TRABAJO FIN DE MÁSTER

MÁSTER UNIVERSITARIO EN LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN EN LA ENSEÑANZA Y EL TRATAMIENTO DE LENGUAS

SOPORTE BILINGÜE EN EL ANÁLISIS SENTIMENTAL BASADO EN ASPECTOS Y GUIADO POR ONTOLOGÍAS

JAVIER BILBAO VILLAVERDE

TUTOR ACADÉMICO: Olga Borik

FACULTAD DE FILOLOGÍA

CURSO ACADÉMICO: 2020-21- Convocatoria: Septiembre

Agradecimientos

A mi tutora, por su gran dedicación. Su orientación y cuidadosas lecturas han sido la más valiosa ayuda para dar forma a este proyecto.

A mi familia y mi pareja, por apoyarme siempre, en los pequeños y los grandes pasos.

Índice

Índice de abreviaturas y siglas

API: Application programming interface

AS: Análisis sentimental

CRiSOL: Combined Resources in iSOL

EL: Entity linking

et al.: et alii, 'y otros'

ibíd.: ibídem, 'en el mismo lugar'

iSOL: improved Spanish Opinion Lexicon

JSON: JavaScript Object Notation

MCR: Multilingual Central Repository

MO: Minería de opinión

MWT: Multi-word tokens

OSPM: Ontology-supported polarity mining

PLN/NLP: Procesamiento lenguaje natural

POS: Part of Speech

RDF: Resource Description Framework

SOL: Spanish Opinion Lexicon

SWN: SentiWordNet

TIC: Tecnologías de la Información y la Comunicación

WSD: Word sense disambiguation

XML: Extensible Markup Language

Lista de figuras

1. Introducción

1.1 Presentación del proyecto

Este proyecto se desarrolla como Trabajo de Fin de Máster, correspondiente a los estudios de Tecnologías de la Información y la Comunicación en la Enseñanza y el Tratamiento de Lenguas, concretamente en lo relativo a las aplicaciones de las TIC al procesamiento del lenguaje y en el contexto del estudio de la diversidad lingüística y cultural.

El trabajo busca presentar el diseño teórico de un sistema de análisis sentimental con soporte bilingüe, esto es, un sistema encargado de examinar textos y automáticamente determinar la polaridad de la opinión expresada en ellos, ya sea positiva, negativa o neutra. El sistema está diseñado para analizar opiniones pertenecientes al ámbito musical.

En concreto, el planteamiento introducido en este proyecto busca expandir la investigación de la orientación aspectual del análisis sentimental y, en especial, los planteamientos basados en el uso de ontologías. Estas líneas, que analizaremos en detalle, se alinean con los contenidos del máster, al enfocar múltiples tareas de procesamiento del lenguaje natural desde una perspectiva más cercana a la lingüística aplicada, centrándose en el desarrollo de recursos lingüísticos específicos, como ontologías o lexicones de polaridad, más que en la obtención de unos resultados por medios puramente computacionales.

Si bien las investigaciones anteriores en el análisis sentimental basado en aspectos y guiado por ontologías han establecido un modelo a seguir, con una arquitectura de sistema sólida y resultados experimentales positivos, poco se ha profundizado en la posibilidad de introducir un funcionamiento multilingüe en este tipo de sistemas. Por ello, nuestro planteamiento busca integrar en un mismo sistema de análisis sentimental el soporte de dos lenguas distintas, mediante recursos lingüísticos específicos pero equivalentes entre sí, de modo que el funcionamiento del sistema no cambie en función del lenguaje.

Este tipo de funcionalidad viene justificada por la realidad multilingüe e intercultural de nuestro mundo actual y, en concreto, de la comunicación desarrollada en internet y las redes sociales. La línea de investigación en la que se enmarca este trabajo, aplicaciones de las TIC al estudio de la diversidad lingüística y de las variedades de las lenguas, busca profundizar precisamente en estas cuestiones, por lo que se ha hecho énfasis a lo largo del proyecto en la necesidad de prestar atención y reconocer la importancia del tratamiento adecuado de las lenguas integradas en el sistema. Por ello, el soporte bilingüe del sistema es la principal aportación de nuestro estudio en busca de complementar los diseños propuestos en

investigaciones anteriores con nuevas funcionalidades encaminadas a mejorar su funcionamiento en contextos reales de comunicación online.

1.2 Justificación teórica

A la hora de tomar una decisión, especialmente en nuestro rol de consumidores, es considerablemente difícil procesar y valorar correctamente toda la información disponible, por lo que utilizamos diversos procesos heurísticos para facilitar esta tarea (Duhan *et al*., 1997). Algunos de los mecanismos que seguimos se basan en procesos autónomos para simplificar la información, pero otros se ayudan de recomendaciones externas.

De este modo, es común considerar las opiniones ajenas como una fuente importante de información mediante la que orientar nuestro juicio (Pang y Lee, 2008). Antes de la popularización de Internet, el boca a boca era uno de los medios principales por los que recibíamos recomendaciones y que, además, se ha demostrado efectivo a la hora de influenciar las decisiones (East *et al*., 2005). Al mismo tiempo, las recomendaciones de expertos a través de los medios de comunicación tradicionales tenían bastante peso. Sin embargo, la popularización de la comunicación a través de la *web* ha cambiado significativamente la situación. Hoy encontramos en la red una gran cantidad de opiniones vertidas por personas que no pertenecen a nuestro grupo de conocidos ni son críticos profesionales. Estas experiencias anónimas, referidas tanto al consumo de bienes y servicios como a ámbitos tan alejados como la política, son tenidas en cuenta cada vez más (Pang y Lee, 2008). Por ello, también surge un nuevo interés por parte de las compañías sobre las opiniones vertidas en internet a gran escala.

En especial desde la llegada de la Web 2.0, que inicia el cambio hacia un paradigma más colaborativo (O'Reilly, 2009), hasta hoy que las redes sociales son parte de nuestro día a día, las opiniones que como usuarios compartimos de manera libre y pública sobre productos o temas de ámbitos de gran diversidad han crecido a un ritmo desorbitado. La también llamada Web Social acoge una multitud de plataformas dedicadas a intereses muy variados donde se fomenta la participación libre. En plataformas como Amazon[1] los compradores pueden dejar reseñas de los productos que han adquirido; en IMDb[2], de sus películas o series favoritas; o en Discogs[3], de los discos de su colección. Actualmente, encontramos un sitio en Internet

[1] En su version inglesa: http://www.amazon.com
[2] Internet Movie Database: https://www.imdb.com/
[3] Disponible en https://www.discogs.com/

para prácticamente todo ámbito de interés, ya sea en una *web* dedicada como las citadas, a través de foros genéricos como Reddit[4] o blogs de opinión, donde sus creadores comparten con el público sus impresiones. Por su parte, las redes sociales más extendidas, como Facebook[5] o Twitter[6], permiten a sus usuarios compartir todo tipo de contenido, por lo que se genera en ellas a diario una incalculable cantidad de opiniones.

Sin embargo, la amplia variedad de estas opiniones también añade una gran cantidad de ruido adicional que dificulta su valoración. Enfrentados a las opiniones dispares, confusas, incompletas o poco formadas, los usuarios pueden beneficiarse de sistemas que permitan utilizar esa información de manera más eficiente (Pang y Lee, 2008). Por ello, ante la grandísima cantidad de información existente, imposible de revisar manualmente, han surgido propuestas manejar computacionalmente los datos para extraer los aspectos más relevantes de los comentarios, mediante prácticas como el análisis sentimental. Esta disciplina, a medio camino entre la recuperación de información y la lingüística computacional, no se encarga de analizar el tema de una reseña, sino la opinión expresada en ella (Peñalver *et al*., 2014). De este modo, se busca determinar si un texto concreto tiene un carácter positivo, negativo o neutral.

En un contexto de suspicacia por la publicidad y una mayor confianza en las recomendaciones personales (Kennedy, 2012), estas prácticas han ganado un terreno considerable. Así, por un lado, mucha gente puede favorecerse de conocer las opiniones sobre un producto antes de comprarlo, un servicio antes de contratarlo, o incluso sobre cuestiones de carácter más amplio (como la política) antes de tomar una decisión delicada (como elegir un partido al que votar) (Liu, 2012). Por otro lado, el análisis sentimental es del interés de muchas empresas interesadas en recolectar aquello que los usuarios de sus productos dicen sobre ellos en Internet, con la perspectiva de utilizar esas opiniones para entender cómo se percibe su marca, productos y servicios, y orientar de manera consecuente su estrategia empresarial (Pang y Lee, 2008). Tradicionalmente, los métodos empleados por las compañías para adquirir información de este tipo pasaban por realizar encuestas o sondeos de opinión; hoy, gracias a técnicas como el análisis sentimental, se puede explotar el material disponible libremente en la red. Para ello se han intentado diferentes

[4] Donde existe una grandísima variedad de comunidades (llamadas *subreddits*) dedicadas a los temas más diversos. Disponible en https://www.reddit.com/
[5] Disponible en https://www.facebook.com/
[6] Disponible en https://twitter.com/

procedimientos para automatizar este análisis de opiniones y valorar los sentimientos expresados en ellas, como detallamos más abajo.

Puesto que el análisis sentimental parte de las opiniones textuales, se han implementado sistemas dedicados a estudiar sectores tan diversos como el del cine (Zhou y Chaovalit, 2008; Zhao y Li, 2009; o Peñalver *et al*., 2014), las finanzas (Salas Zárate, Valencia García *et al.,* 2017) o la medicina (Salas Zárate, Medina Moreira *et al.,* 2017). De esta manera, teóricamente se podría abarcar cualquier ámbito deseado, con el único requisito de disponer de los recursos lingüísticos necesarios para la tarea.

A pesar de que el ámbito musical no haya sido apenas tratado en el contexto de la minería de opinión, creemos que este tipo de investigación es interesante, debido a la ubicuidad de la música en la sociedad y especialmente en las redes sociales. Un sistema de este tipo podría ser de gran utilidad para agencias musicales o discográficas interesadas en estudiar de manera detallada la opinión del público.

La industria musical transnacional es la principal fuerza en el contexto global de distribución musical; esta, además, se encuentra concentrada en unas pocas manos que controlan un porcentaje muy alto del mercado mundial (Gebesmair, 2017). Sin embargo, esta expansión global no implica una homogeneidad, sino que existen diferentes demandas locales que responden a las culturas de cada región o estado. De este modo, las filiales de las grandes discográficas definen los mercados locales, que en algunos países tienen un carácter más marcadamente local, como en Reino Unido o Brasil (*ibíd.*: 3). Además, estas compañías subsidiarias se enfrentan entre sí para posicionar a sus artistas en posiciones más altas en el panorama mundial. Estas tendencias han ido definiendo el funcionamiento de la industria en las primeras décadas del siglo XXI, como reacción a los cambios desatados por la digitalización y la piratería.

Atendiendo a esta situación, un sistema de análisis sentimental podría aportar información detallada de las opiniones del público sobre los artistas de una discográfica para poder estudiar el desarrollo de sus actividades en el mercado. En una época donde las decisiones vienen determinadas cada vez más por los datos, este tipo de análisis permite tomar en consideración datos sobre aspectos muy subjetivos, que no vienen marcados por las métricas tradicionales de otras plataformas, como el número de reproducciones, seguidores o interacciones. Estos datos cuantitativos no toman en consideración aspectos que sí pueden estudiarse mediante un análisis basado en la subjetividad, como en análisis sentimental.

Así, una canción puede generar mucha tracción a pesar de que no se valore positivamente, por ejemplo, si es muy polémica; de este modo, un artista que genere una gran atención a base de escándalos puede no ser interesante para una compañía. En consecuencia, un estudio detallado de las opiniones puede ayudar a considerar este tipo de cuestiones y tomar decisiones más adecuadas. Decidimos de esta manera aplicar un sistema de análisis sentimental al ámbito musical con el propósito de dar respuesta a este tipo de necesidades.

De manera adicional, nuestro estudio trata de enfocar este tipo de sistemas desde una perspectiva multilingüe, centrada en dos de los idiomas más relevantes en la industria musical, el inglés y el español[7]. Si bien las investigaciones sobre la minería de opinión se han realizado mayoritariamente en inglés hasta el momento, la posición emergente del español como lengua musical global en la actualidad[8] motiva el desarrollo de sistemas que no solo permitan operar en este idioma, sino que empleen recursos específicamente diseñados para ello. Nuestro diseño trata de integrar ambas lenguas de manera equitativa para permitir un análisis plenamente multicultural que responda a la nueva realidad del ámbito.

1.3 Objetivos

El propósito de este trabajo es el desarrollo del diseño teórico de un sistema de minería de opinión bilingüe (inglés y español) con base fundamentalmente lingüística, siguiendo el marco del análisis sentimental con base aspectual y guiado por ontologías, diseñado para analizar comentarios de los usuarios de redes sociales relacionados con el ámbito musical, y por tanto relativos a cantantes grupos, discos, conciertos o festivales.

