante los doce días que estuve ingresado.La atención posoperatoria fué intachable. Tanto por la parte facultativa como por atención clínica.Reitero, artículos alimentarios muy deficientes."*
- **Mensajes incompletos o con información no relevante:**
  - "*También se tendrían que evaluar a los otros profesionales que están allí trabajando como las enfermeras, administrativos del mostrador, auxiliares y personal de limpieza. "*
  - *"De parte de isabel jackson"*
  - *"MI MEDICO SE LLAMA NATIVIDAD"*
  - *"tiene que estar mas hacia al parqueeasi la gente lo encuentra mas bien"*
  - *"sssssssssssssdfffffffffffffffffffffffffffdccccccccccccccccccccvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv"*
- **Valoraciones que contienen información objetiva:**

- "*Este Hospital atiende a una población de casi 200.000 personas de Ferrol y comarca.*"
- *"A este parking se entra por la calle que está junto al Loewe de Plaza nueva."*
- "*Me llamo Nicolás y me han operado dos veces del aparato digestivo*."
- "*He acudido en varias ocasiones a la ya menciona farmacia bon dia' en sant feliu y me han atendido tres empleados distintos."*

- **Valoraciones en otro idioma distinto al castellano:**
  - "Bon dia, Avui he anat al ginecoleg, on en teoria m'havian de trucar fa quatre mesos, alfinal el meu metge ho va demanar amb urgencia, i avui tenia la visita. He estat una hora esperant i sense ser atesa, gracies a la presa que tenen els ginecolegs per dinar, donat q jo estava alla a la 13.15 i va entrar una noia... als 15min va sortir, 13.30, com l'ho correcte es no picar i molestar, m'espero... 13.35 ... 13.40...no hi ha ningu a dins, em sap greu, pero havia de picar a la porta..."
  - *"Feia temps que no necessitava els serveis del centre. He vist un gran canvi, amb una remodelació bona. Abans el centre era insuficient i començava a estar deteriorat. Ara és lluminós i ampli."*

Con esta información vemos que, efectivamente existen opiniones relevantes que son candidatas a ser publicadas, negativas o no negativas, ya que cumplen con el propósito general de *sanidadysalud.com*, y otras que no lo cumplen y/o contienen errores y no pueden ser publicadas.

Por ello, el inventario de etiquetas que se propone para la primera fase (anotación de comentarios) es:

$$\mathbf{P_1 = \{relevante, no\text{-}relevante, error, otro\text{-}idioma\}}$$

De esta forma, el procedimiento propuesto para el proceso de anotación a nivel de documento sería:

1. Confirmar si la opinión está escrita en castellano. Si no lo está, se le asignará la etiqueta **{otro-idioma}.**
2. Verificar si la opinión está completa y si se entiende su significado. Si no está completa, de tal forma que no se entiende su significado, se le asignará la etiqueta **{error}.**
3. Si el significado de la opinión se entiende y aporta suficiente información para determinar el nivel de servicio del centro sanitario, entonces se ale signará la etiqueta **{relevante}.**
4. En último caso, si la opinión es una exposición de una anécdota, un caso, una experiencia particular, un mensaje directo para el centro o los profesionales, una firma o, en general, cualquier texto del que no se puede extraer significado relevante para determinar el nivel de servicio del centro sanitario, se le asignará etiqueta **{no-relevante}.**

Por otra parte, también cabe destacar que, al estar los datos anonimizados el anotador en ningún momento conoce el centro al que se hace referencia, ni la valoración numérica que le ha dado el usuario del cuestionario junto con su opinión. De esta forma, centramos la tarea del

anotador exclusivamente en interpretar el sentido de la opinión escrita por el usuario desde el punto de vista semántico.

Así, las **opiniones candidatas a ser publicadas o relevantes** serán las que pasarán a la siguiente fase (anotación de frases) para determinar su polaridad.

Etiquetas de relevancia

En el siguiente apartado se presentan diferentes ejemplos, para cada una de las cuatro etiquetas anteriores del conjunto **P_1.**

- **otro-idioma**: las opiniones que estén escritas en otro idioma diferente al castellano quedan fuera del alcance del presente estudio. Por tanto, se marcan con el objetivo de excluirlas posteriormente del corpus definitivo.

  a. *"El doctor Jesús Liesa, es de una professionalitat i humanitat excel%u2022lent."→**{otro-idioma}***
  b. *"No marxis mai.. Et mereixes gaudir molt del teu nou fill." →**{otro-idioma}***
  c. *"I really miss those days when we were both young and inexperienced" →**{otro-idioma}***

- **error:** En algunos casos, es posible encontrar opiniones que estén cortadas de tal forma que no se entienda el significado del mensaje. El objetivo de etiquetar dichas opiniones es excluirlas posteriormente del corpus definitivo.

  a. *"Mi hermana e"→**{error }***
  b. *"sssssssssssssdfffffffffffffffffffffffffdccccccccccccccccccccvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv" →**{error }***
  c. *"Mi novio ingreso en el ala de psiquiatria y me dieron un numero de telefono en el q supuestamente me podian pasar con la habitacion. Yo lo intentaba pero se cortaba. Llame al telefono de informacion y me volvieron a decir lo mismo hasta q ovolvi a llamar y me exicaron q en el ala.de psiquiatria no se pasaban llamadas a los pacientes pero q el horario de visitasera ds cinco a siete y media. Asique yo q vivo en torrevieja me coji un taxi hastael hospital q me costo setenta euros y cuando llegue me" →**{error}***
  d. *"BUENOS DIAS PASO A EXPONER LO OCURRIDO EN MI VISITA A LA CITADA CLINICA: EL PASADO MIERCOLES 21 DE AGOSTO DE 2013, ASISTI A CONSULTA DE NEUROLOGIA CON LA Dª MARMOL. LA CITA LA HABIA SOLICITADO MEDIANTE CORREO ELECTRONICO, ESPECIFICANDO QUE LA CITA LA QUERIA POR LA TARDE. AL POCO RATO DE SOLICITARLA, UNA SEÑORITA ME LLAMA Y ME DICE QUE LAS CITAS SON POR LA MAÑANA; QUE POR LAS TARDES NO HAY CITA PARA NEUROLOGIA. LE COMENTO QUE TRABAJO DE MAÑANA Y QUE PARA ASISTIR POR LA MAÑANA AL MEDICO ME TIENEN QUE D" →**{error}***

- **relevante:** en el contexto del presente trabajo, la etiqueta relevante se asignará a toda opinión que cumpla con el propósito general de sanidadysalud.com, de forma que sea una

opinión que aporte información importante para el resto de usuarios sobre el servicio que se presta en los centros sanitarios.

a. *"las enfermeras son de una cualidad humana impresionante, ojalá todo el resto de enfermería masculina les llegase a su altura"* →***{*** **relevante** ***}***
b. *"hay excepciones, pero en general unos grandes profesionales"* →***{*** **relevante** ***}***
c. *"Me echaron de ginecología de mala manera, y encontré en cirugía a mi ángel de la guarda."* →***{*** **relevante** ***}***
d. *"buena persona pero mal médico"* →***{*** **relevante** ***}***
e. *"Y dicen que los mejores médicos del mundo están en España."* →***{*** **relevante** ***}***
f. *"Lamentable"* →***{*** **relevante** ***}***
g. *"Mal"* →***{*** **relevante** ***}***
h. *"Sin personal."* →***{*** **relevante** ***}***
i. *"Esto no es un banco."* →***{*** **relevante** ***}***
j. *"Creí morir."* →***{*** **relevante** ***}***
k. *"En navidad se transforma en un hipermercado chino, hay más luces de navidad dentro de la farmacia que en portaferrisa."* →***{*** **relevante** ***}***
l. *"Por falta de ayudas, se nota que el Hospital está dejado de servicios y esto repercute en la atención."*→***{*** **relevante** ***}***
m. *"Mi nieto nació hace unos días, el trato la información por parte de las matronas fue inexistente, con oscurantismo, con comentarios como 'no aguantas nada de dolor y quieres tener hijos'; al marido 'por culpa de ella estamos sin comer'. Después de romper aguas estuvo 32 h, al insistir el marido le hicieron la cesarea, entonces la cosa cambio con los otros profesionales al cien por cien. Excepto por lo comentado el resto muy bien."* →***{*** **relevante** ***}***
n. *"GRACIAS."* →***{*** **relevante** ***}***
o. *"doy las gracias al personal de la 3 planta por como han atendido a mi padre y a la auxiliar mari carmen ."* →***{*** **relevante** ***}***
p. *"Muchas gracias por todo."* →***{*** **relevante** ***}***
q. *"Gracias doctor Melón"* →***{*** **relevante** ***}***
r. *"UN BESAZO A TODOS ELLOS."* →***{*** **relevante** ***}***
s. *"SALUD PARA ELLOS"* →***{*** **relevante** ***}***
t. *"AGRADECER AL CIRUJANO DR, FUEYO Y TODO EL EQUIPO DE LA PLANTA 10 , SU INTERÉS POR MI ENFERMEDAD. BESOS A TODOS . CHONI FERNÁNDEZ"*→***{*** **relevante** ***}***
u. *"Fantástico equipo de cirugía cardiovascular"*→***{*** **relevante** ***}***
v. *"El personal anestesista es muy profesional."* →***{*** **relevante** ***}***
w. *"El cirujano aparte de ser um muy buen profesional, es una persona amable y simpática."* →***{*** **relevante** ***}***
x. *"Está cerca de todo."* →***{*** **relevante** ***}***
y. *"Para apreciar lo que se tiene a veces hay que perderlo. No soy de Madrid y cuando la sanidad publica pierda este hospital nos daremos cuenta de la perdida. El personal muy amable y eficiente. Es de agradecer. Ya quisiéramos en el resto de España hospitales así. De pena la educación de las personas, fuman en las instalaciones e incluso dentro del hospital. La mala educación de las personas no se cura en hospitales. la educación y la sanidad no se debieran de tocar, así nos va."* →***{*** **relevante** ***}***