De este modo, el proyecto tiene dos orientaciones principales diferenciadas: el diseño de un sistema de minería de opinión adaptado al sector de la música y el tratamiento paralelo de dos lenguas dentro de un mismo sistema.

[7] https://www.musicbusinessworldwide.com/english-language-music-is-losing-its-stranglehold-on-global-pop-charts-and-youtube-proves-it/
[8] https://www.rollingstone.com/music/music-latin/latin-pop-urban-reggaeton-trap-755772/

2. Marco teórico

El análisis sentimental (AS) existe en la actualidad como actividad destinada al estudio y el análisis de las opiniones emitidas en el contexto de comunicación *online* generado a partir de la *web* 2.0 y que engloba múltiples plataformas (redes sociales, *blogs*, foros, *webs* colaborativas, etc.). Sin embargo, en un intento de formalizar su labor, podemos seguir la siguiente definición:

> Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. (Liu, 2012: 7)

Esta es un área de investigación compleja y variada, lo que ha ocasionado que surjan términos igualmente diversos para referirse a ella, como "*sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc*." (*ibíd*.: 7). Sin embargo, suelen recogerse bajo los términos *minería de opinión* y *análisis sentimental*, que a su vez suelen usarse de manera intercambiable (Pang y Lee, 2008; Liu, 2012). De modo que aquí se utilizan ambos indistintamente para referirse a esta disciplina.

A pesar de que la minería de opinión (MO) ha generado un gran interés académico y práctico en los últimos años, sus investigaciones son relativamente recientes. Pang y Lee (2008) marcan el año 2001 como punto de inicio a gran escala de la actividad en este ámbito, coincidente con la aparición de una nueva conciencia de las oportunidades prácticas de estas tareas, así como de los problemas existentes en el proceso. Entre los principales factores que facilitaron esta expansión encuentran:

> • the rise of machine learning methods in natural language processing and information retrieval;
> • the availability of datasets for machine learning algorithms to be trained on, due to the blossoming of the World Wide Web and, specifically, the development of review-aggregation web-sites; and, of course
> • realization of the fascinating intellectual challenges and commercial and intelligence applications that the area offers. (Pang y Lee, 2008: 7)

La gran abundancia de opiniones públicas propiciada por la *web* social no puede ser por tanto aprovechada ampliamente de no ser por la implementación de técnicas informáticas específicas que permitan el manejo de estos datos de manera automática.

La recuperación de información se realizó tradicionalmente mediante algoritmos que utilizaban una representación textual de las páginas que analizaban (Cambria *et al.,* 2013).

Sin embargo, estos tan solo eran capaces de capturar los textos, dividirlos en partes o realizar cálculos numéricos sobre ellos. Su capacidad de análisis interpretativo era reducida.

Algunas de las primeras propuestas de minería de opinión partieron de planteamientos sencillos e inmediatos, como el listado y clasificación de palabras por su carácter positivo o negativo y evaluación textual de las opiniones en función de la presencia de estas palabras en ellas. Otros esfuerzos más complejos todavía en el ámbito de la palabra han empleado bases de datos léxicas o cálculos probabilísticos basados en las asociaciones de palabras (Peñalver *et al*., 2014).

Actualmente las técnicas empleadas son de tipología variada y mucho más complejas. Ravi y Ravi (2015) realizan una revisión bibliográfica de las tareas, enfoques y aplicaciones de la minería de opinión desde seis ángulos distintos: "*subjectivity classification, sentiment classification, review usefulness measurement, lexicon creation, opinion word* and *product aspect extraction,* and various applications of *opinion mining*" (*ibíd*.: 8). En este estudio detallado, se describen las técnicas utilizadas en las distintas tareas o etapas en que se puede dividir el AS. Como muestra no exhaustiva de estas, podemos mencionar como ejemplo la clasificación de sentimientos (*ibíd*.: 10-11), en la que se utilizan tanto técnicas de *machine learning* como basadas en léxico. Entre las primeras se distingue entre las supervisadas (como árboles de decisiones, máquinas de vectores de soporte o redes neuronales) y las no supervisadas, según utilicen datos de entrenamiento anotados o no. Entre las segundas, las basadas en diccionarios (lexicones que contienen el valor sentimental de sus términos) o en corpus (grandes recopilaciones textuales que se utilizan para realizar cálculos probabilísticos de la ocurrencia de ciertas palabras en opiniones positivas o negativas).

Los acercamientos a la tarea de MO han ido por tanto de la mano del desarrollo y aplicación de métodos informáticos y estadísticos, así como de recursos de lingüística computacional. Como consecuencia, esta disciplina es eminentemente técnica y puede entenderse en la misma línea que otras aplicaciones de PLN, como una tarea de base lingüística pero cuya resolución pasa por planteamientos informáticos y de inteligencia artificial.

Al considerar la MO, Liu (2012) clasifica los tipos de investigación existentes en función del nivel en que se ponga el foco:

-A nivel de documento se trata de clasificar la opinión expresada globalmente a lo largo del texto según exprese sentimientos positivos o negativos. Este planteamiento parte de la idea

de que cada documento expresa opiniones respecto a una entidad única, por lo que su aplicabilidad queda reducida a textos que cumplan esta condición.

-A nivel oracional se intenta determinar el valor de la opinión en cada oración. Su carácter no es únicamente binario, sino que se toman en consideración las oraciones que incluyan información factual y por tanto impliquen neutralidad. Sin embargo, no solo aquellas oraciones que presenten visiones claramente subjetivas pueden afectar la orientación positiva/negativa, sino que algunas oraciones objetivas pueden igualmente entenderse opiniones, como puede ser el caso de "compré el disco de Justin Bieber y el libreto tenía muchas erratas", donde se expresan hechos que, no obstante, implican que la compra no ha sido del todo satisfactoria (opinión negativa).

-A nivel aspectual o de características se busca suplir las carencias de los niveles anteriores, mediante análisis más detallados, considerando las opiniones en sí en lugar de las estructuras lingüísticas (textos, párrafos, oraciones, *etc*.). Este nivel parte de la idea de que una opinión engloba un *sentimiento* (ya sea positivo o negativo) y un *objeto*, sobre el que se opina; por lo que, sin identificar el objeto, de poco valor es el sentimiento (*ibíd*.: 11). Al mismo tiempo, al poner el foco en el objetivo de la opinión, puede comprenderse el proceso de análisis sentimental de una manera más completa. Al observar la oración "me encantó el disco de Justin Bieber a pesar de que el libreto tuviera erratas" se entiende que la opinión general es positiva, aunque no todo lo que en ella se diga lo sea; ya que se valora positivamente *el disco,* pero negativamente *el libreto*. Como señalan Peñalver *et al.* (2014):

> In fact, classifying opinions at the document or sentence level does not indicate what the user likes and dislikes. A positive document on an object does not mean that the user has positive opinions on all aspects or features of that object. Likewise, it is impossible to ensure that a negative document signifies that the user dislikes everything about the object. In a document (e.g., a product review), the user typically writes about both the positive and negative aspects of the object. (*ibíd*.: 2)

Así podemos observar que, a menudo, los objetos de una opinión se describen por sus entidades o diferentes aspectos, por lo que este nivel de análisis busca estudiar los sentimientos en dichas entidades o aspectos (Liu, 2012:11). El objetivo es, por tanto, la estructuración de la opinión existente en un texto no estructurado, mediante la organización de las entidades y aspectos de ellas a los que se aludan:

> Based on this definition, an entity is represented as a tree or hierarchy. The root of the tree is the name of the entity. Each non-root node is a component or sub-component of the entity. Each link is a part-of relation. Each node is associated with a set of attributes. An opinion can be expressed on any node and any attribute of the node. (Liu y Zhang, 2012: 417)

Ante estos enfoques, Peñalver *et al*. (2014) se decantan por la aproximación basada en características, pero subrayan sus dificultades, debidas tanto a la variabilidad semántica de los textos de opinión como a la gran diversidad de aspectos que pueden definir los productos y las palabras utilizadas para opinar sobre ellas. La resolución de este tipo de problemas requiere técnicas de procesado de lenguaje natural más avanzadas, pero produce resultados mucho más precisos (Liu y Zhang, 2012). Por esto, han surgido nuevos métodos empleando corpus especializados y tecnologías de la *web* semántica para afrontar el análisis sentimental a nivel aspectual (Cambria *et al*., 2013).

La *web* semántica es un intento de dar una estructura formalizada, y por tanto manejable de manera automática por máquinas, a los datos e información existentes en la *web*. Para lograrlo, el reto es disponer de un lenguaje que permita describir tanto los datos como una serie de reglas de razonamiento sobre ellos y que además permita que las reglas de una representación del conocimiento existente se exporten y utilicen en la *web* (Berners-Lee, Hendler y Lassila, 2001: 38). Dos de las tecnologías desarrolladas con este objetivo son el lenguaje de marcado XML (que permite anotar y estructurar los documentos) y la especificación RDF (que establece formas de expresar el significado). Sin embargo, la tecnología más relevante a la hora de asistir el análisis sentimental es la ontología, un tipo de archivo que conceptualiza y define de manera formalizada las relaciones entre términos de un campo de conocimiento. En la *web*, la forma más típica de ontología tiene una taxonomía y una serie de reglas de inferencia (*ibíd.*: 40). La ontología, entendida como una colección de información, establece definiciones de las clases de objetos y sus relaciones mutuas: "Classes, subclasses and relations among entities are a very powerful tool for Web use. We can express a large number of relations among entities by assigning properties to classes and allowing subclasses to inherit such properties" (*ibíd.*: 40). Por último, la existencia de reglas de inferencia o razonamiento permite aportar mayor poder operacional a los ordenadores, que, si bien van a seguir sin comprender plenamente la información que manipulen, la van a poder manejar de un modo mucho más efectivo y en maneras que resulten de mayor utilidad y sentido para el usuario humano en el desarrollo de sus actividades (*ibíd.*: 40). Es por esta razón que:

> In this context, we believe that the already mature Semantic Web technology may be a valuable addition to traditional opinion mining approaches. More concretely, ontologies constitute the standard knowledge representation mechanism for the Semantic Web and can be used to structure information. The formal semantics underlying ontology languages enables the automatic processing of the information in ontologies and allows the use of semantic reasoners to infer new knowledge. (Peñalver *et al*., 2014: 3)

Esta tendencia desde planteamientos tradicionales con base en la palabra hacia un análisis sentimental a nivel aspectual semánticamente más sólido y centrado en los conceptos ("semantically rich concept-centric aspect-level sentiment analysis" según Schouten y Frasincar, 2015: 17) permite dar una respuesta a cuestiones lingüísticas sin resolver hasta ahora, como las implicaciones o inferencias (que requieren cierta capacidad de razonamiento), puesto que los enfoques semánticos integran de manera natural información de sentido común y conocimiento tanto del mundo general como de un dominio (*ibíd*.: 17). De este modo:

> Combining concept-centric approaches with the power of machine learning will give rise to algorithms that are able to reason with language and concepts at a whole new level. This will allow future applications to deal with complex language structures and to leverage the available human-created knowledge bases. Additionally, this will enable many application domains to benefit from the knowledge obtained from aspect-level sentiment analysis. (*ibíd*.: 17)

Las técnicas basadas en ontologías permiten procesar la información con mayor finura, especialmente al poder manipular automáticamente el conocimiento referido a un dominio concreto en vez de contemplar el lenguaje como una sola entidad general:

> When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names are meant to denote, and formal axioms that constrain the interpretation and well-formed use of these terms. (Gruber, 1993: 2)

Por tanto, las ontologías pasan a utilizarse como recurso clave gracias a que proporcionan una representación del conocimiento formal y estructurada en relación con un dominio, estableciendo formalmente un lenguaje común para el mismo que involucra las diferentes entidades, atributos y sus relaciones mutuas, algo que se adapta a las necesidades del proceso de AS aspectual descrito arriba. Esto permite su manejo por parte de máquina y ser humano además de funcionar como "sound semantic ground of machine-understandable description of digital content" (Zhou y Chaovalit, 2008: 101). En esta línea, Peñalver *et al*. (2014) apuestan por el uso de técnicas de la *web* semántica para la extracción e identificación de características de una ontología de dominio que, según Ravi y Ravi (2015) han presentado resultados más precisos que las que no utilizan ontologías (dentro de las investigaciones a nivel aspectual).