- **no-relevante:** esta etiqueta normalmente será asignada a opiniones que representen anécdotas o experiencias personales y que, por sí mismas, no ofrezcan un mínimo de información para saber cómo es el servicio del centro; o bien serán frases pertenecientes a una opinión cuya información no está relacionada directamente con el centro sanitario, o que se salen del ámbito del portal *sanidadysalud.com,* como por ejemplo los intentos de comunicación directa con los centros o profesionales o el *spam*. Cabe resaltar también que los textos que se reducen a los nombres propios de una firma, o a saludos, en todas sus formas, se considerarán no relevantes, por no aportar información suficientemente relevante para el resto de usuarios.

  a. *"Mi eposa ha presentado una demanda judicial, para lo que le agradeceria fuera tan amable deenviarme un informe los mas completo que pueda soble el desarrollo de mi enfermedad, desde que me atienda."* →***{no-relevante}***
  b. *"ME PARECE MUY MAL EL RUMOR QUE ACABO DE LEER QUE PROBABLEMENTE CIERREN LAS CONSULTAS DE SALU MENTAL... COMO SE NOTA QUE ELLOS NO TIENEN 'ESOS' PROBLEMAS..... QUE SEPAIS QUE NOS DEJAIS COMO.... HUERFANOS..."* →***{no-relevante}***
  c. *"Este centro 'social' es capaz de despedir del centro a un discapacitado del 75%, sin familia que pueda atenderle, con sólo 48 horas de antelación y sin ofrecerle una alternativa. Resumiendo a la calle peor que un perro. Y todo por no haber aceptado que faltara un dia del centro porque tenía una cita a 30 Km de distancia con su asistenta social. En fin enhorabuena por lo humanos, sociales y éticos que son"* →***{no-relevante}***
  d. *"Especialista O. R L. Turno de tarde doctor Ali."* →***{no-relevante}***
  e. *"Respecto al comentario del señor de 43 años, decirle que dudo que eso sea verdad puesto que si el asunto es tan grave como dice debería denunciarlo al Juzgado o donde corresponda y no ponerlo en una web que es libre de leerlo todo el mundo y puede causar perjuicios importantes a la farmacéutica de manera infundada. Un saludo."* →***{no-relevante}***
  f. *"Vicky."* →***{no-relevante}***
  g. *"Saludos."* →***{no-relevante}***
  h. *"UN SALUDO DRA"* →***{no-relevante}***
  i. *"(SI UNO VA AL MÉDICO ES PORQUE NO SE ENCUENTRA BIEN)."* →***{no-relevante}***
  j. *"He sido atendida por urgencias esta misma mañana."* →***{no-relevante}***
  k. *"POR FAVOR TENGAN MIS COMENTARIOS EN CUENTA"* →***{no-relevante}***
  l. *"Bernarda M. S., ocupo la habitación 154."* →***{no-relevante}***
  m. *"urgencias, 7 de agosto de 2012."* →***{no-relevante}***
  n. *"En esta farmacia hay tres empleados."* →***{no-relevante}***
  o. *"La valoración se refiere al Pediatra de mi hija"* →***{no-relevante}***
  p. *"MI MEDICO SE LLAMA NATIVIDAD"* →***{no-relevante}***
  q. *"Esta opinion sé ha hecho sobre el doctor Agustín"* →***{no-relevante}***
  r. *"me gusta una de las chavalitas ke hay ahi es la ke tiene 24 años y ojos claros"* →***{no-relevante}***