Del mismo modo, en el presente proyecto, nos centraremos en los enfoques basados en ontologías, por ser esta una sólida línea de trabajo, en alza en los últimos años y que ha presentado resultados bastante satisfactorios, como muestran Zhou y Chaovalit (2008) y Zhao y Li (2009). En la misma línea, seguimos aquí los desarrollos producidos en los trabajos de Peñalver Martínez *et al*. (2011 y 2014) y Salas Zárate, Valencia García *et al*. (2017), que introducen la valoración de la polaridad basada en el análisis vectorial, una técnica concisa y novedosa con gran precisión en los experimentos realizados.

El planteamiento del AS basado en ontologías se introdujo por primera vez en Zhou y Chaovalit (2008), presentado como *ontology-supported polarity mining* (OSPM), partiendo de la tesis de que las ontologías pueden potenciar el proceso de MO aportando información específica del dominio con mayor granularidad semántica.

Para ello, se llevó a cabo un experimento en el que se seleccionó un corpus de opiniones de películas extraídas de la plataforma online IMDb[9] (180 reseñas en total, agrupadas como positivas o negativas) y se aplicaron dos enfoques representativos de las técnicas supervisadas y no supervisadas empleadas hasta entonces en la tarea de minería de polaridad, implementando a partir de ellos su propuesta OSPM.

Su formulación de la predicción de polaridad en términos vectoriales ilustra claramente la orientación de AS aspectual:

> Suppose $t \in T$ is a text about Domain d and T is a collection of texts. In addition, c is a concept in an ontology of d and is described with n properties: $p_1, p_2, \ldots, p_n$, about which sentiment orientations could be expressed in t. According to the number of related properties of c, t can be decomposed into m ($m \leq n$) segments.
> As a result, the predicted polarity of t is not a single value but rather a vector [p_i, v_i, i = 1 . . . n], consisting of polarity value $v_i \in$[-1, 1] of property p_i of c. The values of -1, 0, and 1 indicate extremely negative, neutral, and extremely positive, respectively. (*ibíd*.: 102)

Otros estudios posteriores también tomarán interpretaciones vectoriales del problema como parte de su enfoque metodológico.

La arquitectura de su sistema también define el esquema básico propuesto en varios de los estudios posteriores de este tipo: partiendo de la recolección de textos en bruto (corpus de opiniones), se realiza un pre-procesado y una segmentación del texto (tareas de procesamiento de lenguaje natural que permiten su manipulación automática en las fases

[9] Base de datos en línea que recopila información y opiniones relacionadas con el mundo del cine (películas, series, actores…). Disponible en: https://www.imdb.com/

posteriores), para a continuación proceder a vincular las partes del texto con las propiedades de los conceptos de la ontología, lo que permite generar valores de polaridad para cada segmento y después, para el texto en su conjunto. De esta manera, la ontología de dominio se considera el eje del sistema pues participa en la selección de los aspectos a partir de los que se realiza el análisis de polaridad. Para el desarrollo de la ontología emplean un enfoque híbrido:

> Ontology development can follow a bottom-up, a top-down, or a hybrid approach. The top-down approach starts with the high-level ontological concepts, which is then gradually expands into a full-fledged ontology. The bottom-up approach starts with textual documents and extracts ontological knowledge from the documents. A hybrid approach simultaneously derives knowledge from the top-level ontology and extracts low-level ontologies from documents, and then creates mappings between the different levels of ontologies. The hybrid approach has become increasingly popular in recent years because it is able to not only take advantage of existing domain resources (e.g., taxonomies) and heuristic knowledge but also discover new information from real documents. (*ibíd*.: 102)

A continuación, para realizar el análisis de propiedad en sí, "the polarity of a text is calculated on the basis of feature weights, and the value of each feature's polarity is estimated using the 'maximum likelihood estimate' approach" (Peñalver *et al.,* 2014: 4). De este modo, a cada característica vinculada gracias a la ontología, se le asigna un peso que influenciará el valor de polaridad final.

Por último, se evaluó la precisión de los desarrollos realizados. De esta manera, se demostró que el enfoque era adecuado en su ámbito de aplicación (AS de reseñas de películas), mejorando los ratios de acierto presentados en estudios previos, al mismo tiempo que expandiendo el entendimiento sobre este ámbito al mostrar que algunas de las propiedades de un dominio tienen un mayor peso e importancia que otras en la determinación de la opinión general de un usuario (Zhou y Chaovalit, 2008: 99), lo que hace ver que la integración de una ontología a un sistema de AS es una nueva oportunidad para descubrir nuevos métodos heurísticos para la minería de opinión (*ibíd*.: 105).

Posteriormente, Zhao y Li (2009) continuaron con la investigación de MO de reseñas de películas. Su estudio presenta un marco más desarrollado sobre el que se construirán otros sistemas de base ontológica. En él se integran una serie de componentes principales que sintetizan las etapas anotadas en Zhou y Chaovalit (2008): pre-procesamiento (segmentación de palabras y anotado de partes del discurso, asistido por ontología, ya que algunos conceptos del dominio no aparecen en diccionarios generales), identificación de características (donde se integra plenamente la ontología), identificación de polaridad (mediante un lexicón de términos anotados según los sentimientos que expresen,

SentiWordNet en este caso) y análisis sentimental (donde se obtiene la polaridad general de la opinión).

Además, también se define claramente el papel de la ontología:

> The domain ontology is to describe the concepts of special area, including concepts in special area, attributes of the concepts, relationship between concepts, and constraints among the relationships. The target of constructing domain ontology is to define common terminologies in the area, and give the definition of the relationships among the terminologies. (Zhao y Li, 2009: 207)

De este modo, en el ámbito de las películas el foco son la película y sus diversas características, que se consideran a la hora de desarrollar la ontología partiendo del análisis textual, del que se extraen concepciones de manera iterativa para expandir la ontología.

La aplicación de una ontología de este tipo es importante pues permite seleccionar de entre las opiniones analizadas aquellas oraciones que se relacionen con los términos relativos al dominio estructurado, utilizando las características concretas encontradas para la asignación de polaridad de la opinión.

Otros dos aspectos fundamentales serán también de relevancia en estudios posteriores: por un lado, la utilización de SentiWordNet (Baccianella, Esuli y Sebastiani, 2010) como base de datos léxica de la que se extraen las polaridades (en la forma de vectores, que dan valores numéricos a la polaridad positiva o negativa de una palabra); y, por otro, la utilización de 'reglas lingüísticas' para la aplicación de las polaridades en su cálculo final. De esta manera, se implementan técnicas como, por ejemplo, la consideración de las expresiones de negación como factores que invierten el valor de polaridad expresado, o la asignación de valores automáticos a palabras de polaridad desconocida en aquellos casos en que vayan unidas a otras mediante palabras conjuntivas.

Este estudio asienta la estrategia de MO basada en ontologías y abre las puertas para investigaciones posteriores, en especial en lo relativo al desarrollo de ontologías más precisas y de reglas lingüísticas como las mencionadas para facilitar el tratamiento de oraciones más complejas (Zhao y Li, 2009: 213).

Por su parte, Peñalver *et al.* (2011) supone un sólido paso adelante en las investigaciones basadas en ontologías, puesto que parte de las investigaciones de Zhou y Chaovalit (2008) y Zhao y Li (2009), manteniendo el ámbito de estudio de reseñas de películas, pero aporta un nuevo método de cálculo de análisis sentimental mediante análisis vectorial. En cuanto a su

estructura, comparte los cuatro módulos presentados por Zhao y Li (2009), al mismo tiempo que mantiene SentiWordNet para la obtención de valores de polaridad.

Entre las novedades introducidas, destaca la introducción de un método que asigna importancia variable a las características encontradas en función del número de veces que se mencionen y su posición en el texto, lo que supone una mejora respecto a los trabajos anteriores, que valoraban todas ellas de igual manera. Este nuevo procedimiento permite darles un mayor peso a las características más mencionadas y a aquellas que aparezcan al principio o, especialmente, al final de un texto, pues son las secciones donde generalmente se incluyen los aspectos más relevantes o que sintetizan la opinión.

Sin embargo, su mayor aportación consiste en el método vectorial en la minería de opinión, pues permite fundamentar esta mediante una técnica objetiva, efectiva y fácilmente aplicable. Dado que las características reciben un peso numérico y las polaridades se expresan mediante un vector (que marca los valores de positividad, negatividad o neutralidad de cada término, extraídos de SentiWordNet), se emplea un sencillo cálculo para obtener un vector posición como suma de los vectores de posición ya tasados de cada característica. Este método se expandirá en Peñalver *et al*. (2014).

Los resultados de este estudio muestran que la precisión de este sistema es mayor, en contraste con Zhou y Chaovalit (2008) y Zhao y Li (2009). Además de utilizar una ontología que "contiene no solo conceptos básicos, sino también conceptos específicos, atributos, relaciones e instancias" (Peñalver *et al.,* 2011: 97); se valoran positivamente sus innovaciones metodológicas:

> Después de una reflexión, encontramos que los factores clave para la obtención de estos resultados han sido dos: (i) método de asignación de pesos a características en función de su posición y frecuencia en el texto, y (ii) método de análisis vectorial para la minería de opiniones. (*ibíd*.: 97)

Las aportaciones de Peñalver *et al*. (2014) pueden entenderse como una maduración de Peñalver *et al.* (2011). Se caracteriza por ser un sistema de funcionamiento completamente automático y fácil adaptación a otros dominios e incluso lenguas, mediante el cambio de la ontología y el corpus textual. Esta orientación busca aportar versatilidad a las propuestas previas basadas en ontologías, como las mencionadas arriba, además de proponer algunas mejoras en el diseño de sistemas.

Si bien se mantiene su estructura general respecto a Peñalver *et al.* (2011), se refinan algunos de sus módulos y su funcionamiento, permitiendo un funcionamiento más sólido y versátil.

Destaca en este sentido la adición de la herramienta Stanford POS Tagger (Toutanova *et al.*, 2003) como utilidad de procesamiento del lenguaje natural, que posibilita el análisis lingüístico multilingüe.

Entre otras de sus novedades, sobresale la utilización de los métodos N_GRAM en el proceso de identificación de polaridad. Esta técnica se utiliza para seleccionar qué palabras deben considerarse a la hora de calcular la polaridad y está basada en la cercanía de estas palabras al aspecto que se está considerando. De este modo, el valor asignado a N_GRAM señala el número de palabras próximas al aspecto que van a tenerse en cuenta en el proceso de identificación de polaridad (Peñalver *et al*., 2014: 16), de manera que un menor valor hará que se consideren solo las palabras en el contexto más inmediato de la característica. Existen tres métodos principales *Before*, *After* y *Around,* que determinan si se seleccionan palabras que vaya antes, después o alrededor del aspecto valorado, además de un cuarto método *All Phrase*, que considera las palabras de la oración entera en que se encuentre el aspecto. En este estudio, la aplicación de estos métodos se demostró de gran utilidad, especialmente para el método N_GRAM Before, aunque esto es dependiente de la lengua considerada, por lo que no coincidirá necesariamente en lenguas distintas de la inglesa.

Estas técnicas de identificación de polaridad complementan la fase posterior de asignación de polaridad. Dado que se mantiene el sistema de asignación de pesos a cada característica de Peñalver *et al*. (2011), se calcula la polaridad de cada aspecto mediante el producto escalar de su peso y el vector de polaridad obtenido mediante la media aritmética de las polaridades de las palabras obtenidas mediante los métodos N_GRAM. Al realizar el sumatorio de todos los aspectos valorados, se obtiene un vector de polaridad final, cuya componente de mayor valor determina el resultado del análisis.

Es importante anotar que la comparación de los resultados respecto a otros sistemas de minería de opinión es complicada, debido a la diversidad de recursos y dominios aplicados. Resulta relevante, por tanto, la propuesta de un sistema de estandarización en la evaluación de Schouten y Frasincar (2015), que sin duda podría ayudar a poner en situación esta propuesta en un contexto más amplio:

> We would like to stress that transparency and standardization is needed in terms of evaluation methodology and data sets in order to draw firm conclusions about the current state-of-the-art. Benchmark initiatives like SemEval or GERBIL that provide a controlled testing environment are a shining example of how this can be achieved. (*ibíd*.: 17)

En cualquier caso, la efectividad de la propuesta de Peñalver *et al*. (2014) parece clara:

> Our approach, when using the "N_GRAM Before" method, has achieved an accuracy of 89.6% for the sentiment classification of all the opinions as regards positive, negative and neutral opinions. The best results of our approach are therefore comparable to the results obtained in previous works for the classification of positive and negative opinions. (*ibíd*.: 30)

En síntesis, la contribución de este estudio se basa en:

> First, an ontology-based feature identification permits the reuse of existing vocabularies and ontologies in order to extract feature related information from opinions in different domains. Second, four different configurable methods for feature polarity identification are proposed. These methods can be configured with different parameters to obtain the best polarity identification approach for different domains and languages. Finally, the vector analysis based opinion mining approach permits the sentiment classification of a document to be calculated, and no training phases are necessary. (Peñalver *et al*., 2014: 31)

Sin embargo, se admite que su principal obstáculo es la necesidad de una ontología del dominio específica, pues su desarrollo es una tarea laboriosa. Al mismo tiempo, la estaticidad del conocimiento recogida en esta es otro factor limitante. Por ello, se aboga por la experimentación de métodos de desarrollo de ontologías de manera parcialmente automática, empleando las reseñas como fuente, de modo que sea pueda actualizar el conocimiento representado, mediante la extracción de nuevas entidades y clases directamente de las opiniones utilizadas (*ibíd*.: 32).