Pudiera darse el caso de que el anotador encuentre que, estando el mensaje completo aparentemente, su redacción no permite entender su significado. Estos casos se marcarán también con la etiqueta ***{no-relevante}.***

Anotación de frases

La siguiente tarea de anotación se realiza con la frase como unidad o segmento. Por tanto, la entrada a este proceso requiere una segmentación previa de las opiniones en sus frases constituyentes.

El objetivo de esta fase es etiquetar la polaridad de las frases. Por tanto, al igual que antes, se hizo un muestreo de frases para determinar el inventario de etiquetas. Los principales tipos de frases son:

- **Valoraciones totalmente negativas:**
  - "*HE TENIDO MUY MALAS EXPERIENCIAS Y CUANDO HE IDO A URGENCIAS(MUCHAS VECES)HE TENIDO QUE ESPERAR DEMASIADO TIEMPO…"*
  - *"CADA DIA TE VISITA UN MEDICO DISTINTO, CON LO CUAL LAS EXPLICACIONES HAY QUE INTERPRETARLES POR LAS CARACTERISTICAS DISTINTAS DE CADA PERSONA, EN FIN."*
- **Valoraciones totalmente positivas:**
  - "*Quiero felicitar a TODOS los empleados por la atencion que ha recibido mi madre como paciente de S. SOCIAL PUBLICA*."
  - "*La verdad es que me siento muy orgullosa de nuestro C. médico."*
- **Valoraciones tanto con referencias positivas como con referencias negativas:**
  - "*buena persona pero mal médico*"
  - "*Ahora bien, creo que fundamentalmente depende del médico que tengamos (me consta que todos no son como el mío)... Cuando mi médico me deriva a un especialista fuera del centro, creo que no existe coordinación."*
- **Valoraciones que contienen descripciones de casos, anécdotas o información objetiva:**
  - "*Me llamo Nicolás y me han operado dos veces del aparato digestivo*."
  - "*He acudido en varias ocasiones a la ya menciona farmacia bon dia' en sant feliu y me han atendido tres empleados distintos."*
  - "*Este Hospital atiende a una población de casi 200.000 personas de Ferrol y comarca."*
  - *"A este parking se entra por la calle que está junto al Loewe de Plaza nueva."*
- **Frases incompletas:**
  - ***"Las Dras."***
  - ***"Me han."***
  - ***"para la cual ha sido asig"***
- **Frases con información sin opinión:**
  - "*¡Me he cruzado mensajes con ellos a través de ¡¡¡twitter!!!"*
  - *"De parte de isabel jackson"*
  - *"Hola."*
  - *"Saludos."*

- *"Habitacion 501."*

Con esta información, y teniendo claro que el objetivo principal de este trabajo es detectar las opiniones negativas, se propone el siguiente conjunto de etiquetas:

P_2 = {negativo, no-negativo, error-frase}

Por tanto, se propone el siguiente procedimiento para el proceso de anotación a nivel de frase:

1. Verificar si la frase está completa; si no lo está, de forma que no se puede entender su significado, se le asignará la etiqueta **{error-frase}.**
2. Identificar si contiene alguna referencia negativa, en cuyo caso se le asignará la etiqueta **{negativa}.**
3. Si, en cambio, sólo contiene referencias positivas, objetivas o información sin opinión, entonces se le asignará la etiqueta **{no-negativa}.**

En algunas situaciones, el anotador encontrará que la única forma de esclarecer el sentido de la frase es leerla en su contexto en el comentario original. Por ello, en todo momento, junto a cada frase, el anotador tendrá visible el mensaje original completo. Esto es especialmente importante, en posibles casos de uso de la ironía.

Etiquetas de polaridad

En el siguiente apartado se presentan diferentes ejemplos de uso de cada una de las tres etiquetas anteriores del conjunto **P_2.**

- **error-frase:** es posible encontrar frases que estén cortadas. Esto es debido a que el propio mensaje está cortado o a fallos del algoritmo de detección de frases que se aplica previamente sobre la opinión completa. El objetivo de etiquetar dichas frases es descartarlas, pasada esta fase.

  a. *"que ha habido nunca, se merece mas que 10, me gustaria que tengamos muchos años estas dos profesionales."* →***{error-frase}***
  b. *"A las 5h de haber salido, estaba de vuelta cn ls mismos sintomas, tardaron 7 horas en atenderm llevo 3 dias en el hoital y aun no."* →***{error-frase}***
  c. *"saben lo ke tengo."* →***{error-frase}***
  d. *"Las Dras."* →***{error-frase}***
  e. *"para la cual ha sido asig"* →***{error-frase}***

- **negativa:** Debido al objetivo marcado en este trabajo (identificar las opiniones negativas) cualquier frase que contenga una referencia negativa será etiquetada como negativa, con independencia de si contiene también referencias positivas.

  a. *"las enfermeras son de una cualidad humana impresionante, ojalá todo el resto de enfermería masculina les llegase a su altura"* →***{negativa}***

b. *"hay excepciones, pero en general unos grandes profesionales"* →***{negativa}***
c. *"sin este hándicap, sería un 10 en todo"* →***{negativa}***
d. *"buena persona pero mal médico"* →***{negativa}***
e. *"Y dicen que los mejores médicos del mundo están en España."* →***{negativa}***
f. *"¿Como es posible?."* →***{negativa}***
g. *"Lamentable"* →***{negativa}***
h. *"Mal"* →***{negativa}***
i. *"Sin personal."* →***{negativa}***
j. *"Esto no es un banco."* →***{negativa}***
k. *"Creí morir."* →***{negativa}***
l. *"En navidad se transforma en un hipermercado chino, hay más luces de navidad dentro de la farmacia que en portaferrisa."* →***{negativa}***
m. *"En mi opinión los responsables deben de tomar medidas al respecto."* →***{negativa}***

- **no-negativa:** si la frase no contiene ninguna referencia negativa, esto es, todas sus referencias son positivas, objetivas, descriptivas o no contienen opinión, entonces será etiquetada como no-negativa. Además, los agradecimientos se considerarán no negativos en todas sus variaciones.

a. *"GRACIAS."* →***{no-negativa}***
b. *"doy las gracias al personal de la 3 planta por como han atendido a mi padre y a la auxiliar mari carmen ."* →***{no-negativa}***
c. *"Muchas gracias por todo."* →***{no-negativa}***
d. *"Gracias doctor Melón"* →***{no-negativa}***
e. *"UN BESAZO A TODOS ELLOS."* →***{no-negativa}***
f. *"SALUD PARA ELLOS"* →***{no-negativa}***
g. *"Es alguien que vive por y para su trabajo, tratando a pacientes y familiares como a seres únicos y especiales, aprendièndose sus nombres y no tratando al 26A o al 17D."* →***{no-negativa}***
h. *"En cambio, los exteriores están bien cuidados y son agradables."* →***{no-negativa}***
i. *"Porque bastante tenemos con estar enfermos como para q fueran serios como en otras farmacias!"* →***{no-negativa}***
j. *"El cirujano aparte de ser um muy buen profesional, es una persona amable y simpática."* →***{no-negativa}***
k. *"Viene a casa las veces que la necesitamos y. siempre es muy profesional y cariñosa"* →***{no-negativa}***
l. *"Está cerca de todo."* →***{no-negativa}***
m. *"YO NO VIVO CERCA PERO NO ME IMPORTA,"* →***{no-negativa}***
n. *"EL PROBLEMA NO SON LOS MEDICOS;"* →***{no-negativa}***
o. *":)"*→***{no-negativa}***

Anexo III: Interfaz web de la herramienta de anotación

La interfaz web diseñada presenta al anotador humano un botón por cada etiqueta disponible en el inventario. Una vez pulsado el botón de la etiqueta que desea asociar a cada texto, frase u opinión, dicha asociación se guarda en la BBDD. No obstante, el sistema siempre permite reclasificar cada texto para la corrección de errores o cambios de opinión ante situaciones de ambigüedad.