Salas Zárate, Valencia García *et al.* (2017) parte de la propuesta de Peñalver *et al*. (2014), para su aplicación en el ámbito de noticias financieras. Por ello, el sistema propuesto se compone del mismo tipo de módulos y enfoques, pero se construye una ontología que estructura semánticamente la información del ámbito financiero, atendiendo a la referencia de otras ontologías desarrolladas anteriormente en ese ámbito. Mantiene el mismo planteamiento de Peñalver *et al*. (2014) en lo relativo a la identificación de polaridad y cálculo vectorial, incluyendo cálculos vectoriales y métodos N_GRAM.

A la vista de sus resultados, esta aplicación resulta igualmente alentadora, a pesar del cambio de ámbito. A diferencia de Peñalver *et al*. (2014), el método N_GRAM Around fue el que dio mejores resultados en el dominio financiero (y nuevamente en inglés). Nuevamente, las dificultades en la comparación limitan las comparaciones posibles respecto a otros enfoques, a pesar de lo cual se cita que presenta buenos datos en relación con otros sistemas que se aplican al mismo ámbito (Salas Zárate, Valencia García *et al.,* 2017: 19).

Sus conclusiones van también en la misma línea que las de Peñalver *et al*. (2014) en lo que concierne a sus limitaciones. La intensa labor manual de construcción de ontologías destaca como el mayor inconveniente, por lo que de nuevo se apunta a un posible intento de

automatización de esta tarea. Además, se anota por primera vez como desventaja su funcionamiento en una sola lengua:

> The proposed method is only able to deal with news expressed in English, which is a disadvantage owing to the vast amount of information available in other languages. We shall therefore attempt to apply this method to the Spanish language, since it has been studied less frequently in the opinion mining field. Furthermore, it should be mentioned that Spanish is the third most spoken language in the world, and we therefore firmly believe that the computerisation of Internet domains in this language is of the utmost importance. (Salas Zárate, Valencia García *et al.,* 2017: 19)

Estas ideas vuelven a aparecer en Salas Zárate, Medina Moreira *et al.* (2017), donde se desarrolla una investigación similar a la de Salas Zárate, Valencia García *et al.* (2017): si bien se destina al ámbito médico y sus textos proceden de Twitter, la metodología es la misma. Utiliza SentiWordNet y una ontología específica para el dominio de diagnosis de diabetes (el foco del estudio), así como los mismos métodos de cálculo de polaridad establecidos en Peñalver *et al*. (2014); no obstante, introduce un mecanismo de desambiguación (*word sense disambiguation*, WSD) para mejorar el rendimiento de SentiWordNet, mediante el uso de Babelfy (Moro, Raganato y Navigli, 2014), sistema basado también en estructuras semánticas:

> Our joint solution is based on three key steps: i) the automatic creation of semantic signatures, i.e., related concepts and named entities, for each node in the reference semantic network; ii) the unconstrained identification of candidate meanings for all possible textual fragments; iii) linking based on a high-coherence densest subgraph algorithm. (*ibíd*.: 241)

Sus resultados y conclusiones van por tanto en la misma línea. Así como en Salas Zárate, Valencia García *et al.* (2017), los resultados señalaron que el método de mayor fidelidad fue el N_GRAM Around, "with a precision of 81.93%, a recall of 81.13%, and an *F*-measure of 81.24%" (Salas Zárate, Medina Moreira *et al.,* 2017: 7). Del mismo modo, se señalan nuevamente las limitaciones de la operación únicamente en inglés del sistema y de la necesidad de una ontología específica. Estos dos aspectos han sido señalados reiteradamente como pendientes de resolver, por lo que trataremos de buscar una aproximación a estas cuestiones a lo largo del presente proyecto.

Si bien la cantidad de investigaciones en AS ha crecido considerablemente, la mayoría de estas se han centrado en datos correspondientes a la lengua inglesa, en palabras de Dashtipour *et al*. (2016):

> The majority of current sentiment analysis systems address a single language, usually English. However, with the growth of the Internet around the world, users write comments in different languages. Sentiment analysis in only single language increases the risks of missing essential information in texts written in other languages. In order to analyse data

> in different languages, multilingual sentiment analysis techniques have been developed. (*ibíd*.: 757)

La necesidad de responder a una realidad lingüística y cultural más amplia es la principal motivación del estudio de AS multilingüe:

> In fact, only 28.6 % of the Internet users speak English. It is thus essential to explore or build resources and tools in languages other than English. Moreover, Asia now has the most Internet users (48.2%); followed by Europe (18%). As a result, there is a growing need to work on languages such as Chinese and Japanese. (Lo *et al*., 2017: 501)

Sin embargo, el mayor obstáculo de esta orientación ha sido hasta ahora la falta de recursos específicos disponibles:

> Due to the multilingual nature of social media data, analysis based on a single official language may carry the risk of not capturing the overall sentiment of online content. While efforts have been made to understand multilingual sentiment analysis based on a range of informal languages, no significant electronic resource has been built for these localised languages. (*ibíd*.: 499)

De este modo, en muchas lenguas, el proceso de análisis sentimental se enfoca mediante la transferencia del conocimiento específico de una lengua que disponga abundantes recursos, como el inglés, a otra con mayor escasez, pues no existen recursos disponibles de otra manera en esas lenguas (Dashtipour *et al*., 2016: 757).

Otra de las principales alternativas es el empleo de traducción automática como parte del sistema de MO para convertir las opiniones originales al inglés y poder utilizar módulos dedicados a este idioma. Esto, no obstante, también tiene sus inconvenientes, pues en ocasiones, los sistemas automáticos no traducen correctamente partes importantes de un texto, de modo que se pierden aspectos esenciales y pueden producirse problemas que reduzcan el número de frases bien construidas (*ibíd*.: 758).

En cualquier caso, la posición respecto al valor de la traducción automática es compleja, ya que se observa que el análisis sentimental multilingüe basado en corpus paralelos en lugar de traducción automática puede mejorar la precisión en la clasificación (Lo *et al*., 2017: 514), pero al mismo tiempo se han realizado investigaciones que aplican estas técnicas con buenos resultados y confían en las capacidades de los sistemas actuales de traducción automática:

> While poor performance of multilingual sentiment analysis may be due to the limitation of a machine translation system, Balahur and Turchi (2014) conducted extensive evaluation scenarios to show that machine translation systems are mature enough to obtain multilingual data for supervised sentiment analysis. (*ibíd*.: 514)

Balahur y Turchi (2012, 2013, 2014) desarrollan estudios y evaluaciones sobre el uso de los sistemas de traducción automática de base estadística (*statistical machine translation*, o SMT). Balahur y Turchi (2012) presentan un estudio comparativo de la aplicación de tres sistemas (Bing Translator, Google Translate y Moses) en tres idiomas (alemán, francés y español) en el proceso de AS de base estadística (empleando el algoritmo *Support Vector Machines Sequential Minimal Optimization,* o SVM SMO). En Balahur y Turchi (2013) se emplea el sistema de traducción de Google, con un idioma añadido a los tres de Balahur y Turchi (2012), el italiano, y la implementación Weka de SVM SMO, para el análisis sentimental de tweets.

Por su parte, Balahur y Turchi (2014) propone una implementación y evaluación de esta propuesta de mayor peso:

> The present work studies the possibility to employ machine translation systems and supervised methods to build models able to detect and classify sentiment in languages for which less/no resources are available for this task when compared to English, stressing upon the impact of translation quality on the sentiment classification performance. Our extensive evaluation scenarios show that machine translation systems are approaching a good level of maturity and that they can, in combination to appropriate machine learning algorithms and carefully chosen features, be used to build sentiment analysis systems that can obtain comparable performances to the one obtained for English. (*ibíd*.: 56)

El planteamiento del estudio es similar al de los dos trabajos previos: utiliza metodologías similares, mismo set de datos, idiomas y sistemas de traducción que Balahur y Turchi (2012) y mismo sistema de *machine learning* que en Balahur y Turchi (2013). Sin embargo, en este se realiza un esfuerzo más grande de experimentación y testado de los sistemas.

> Our findings show that SMT systems have reached a reasonable level of maturity to produce sufficiently reliable training data for languages other than English. The gap in classification performance between systems trained on English and translated data is minimal, with a maximum of 12% in favor of source language data. (Balahur y Turchi, 2014: 69)

A pesar de esto, se anota que "working with translated data implies an increment number of features, sparseness and noise in the data points in the classification task" (*ibíd*.: 69), por lo que implementan otros métodos en sus ensayos para lidiar con estos problemas. Finalmente, otro de los inconvenientes de este tipo de planteamientos se refiere a la disponibilidad de sistemas de traducción automática en diversas situaciones:

> The proposed approach clearly depends on the availability of the translation engines for the required languages. Although, commercial engines are able to translate from and into a large number of languages, they cannot be used to freely translate large amounts of data (usually not more than a certain number of characters). On the other hand, the parallel corpora needed for training the open source SMT systems cover only the most used languages, and their sizes are not comparable to the dataset used to train commercial

> engines. These aspects may limit the use of MT in support of other natural language processing sectors, in particular if focused on less resourced languages. (*ibíd*.: 69)

Pese al optimismo mostrado, es un hecho que todavía no existe un acuerdo sobre el grado de validez de los sistemas de traducción automática en la MO (Lo *et al*., 2017).

Ante esta situación, es natural que vayan surgiendo recursos multilingües desarrollados para su empleo en sistemas de MO de manera directa; "however, sentiment analysis corpora and resources, even if created for multiple languages, cannot be used for other languages. More research is required to improve results in the multilingual sentiment analysis discipline" (Dashtipour *et al*., 2016: 758). Pese al creciente interés en este ámbito, su desarrollo actual aún queda lejos del existente en el dominio del inglés monolingüe.

Entre las técnicas principales empleadas, Lo *et al.* (2017) mencionan las basadas en lexicones, corpus o traductores automáticos, junto a otras basadas en conceptos o *sentics* (que incorporan razonamiento e información perteneciente al sentido común). Por su parte, Dashtipour *et al*. (2016) destacan las basadas en corpus y técnicas de *machine learning*, las basadas en lexicones (como el mencionado SentiWordNet) y las híbridas, que combinan ambos métodos. En su estudio, revisan la situación existente en el panorama multilingüe, sin embargo:

> For practical applications and for research work, one would need to choose the best performing approaches. However, direct comparison between those systems is difficult due to a number of factors. First, the original authors report the results on very different datasets, which makes comparison between the reported figures not fair. More importantly, the original authors describe their systems with varying degree of detail and accuracy, which makes the reported results not always reproducible. With this, even if a method showed excellent results in the authors' own evaluation, lack of detail in their publication may render it unusable in practice for the readers. (*ibíd*.: 767)

La ausencia de criterios de referencia en la evaluación de sistemas surge recurrentemente como una de las limitaciones más notables de la MO como disciplina. Observamos que, por un lado, limita la capacidad de comprender las investigaciones y sus logros, ante la imposibilidad de valorar sus conclusiones comparativamente; y, por otro, no existe un estándar de calidad en los estudios mediante el cual poder medir su rigurosidad y que sirva como indicativo para la aplicabilidad de sus desarrollos, pues el verdadero valor que tiene una técnica para la comunidad investigadora reside en los resultados que puedan reproducirse mediante su empleo y no solo en los que supuestamente fueran obtenidos originalmente por sus autores (*ibíd*.: 767). En su investigación, Dashtipour *et al*. (2016) implementan once propuestas ya existentes tratando de seguir de cerca las indicaciones existentes en sus publicaciones originales.

> In the majority of the cases, we obtained lower results than those reported by their corresponding authors. We attribute this mainly to the incompleteness of their descriptions in the original papers. In some cases, though, the methods were developed for a specific domain, so in such cases comparison on our test corpora may not be fair. A lesson learnt was that for a method to be useful for the research community, authors should provide sufficient detail to allow its correct implementation by the reader. (*ibíd*.: 769)

A pesar de todo, concluyen que, como ya se adelantó, la falta de recursos multilingües específicos es la principal traba de este tipo de investigaciones (*ibíd*.), mucho más en los casos de lenguas que tengan escasos recursos lingüísticos (Lo *et al*., 2017: 521).