En cuanto a la información que se le muestra al anotador, es la estrictamente necesaria para hacer su función. Por un lado, el texto a anotar; y por otro, la fecha, texto original, tipo de centro y cuestionario para aclarar el contexto en situaciones de ambigüedad (véase la Ilustración 53).

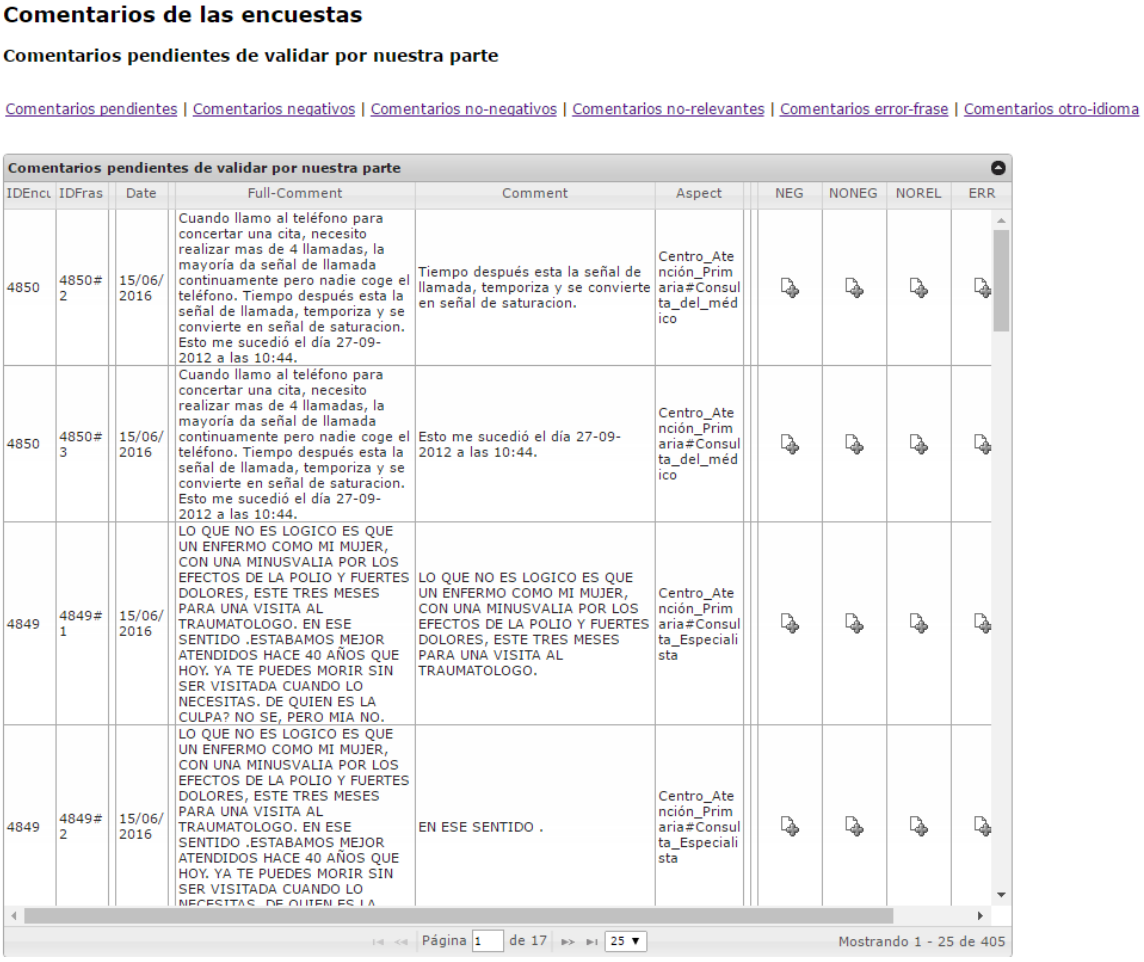


Ilustración 53: Captura de pantalla de la versión 1.0 de la herramienta online desarrollada para gestionar el proceso de anotación