3. Metodología

Entendemos en este trabajo metodología como la sistematización de los criterios a seguir en la definición del diseño de nuestro sistema de minería de opinión, con el objetivo de establecer una serie de guías que determinen tanto su arquitectura y componentes como las funcionalidades buscadas en su utilización. Para ello, trataremos de partir de algunas de las referencias bibliográficas básicas, siguiendo aquellas pautas que hayan resultado exitosas y tratando de replantear algunas de las cuestiones que se han anotado como mejorables, en la busca de mejoras en el funcionamiento del sistema.

La directriz básica del diseño del sistema es que parte del análisis sentimental a nivel aspectual y apoyado mediante el uso de ontologías, siguiendo la investigación de Peñalver Martínez *et al.* (2011 y 2014), Salas Zárate, Valencia García *et al.* (2017) y Salas Zárate, Medina Moreira *et al.* (2017). Nuestro proyecto trata de introducir además un funcionamiento bilingüe en inglés y español donde cada dominio recibe un tratamiento específico independiente, pero siguiendo una misma estructura, lo que permite obtener resultados equivalentes en ambos casos.

Un gran número de los estudios analizados han enfocado sus propuestas hacia el análisis de reseñas de películas, en gran parte gracias a la facilidad de obtener muestras de la página *web* IMDb. Sin embargo, decidimos centrar nuestra atención en el mundo musical, de características similares al del cine, en lo relativo a su conceptualización o sus modos de consumo. Consideramos que este ámbito puede beneficiarse en gran medida del desarrollo de un sistema como este. De cualquier modo, al seguir los modelos marcados por la bibliografía, como en Peñalver Martínez *et al.* (2014: 7), la solución al problema de AS es independiente del dominio, pero puede adaptarse a un ámbito concreto mediante la utilización de una ontología de dominio y un corpus de opiniones específicos a esta área particular. Por tanto, nuestra adaptación concreta está destinada a su aplicación en el dominio mencionado, pero debido a su flexibilidad, podría ser de utilidad en una variedad de dominios deseados.

Siguiendo las directrices de los estudios revisados, la arquitectura del sistema es modular, considerando la actividad de cada módulo como un proceso diferenciado del resto en el que, a partir de la información que recibe de entrada, produce un resultado que se devuelve al siguiente módulo del sistema. La separación de los distintos procedimientos es clara, de modo que cada parte tiene una función detallada e imprescindible para la consecución de la

labor final. A grandes rasgos, la tarea de AS aspectual conlleva tres procesos principales, en las palabras de Schouten y Frasincar (2015):

> The first step is concerned with the identification of sentiment-target pairs in the text. The next step is the classification of the sentiment-target pairs. The expressed sentiment is classified according to a predefined set of sentiment values, for instance positive and negative. Sometimes the target is classified according to a predefined set of aspects as well. At the end, the sentiment values are aggregated for each aspect to provide a concise overview. (*ibíd.*: 2)

En nuestro caso, para su desarrollo, seguiremos una arquitectura de sistema similar a la propuesta en Zhao y Li (2009), Peñalver Martínez *et al.* (2011, 2014), Salas Zárate, Valencia García *et al.* (2017) y Salas Zárate, Medina Moreira *et al.* (2017), muy similar en todos estos casos. De este modo, el análisis sentimental parte del input textual de las opiniones de usuario. A partir de este input, el sistema pre-procesa en primer lugar el lenguaje natural, para posteriormente seleccionar las características que el texto menciona y realizar una evaluación de su polaridad, que finalmente se barema globalmente para obtener un resultado concreto. Los tres procesos señalados por Schouten y Frasincar (2015) en la cita anterior se integran en tres módulos individuales.

Atendiendo a estos requisitos básicos, nuestro sistema se caracteriza por:

-La MO sigue la orientación aspectual y basada en ontologías. Sin embargo, más concretamente, se realizará un análisis de las opiniones a nivel de documento, esto es, considerando cada texto, que puede ser de una extensión variable, de manera conjunta como opinión correspondiente a una persona. De este modo, el resultado del AS obtenido al final del sistema se referirá a la opinión contenida en el documento entero.

-Como soporte ontológico para la identificación de características, empleamos WikiData (Vrandecic y Krötzsch, 2014), la base de conocimiento colaborativa y libre que almacena de manera estructurada los datos de Wikipedia. Su utilización aquí responde a una serie de factores: en primer lugar, organiza la información del dominio de la música de manera actualizada y estructurada; en segundo lugar, su diseño permite el uso multilingüe, ya que los códigos identificativos (o *item IDs*) que se asignan a sus entradas pueden utilizarse como identificadores independientes del lenguaje, de forma que sus esquemas ontológicos son compartidos, lo que facilita el intercambio de información (*ibíd*.: 84); y finalmente, su fácil acceso, puesto que la función básica de Wikidata es permitir que sus datos se utilicen tanto en Wikipedia como otras aplicaciones externas, para lo cual la información se exporta a través de servicios *web* (servicios que permiten el intercambio de datos entre aplicaciones)

empleando formatos habituales en este tipo te tareas, como JSON (JavaScript Object Notation) y RDF (Resource Description Framework) (*ibíd.*: 79). De este modo, las aplicaciones pueden utilizar la API de Wikipedia para buscar, consultar o incluso editar su información. En los casos en que las consultas requeridas son más complejas, se dispone de Wikidata Toolkit, un set de herramientas de código libre, que permiten realizar la extracción y manipulación de datos a escalas mayores, mediante volcados de información que pueden ser actualizados en tiempo real, a modo de imagen espejo de Wikidata, incluyendo los cambios que pudieran realizarse en la plataforma (*ibíd.*: 84).

De esta manera nos proponemos resolver los inconvenientes del desarrollo manual de ontologías, tarea demasiado costosa, al mismo tiempo que dar soporte a más de un idioma a través de la misma estructura ontológica, lo que permite garantizar el funcionamiento análogo en todas las lenguas. Por último, el hecho de que la información se actualice periódicamente es de gran utilidad en un ámbito tan dinámico como el de la música. A esto contribuye que Wikidata asigna identificadores a cada entidad que la vinculan con diferentes plataformas, entre las que destacamos Spotify. Los datos de esta plataforma pueden utilizarse para el cotejado de información específica de la ontología. Para ello se pueden realizar consultas a la API de Spotify (2021), que permite el acceso y la utilización del amplio catálogo de obras y *metadata* asociado (como artistas, álbumes, canciones, fechas, géneros, etc.) de la plataforma musical.

-Como base para el análisis sentimental se emplea un lexicón de palabras sentimentales, concretamente SentiWordNet (Baccianella, Esuli y Sebastiani, 2010), ampliamente utilizado en la bibliografía revisada en las investigaciones de esta misma línea. Para dar soporte a las opiniones en lengua española, hemos decidido emplear CRiSOL (Molina González, Martínez Cámara y Martín Valdivia, 2015), un recurso léxico de similar planteamiento y en español que, además, asigna a sus contenidos los valores de polaridad originales de SentiWordNet, por lo que el tipo de información se mantiene constante.

-En la clasificación de opiniones, continuamos la propuesta iniciada en Peñalver Martínez *et al.* (2011, 2014), consistente en cálculos y análisis vectoriales:

> The methodology described in this paper combines the use of domain ontologies with a new technique based on vector calculation using spatial distances in R^3 to improve the sentiment analysis process. Overall, our proposed architecture results in a more detailed analysis of user opinions in two dimensions: (i) the general user opinion and (ii) all the features identified in the user's opinion. Our approach is not constrained to the use of a single very specific topic, since the proposed method is capable of analyzing and classifying the sentiments of each topic discussed by users in their opinion set. (*ibíd.*: 5)

Del mismo modo, al seguir este paradigma, el propósito es desarrollar un sistema completamente automático para enfrentar la labor de MO.

-Respecto al lenguaje, sucede de la misma manera que con el dominio, a pesar de que se remarca en Peñalver Martínez *et al.* (2014) que su propuesta de AS es independiente del idioma y que se puede adaptar a uno específico mediante una ontología del dominio y un corpus de opiniones en el lenguaje objetivo (*ibíd*.: 7), no hemos encontrado hasta ahora propuestas de AS aspectual con base ontológica con funcionamiento multilingüe que continúen las orientaciones propuestas por Peñalver Martínez *et al.* (2014), Salas Zárate, Valencia García *et al.* (2017) o Salas Zárate, Medina Moreira *et al.* (2017)

Además, si bien es cierto que en el sistema propuesto en Peñalver Martínez *et al.* (2011, 2014) el corpus de opiniones y la ontología son los elementos más determinantes, no se considera en ese análisis el funcionamiento de SentiWordNet, como lexicón en inglés. En algunos proyectos, como Silva, Bastos y Rocha (2018), se ha optado por traducir de manera automática el contenido de SentiWordNet para su aplicación en el AS; no obstante, este proceso debe realizarse de manera supervisada, para garantizar la correcta traducción y aplicación de los términos. No podemos confiar en que una simple traducción automática vaya a permitir su utilización en condiciones similares. Por lo que, por razones de rigor, decidimos emplear recursos desarrollados de manera específica para este fin.

Ante la bibliografía revisada, cabría también la opción de desarrollar un sistema multilingüe impulsado por un sistema de traducción automática y recursos lingüísticos específicos para la lengua inglesa. Sin embargo, ante la falta de acuerdo y las cuestiones que arroja el uso de traducción automática en el contexto de textos de opinión, optamos por un desarrollo multivía, en el que un detector de idioma selecciona el camino que se debe recorrer para realizar el análisis. De este modo, cada lengua puede tener un tratamiento específico mediante la utilización de recursos desarrollados para sus características. Confiamos en que este tratamiento lingüístico personalizado produzca una mayor eficacia en los resultados.

En nuestro caso, debido a que el inglés y el español son los principales idiomas en la industria musical global, decidimos diseñar un sistema que soporte el funcionamiento en ambos indistintamente, pero teniendo un tratamiento diferenciado, como acabamos de exponer. Los módulos empleados son similares en ambos casos, salvo en lo relativo a los recursos lingüísticos específicos para cada idioma.

4. Descripción del sistema

En los siguientes subapartados vamos a analizar con detalle cada uno de los módulos que configuran nuestro sistema de minería de opinión, para describir detalladamente su utilidad y funcionamiento dentro del conjunto, siguiendo el esquema definido en la Ilustración 1:

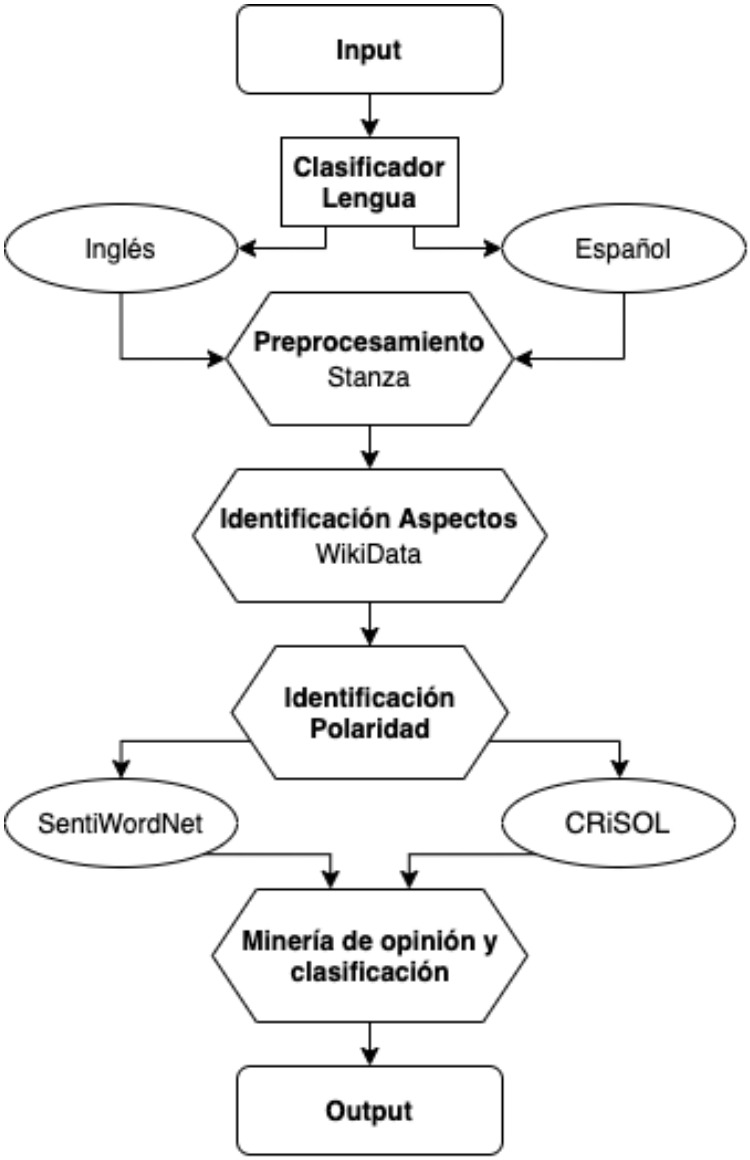


Ilustración 1: Diagrama de flujo del sistema

El sistema parte de una entrada textual que se corresponde con la reseña a analizar. El diseño se configura para el tratamiento de textos procedentes de diferentes fuentes de internet. Para el ámbito musical, existen plataformas dedicadas como AllMusic.com o Discogs, además de redes sociales genéricas como Youtube o Twitter, donde se puede recuperar contenido de opinión a través de sus APIs correspondientes, o librerías como Twitter4J, utilizada en Salas Zárate, Medina Moreira *et al.* (2017).

Cualquier página similar que permita recuperar sus comentarios o *posts* mediante solicitudes a una API puede ser integrada en el flujo del sistema. Contemplamos por tanto en el diseño del sistema diferentes tipos de configuración según las características de la plataforma de la que se recuperen las opiniones.

4.1 Clasificación de lengua

El primer módulo del sistema es el encargado de detectar el lenguaje del texto de entrada, para poder tratarlo adecuadamente en el resto de los módulos. Así, si se detecta inglés o español, se procederá a seguir su ruta correspondiente; en cualquier otro caso, el sistema arrojará un error, pero anotará el lenguaje correspondiente. Mediante el conteo de las lenguas encontradas y no soportadas por el sistema se podrá analizar periódicamente si existen problemas en el diseño, es decir, si surgen fallos en la clasificación, o si un idioma aparece con abundante frecuencia, lo que podría sugerir nuevos desarrollos o ampliaciones para nuestro sistema de AS.

La herramienta implementada para realizar esta clasificación es LingPipe Language ID, que forma parte de una extensa suite de aplicaciones java desarrolladas para ejecutar diversas tareas de PLN (Carpenter, s.f.). Para realizar esta tarea, LingPipe Language ID enfoca la detección de lenguas como un problema de clasificación, para el cual desarrolla modelos de cada lenguaje empleando textos de entrenamiento, extraídos del Leipzig Corpora Collection (Denecke, 2008):

> The classifier for language identification learns the distribution of characters per language using language models. A language model assigns a probability to a sequence of words $P(w_{1...n})$ by means of a probability distribution. It aims to predict the probability of natural word sequences: Word sequences that actually occur achieve a high probability whereas those sequences that never occur get a low probability.
> An n-gram model approximates these probabilities by assuming that the only words relevant to predicting $P(w_i|w_1...w_{i-1})$ are the previous n-1 words:
> $$P(w_i|w_1...w_{i-1}) = P(w_i|w_{i-n+1}...w_{i-1})$$
> Like text classifiers that attempt to identify attributes which distinguish documents in different categories, language models also attempt to capture such regularities.
> For text classification using an n-gram language model, a language model for each category is trained based on a set of training data. To classify a new (unknown) document, a language model is calculated and compared to the trained language models. The category of the language model which is most similar to that of the unknown document is assigned.
> For language classification with LingPipe, n-grams of size eight are used. (*ibíd.*: 1)

Una descripción más detallada de los modelos de lenguaje empleados en LingPipe se encuentra en Carpenter (2005).

De esta manera, se analiza la distribución estadística de caracteres en cada lenguaje, estableciendo patrones de comparación empleados en el proceso de clasificación,

comparando las probabilidades de que los caracteres de un texto dado se correspondan con un lenguaje u otro. Como indica su *web*, este tipo de modelos lingüísticos están entre los más precisos en la tarea de clasificación textual en general y, en particular, su uso en la identificación de lenguas es bastante oportuno, ya que no requiere el procesado de la entrada textual y tareas como el *tokenizado* son dependientes de la lengua (Carpenter, s.f.). Por esta misma razón decidimos emplearlo en el sistema, ya que toma parte en una fase inicial del proceso, mientras que el procesado lingüístico en sí solo se efectúa a continuación.

Otro aspecto positivo de la herramienta es la posibilidad de generar nuevos modelos, algo que podría ser interesante para optimizar su funcionamiento, utilizando corpus de opiniones de internet como material de entrenamiento, pues quizá sus características difieran de los contenidos de Leipzig Corpora Collection y pudieran aportar algo más de precisión al sistema de clasificación.

Como resultado de su ejecución, LingPipe Language ID ofrece un ranking ordenado de las lenguas que estima que pueden asignarse al texto de entrada. Así, siguiendo el ejemplo mostrado en su *web*, al analizar el siguiente texto en catalán "per poder jutjar l'efectivitat d'una novetat Acs imprescindible deixar" (*ibíd.*), se obtiene el siguiente listado:

```
Input=Per poder jutjar l'efectivitat d'una novetat Acs imprescindible deixar
Rank  Category  Score  P(Category|Input)   log2 P(Category,Input)
0=cat -2.023136071289025 1.0 -145.6657971328098
1=fr -4.321846555117093 1.504464337461871E-50 -311.17295196843065
2=it -4.581659915472809 3.516783524040892E-56 -329.87951391404226
3=nl -4.696851550631136 1.1206338610900442E-58 -338.1733116454418
4=en -4.766148642045668 3.52782891712393E-60 -343.16270222728804
5=no -4.958001178925975 2.4505580508882693E-64 -356.9760848826702
6=se -4.987497723277155 5.622794125159236E-65 -359.0998360759552
7=dk -5.104818791491201 1.6110781779891577E-67 -367.54695298736647
8=tr -5.219139511312786 5.361805775146849E-70 -375.7780448145206
9=de -5.265340898319162 5.344841592730491E-71 -379.10454467897966
10=sorb -5.518926633655108 1.704814443523678E-76 -397.3627176231678
11=ee -5.579930924219301 8.118211859258627E-78 -401.75502654378965
12=fi -6.07538483408664 1.4822218818370195E-88 -437.4277080542381
13=jp -10.067560120929357 4.404215405114351E-175 -724.8643287069137
14=kr -10.728373897912512 2.0954882281132754E-189 -772.4429206497009
```

Ilustración 2: Resultados de LingPipe Language ID (Carpenter, s.f.)

Su interpretación es la siguiente:

> The output is presented as a rank-ordered of predicted categories plus statistics. The language predicted by the classifier is cat (Catalan); the second-best match is fr (French). After the ranks and categories, there are three numbers per line. The second is perhaps the most useful, as it is a conditional probability estimate of the category given the input. For the input given, the conditional estimate is that (within rounding error), the choice of Catalan is 100% certain. The chance of the input being French, according to the classifier's

> estimate, is very very very low (roughly 1/1050). The last column is the log (base 2) joint probability estimate of the category and the input. The first column is the score, which is a kind of entropy rate, which is roughly the log joint probability estimate divided by the length of the input. (*ibíd.*)

De modo que se detecta con total confianza (100% de probabilidad) que el texto está escrito en catalán y el resto de los idiomas propuestos tienen unos valores de probabilidad despreciables en comparación.

Dado que la implementación inicial de LingPipe Language ID utiliza solo corpus textuales de 15 lenguas entre las que no se encuentra el español, será necesario reentrenar el sistema inicialmente para permitir también la funcionalidad en esta lengua. Este proceso, no obstante, es relativamente trivial y queda claramente reflejado en la documentación de la herramienta en la página web de LingPipe[10]. Tan solo habría que descargar los datos del Leipzig Corpora Collection[11] relativos a las lenguas que se quiera soportar y ejecutar una serie de comandos para, en primer lugar, transformar los datos al formato requerido por el sistema y, luego, reentrenar y evaluar el nuevo modelo lingüístico.

Esta tarea solo habría que realizarla una vez en la fase de desarrollo del sistema. Una vez disponible el modelo lingüístico, se puede operar reiteradamente sin que sea necesario realizar un mantenimiento de la herramienta. Además, como comentamos anteriormente, sería posible realizar este entrenamiento con un corpus diferente del propuesto, pero consideramos que en una primera etapa de desarrollo, su funcionamiento por defecto es lo suficientemente útil para diferenciar entre los dos idiomas soportados por el sistema, por lo que descartamos realizar un entrenamiento más detallado, que requeriría un tiempo adicional de desarrollo e implementación, pero difícilmente produciría resultados mucho mejores en el tipo de aplicación que buscamos darle.

Una vez configurado de esta manera, este módulo será capaz de detectar si los textos presentados están escritos en inglés o español. De esta manera, en su aplicación en nuestro sistema, los textos pertenecientes al inglés continuarán una ruta distinta de los del español, y podrán así analizarse mediante herramientas propias de cada lengua. Todo texto detectado como correspondiente a otro idioma distinto de estos dos no será procesado y se descartará. De esta manera aseguramos el funcionamiento del sistema para los lenguajes soportados.

---

[10] Disponible en: http://www.alias-i.com/lingpipe/demos/tutorial/langid/read-me.html
[11] La descarga se puede realizar libremente desde: https://corpora.uni-leipzig.de/es?corpusId=deu_news_2020

4.2 Módulo de preprocesamiento

En esta etapa, el lenguaje natural se analiza y convierte en lenguaje formal, lo que posibilita el tratamiento informático y automático del texto en las etapas posteriores. Para ello utilizamos técnicas de procesamiento del lenguaje natural (PLN) para, primero, determinar las unidades básicas del texto y, después, extraer la información morfológica y sintáctica del documento.

Entre las tareas realizadas en este módulo se encuentran:

a) Normalización
b) Tokenización y separación de frases
c) Detección de POS (*parts of speech*)
d) Lematización

En este módulo contamos con la ayuda de las herramientas de procesamiento de lenguaje natural del Stanford NLP Group. En concreto, utilizaremos Stanza (Qi *et al*., 2020), un paquete de análisis de lenguaje natural *open source*, que introduce una adaptación de su librería CoreNLP (Manning *et al*., 2014) a Python, y tiene un soporte multilingüe completo de las tareas a realizar en esta etapa:

> Stanza, an open-source Python natural language processing toolkit supporting 66 human languages. Compared to existing widely used toolkits, Stanza features a language-agnostic fully neural pipeline for text analysis, including tokenization, multiword token expansion, lemmatization, part-of-speech and morphological feature tagging, dependency parsing, and named entity recognition. (Qi *et al*., 2020: 1)

A pesar de que existe actualmente una considerable variedad de herramientas de PLN disponibles libremente en la red, Stanza fue creada para superar algunas de las limitaciones que estas muestran.

> First, existing toolkits often support only a few major languages. This has significantly limited the community's ability to process multilingual text. Second, widely used tools are sometimes under-optimized for accuracy either due to a focus on efficiency (e.g., spaCy) or use of less powerful models (e.g., CoreNLP), potentially misleading downstream applications and insights obtained from them. Third, some tools assume input text has been tokenized or annotated with other tools, lacking the ability to process raw text within a unified framework. This has limited their wide applicability to text from diverse sources. (*ibíd*.: 1)

Así, Stanza no solo puede utilizarse en 66 idiomas distintos partiendo de texto en bruto, sino que su funcionamiento está completamente desarrollado y adaptado a los criterios de calidad actuales. Stanza se describe como una "neural multilingual NLP pipeline" (*ibíd.*: 2), un sistema modular que permite convertir un texto de entrada en un documento anotado a través

de una serie de componentes que realizan una serie de tareas de PLN, como se observa en la Ilustración 3:

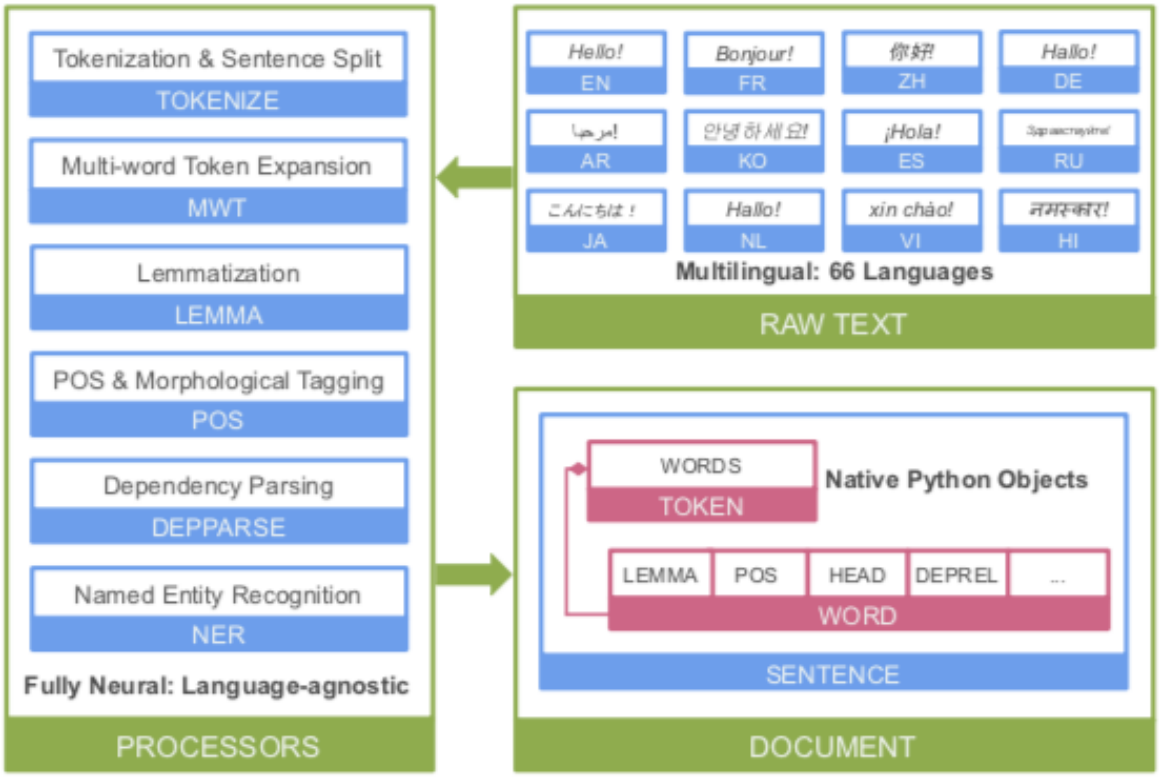


Ilustración 3: Estructura de Stanza (Qi et al.*, 2020: 1)*

Los procesos disponibles en Stanza son los siguientes (*ibíd.*: 2-3):

- Tokenización y separación de oraciones. En un primer paso, entre las secuencias de caracteres del texto de entrada se reconocen las palabras y se separan en oraciones; también deben reconocerse aquellos *tokens* que estén compuestos por varias palabras (*multi-word tokens,* o MWT).
- Una vez identificados, los MWT se expanden mostrando su estructura subyacente, como puede ser en: *del* > *de el.*
- Anotación de partes del discurso (*part-of-speech*, POS) y características morfológicas. A cada una de las palabras separadas en la fase anterior, Stanza le detecta y asigna su POS, así como sus características morfológicas universales, llamadas *UFeats*, entre las que se encuentran rasgos como *singular/plural* o *1st/2nd/3rd person* (*ibíd.*: 2). Así, los *tokens* no solo reciben una etiqueta con su parte de la oración correspondiente, sino que también se anotan sus atributos morfológicos fundamentales.

- Lematización. En esta etapa, se devuelve cada palabra a su forma base, es decir, el lema, extrayendo los afijos que pudieran encontrarse. De este modo, se consideran conjuntamente las distintas formas que pudieran ocurrir para un mismo lema.
- Análisis de dependencias. Como paso posterior en el análisis, se extrae la estructura sintáctica de la oración, señalando las relaciones existentes entre palabras.
- Reconocimiento de nombres de entidades. Se identifican dentro de cada una de las frases las entidades nombradas según una variada tipología (nombres de persona, organizaciones, lugares, obras de arte…).

Sin embargo, puesto que el sistema va a manejar opiniones extraídas de internet, hemos decidido implementar una serie de procesos previos al procesado de Stanza para asegurar que el texto puede ser procesado adecuadamente, tal como se hizo en Salas Zárate, Medina Moreira *et al.* (2017). Esta tarea se realizará en la primera de las fases descritas al inicio de este apartado, normalización. Su objetivo será por tanto certificar que todos los textos siguen unos mismos estándares en cuanto al tipo de caracteres o expresiones encontradas, para evitar irregularidades en el proceso.

Por lo tanto, atendiendo a esto el funcionamiento final de este módulo será el siguiente:

a) Normalización

Se procesa el texto en primer lugar para eliminar caracteres especiales procedentes de etiquetado html, urls o caracteres procedentes de la plataforma en cuestión (como hashtags). De manera análoga, los signos de puntuación repetidos (como múltiples exclamaciones o interrogaciones) serán eliminados (Balahur y Turchi, 2013). En caso de encontrar menciones a usuarios de la plataforma (mediante el símbolo @), se realizará un tratamiento particular.

Puesto que las opiniones pueden estar destinadas a un usuario concreto de la plataforma, pero también pueden mencionar en ellas a alguna entidad a la que refieran (como en "*Still loving @rihanna's 'Cry'. Such an amazing song"),* establecemos un sencillo mecanismo para elegir qué tipo de tratamiento darle a cada mención de usuario. Dado que Wikidata vincula los perfiles de artistas a sus redes sociales, cotejaremos el usuario mencionado con las entradas de la ontología y en caso de que exista una correspondencia, la sustituiremos en el texto de entrada para que pueda ser reconocida posteriormente. En caso de no encontrar ninguna coincidencia, consideramos que la mención no se refiere a entidades de nuestro dominio y no se considerará posteriormente en el análisis sentimental.

En caso de encontrar emoticonos o *emojis* en el texto, dado que estos sí que tienen incidencia en el contenido sentimental de la opinión, implementamos una vía adicional para procesarlos. En cuanto a los emoticonos, pueden clasificarse en dos tipos, los occidentales como :) o :'(y los orientales como (>_<) o (^_^). Siguiendo la propuesta de Elfajr y Sarno (2018), construimos un documento que funciona como diccionario de emoticonos, pues los agrupa en diferentes categorías y les asigna valores de polaridad siguiendo el mismo formato de SentiWordNet.

Para los *emojis*, procedemos de manera similar, utilizando el recurso desarrollado en Kralj Novak *et al.* (2015), de nombre Emoji Sentiment Ranking[12], para construir un diccionario de *emojis* dedicado, donde se relacionarán los *emojis* con sus valores de polaridad correspondientes. De este modo, en las fases posteriores del análisis se podrá considerar el valor de polaridad de estos símbolos de la misma manera en que se consideran el resto de las palabras e independientemente del idioma.

A continuación, se procesa el texto mediante un diccionario Hunspell (Németh, s.f.), herramienta multilingüe con utilidades de corrector ortográfico, con el objetivo de detectar y corregir las posibles faltas ortográficas existentes en los comentarios, ya que impedirían el funcionamiento adecuado del resto del sistema. Además, aquellas expresiones no identificadas por esta herramienta recibirán un tratamiento doble para tratar de encontrar un término que sí pueda ser utilizado a lo largo del sistema.

En primer lugar, de un modo similar a Balahur y Turchi (2013: 52), en caso de existir letras repetidas secuencialmente en una palabra (como podría ser en 'geniaaaal'), estas se reducen a dos (o una) hasta que se pueda encontrar una correspondencia en Hunspell. En un segundo caso, se compararán las palabras con los términos de un tesauro de terminología de internet, como risas ('hahaha', 'lol' o 'jeje') (Jiménez Zafra *et al.*, 2019), abreviaturas ('mñn' en lugar de 'mañana') (Salas Zárate, Medina Moreira *et al.,* 2017) o jerga (Balahur y Turchi, 2013), desarrollado específicamente a partir de las expresiones más comunes, utilizando como referencia diversos recursos de Internet[13]. Este recurso tendrá dos variedades distintas dependientes de la lengua del texto, detectada en el módulo anterior.

[12]Disponible en http://kt.ijs.si/data/Emoji_sentiment_ranking/
[13] Como https://slangit.com/, https://www.urbandictionary.com/, https://www.slanglang.net/abbreviations/ o https://global.oup.com/booksites/content/9780199543403/resources/spanish_sms

Una vez realizados estos pasos, se podrá enviar el texto a Stanza y realizar el análisis lingüístico en sí. Esta fase correctiva inicial, si bien no se menciona en muchas de las investigaciones que hemos revisado, como por ejemplo Peñalver Martínez *et al.* (2011, 2014), resulta importante en un sistema que analiza comentarios de redes sociales, a menudo escritos de manera informal. Por ello, la tarea de construcción manual de los lexicones indicados debe ser considerada como parte del desarrollo del sistema, aunque gracias a los recursos específicos disponibles en la red, su recopilación es mucho menos costosa.

La implementación de recursos lingüísticos adicionales también aumenta la complejidad del sistema y ralentiza su funcionamiento, ya que, aunque estas tareas se realizan de forma automática, las consultas a los recursos adicionales necesariamente incrementarán el tiempo de operación y necesidades de procesamiento; sin embargo, las ganancias en solidez del sistema nos parecen suficientes para justificar estas pérdidas.

b) Tokenización y separación de frases

Partimos de una reseña del álbum Take Care de Drake, extraída de la *web* Discogs[14], y referida a la compra de un doble LP en vinilo:

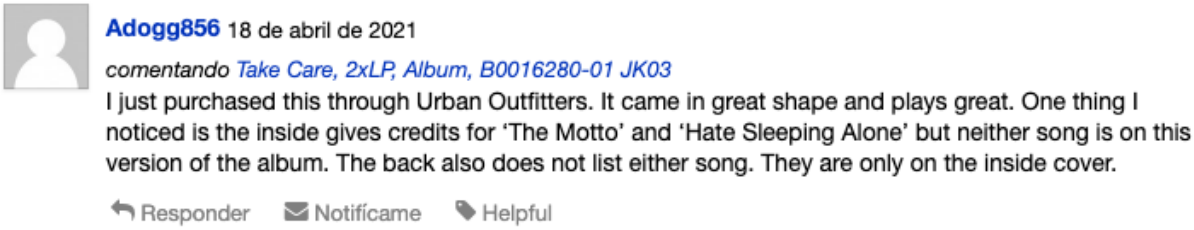


Ilustración 4. Ejemplo de reseña de Discogs

Las etapas anteriores del sistema se encargarían de detectar que se trata de un texto en inglés y seleccionar solo el comentario en sí, extrayendo el contenido de la categoría de html correspondiente, de forma que no se procese el texto contenido como parte de otras etiquetas, como el tema, el nombre, la fecha o las opciones ofrecidas bajo el texto. En este caso, el texto no contiene enlaces, emoticonos u otras cuestiones a resolver, por lo que en la fase de normalización no se realiza una gran cantidad de procesado.

Al entrar el texto normalizado en Stanza, como se explicó en su descripción, la herramienta reconoce entre los caracteres los *tokens* existentes y separa el texto en cinco oraciones:

[14] https://www.discogs.com/

1. I just purchased this through Urban Outfitters.
2. It came in great shape and plays great.
3. One thing I noticed is the inside gives credits for ‘The Motto’ and ‘Hate Sleeping Alone’ but neither song is on this version of the album.
4. The back also does not list either song.
5. They are only on the inside cover.

A partir de esta selección, realizará el resto de los procesos.

c) Detección de POS (*parts-of-speech*)

En esta fase, cada palabra se acompaña por el indicador de la función que desempeña, por ejemplo, en la primera oración (“I just purchased this through Urban Outfitters”) los marcadores de *universal parts-of-speech* (UPOS) son:

word: I	upos: PRON
word: just	upos: ADV
word: purchased	upos: VERB
word: this	upos: PRON
word: through	upos: ADP
word: Urban	upos: PROPN
word: Outfitters	upos: PROPN
word: .	upos: PUNCT

d) Lematización

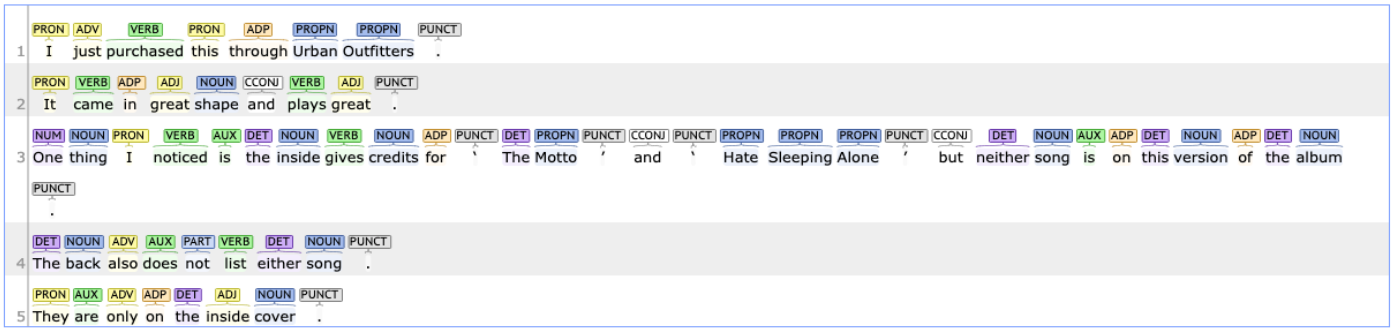
Para la tercera oración (“One thing I noticed is the inside gives credits for ‘The Motto’ and ‘Hate Sleeping Alone’ but neither song is on this version of the album”), se obtendrían los siguientes cambios:

> One thing I notice be the inside give credit for ‘The Motto’ and ‘Hate Sleeping Alone’ but neither song be on this version of the album.

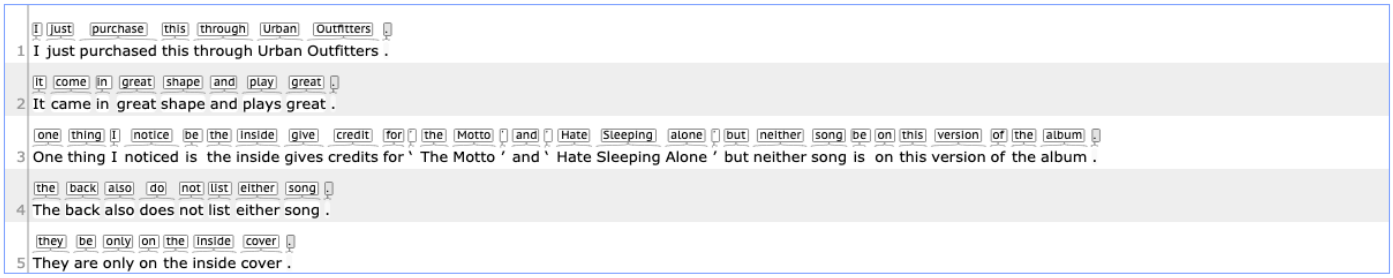
Los verbos pasan a su forma base (*noticed* > *notice, is* > *be, gives* > *give*), así como uno de los sustantivos (*credits* > *credit*). Podemos ver en la Ilustración 5 una muestra de las anotaciones que realiza Stanza a través de su demo interactiva implementada en formato *web*[15] (Qi *et al*., 2020):

[15] Disponible públicamente en http://stanza.run/

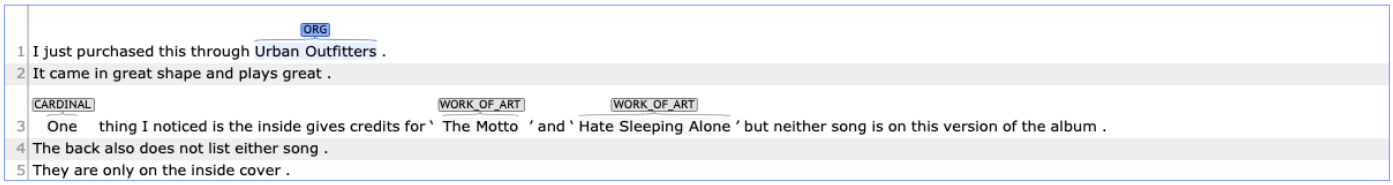
Universal Part-of-Speech:



Lemmas:



Named Entity Recognition:



Universal Dependencies:

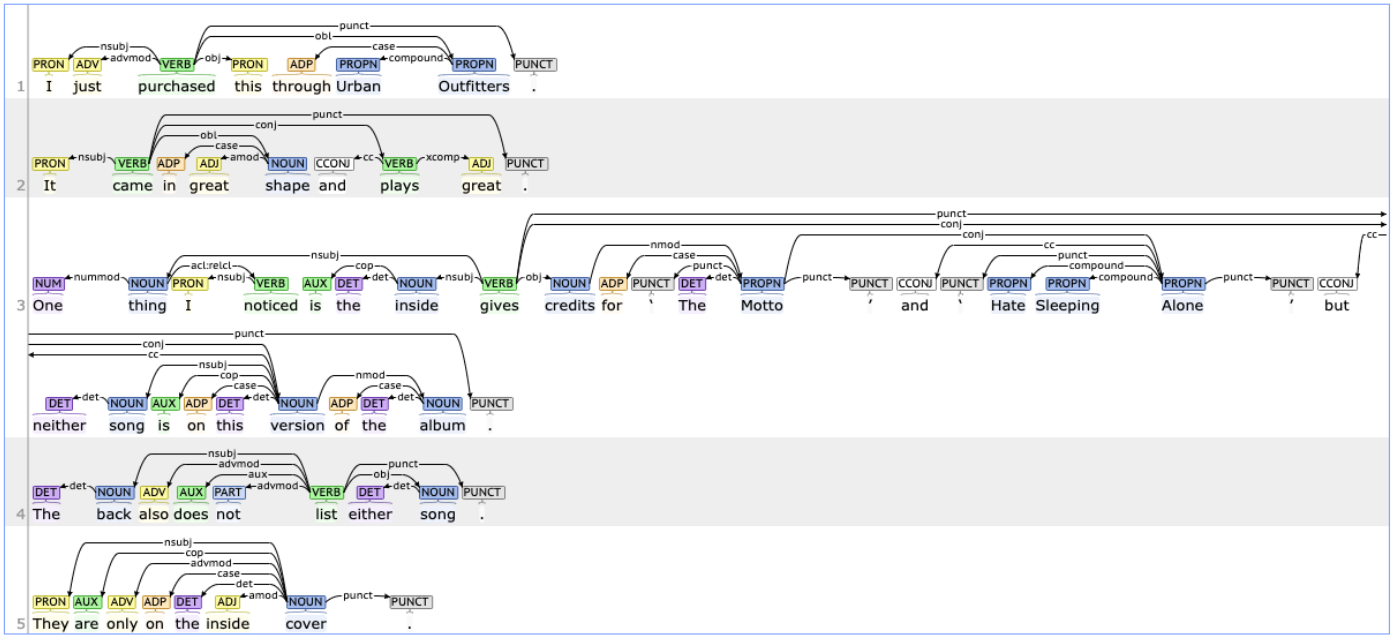


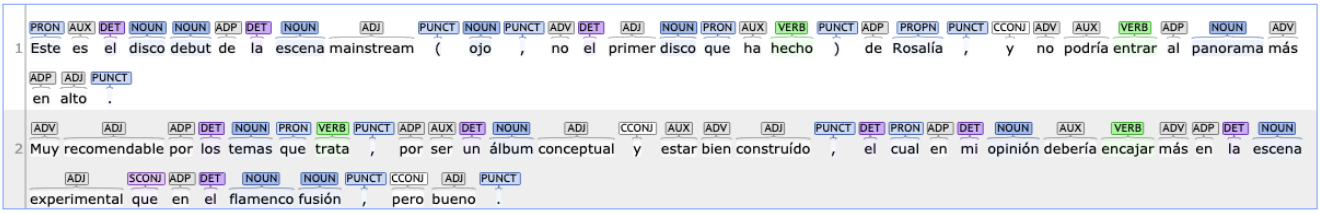
Ilustración 5: Anotaciones de Stanza en inglés

Este módulo opera de manera semejante con textos en español, ya que tiene un soporte multilingüe nativo en el que se incluye esta lengua, como ya anotamos. A modo de ejemplo,

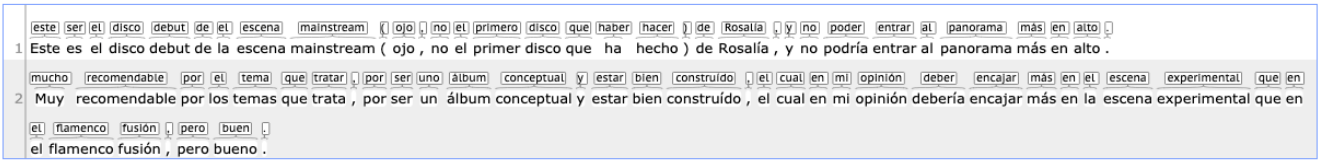
mostramos a continuación el análisis realizado en Stanza para el siguente comentario, también extraído de Discogs:

> Este es el disco debut de la escena mainstream (ojo, no el primer disco que ha hecho) de Rosalía, y no podría entrar al panorama más en alto. Muy recomendable por los temas que trata, por ser un álbum conceptual y estar bien construido, el cual en mi opinión debería encajar más en la escena experimental que en el flamenco fusión, pero bueno.

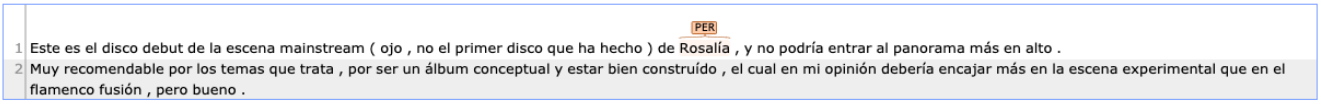
Universal Part-of-Speech:



Lemmas:



Named Entity Recognition:



Universal Dependencies:

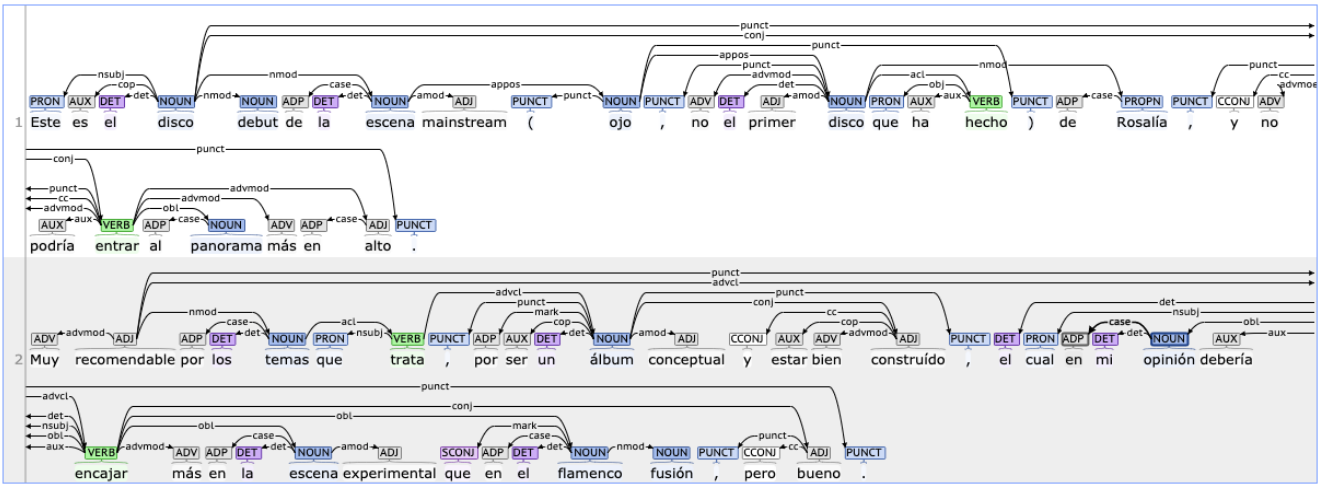


Ilustración 6: Anotaciones de Stanza en español

De este modo, observamos que no existen diferencias notables en el funcionamiento de la herramienta en ambas lenguas, por lo que pasaríamos al siguiente módulo con textos procesados de manera equivalente independientemente de la lengua detectada en el módulo anterior. El paso de normalización busca limpiar el texto para evitar problemas en su procesamiento posterior, la tokenización y separación de frases se encargan de establecer las partes del texto que seguidamente se clasifican según su parte del discurso y morfología y, finalmente, se llevan a su forma base en la etapa de lematización.

4.3 Módulo de identificación de aspectos

El objetivo de este módulo es la selección de los aspectos o características (*features*) referidas en las opiniones de entrada. Para ello se emplea una ontología de dominio, mediante la cual se busca extraer todos los aspectos incluidos en las opiniones expresadas por los usuarios que se vinculen con el dominio[16] en cuestión.

Se señala en Peñalver Martínez *et al.* (2014) que esta ontología de dominio puede ser general, de modo que no es necesariamente un requisito construirla *ad hoc* a partir del corpus textual de las opiniones de usuario (*ibíd.*: 13). Esto facilita la tarea de selección de ontología, permitiendo reutilizar una ya existente.

Si bien en la recopilación de las reseñas probablemente se introduzcan términos de búsqueda que indiquen sobre qué se está opinando (como el disco sobre el que se haya publicado la reseña o la ocurrencia de un término en un *tuit*), la identificación de aspectos permite examinar más profundamente el comentario y realizar un análisis sentimental de grano mucho más fino, como describimos arriba.

Para este trabajo de anotación semántica, empleamos como recurso fundamental la ontología Wikidata (Erxleben, 2014), una base de conocimiento general creada de forma colaborativa, multilingüe y con licencia de dominio público. Esta recoge información estructurada y accesible de manera dinámica, de modo que constituye el repositorio principal de proyectos como Wikipedia. Su carácter libre, multidominio y versátil hace que sea una plataforma en continuo crecimiento, actual y fácilmente accesible, aspectos importantes para un proyecto como el nuestro, centrado en información cultural en continua actualización y expansión.

La estructura de datos de Wikidata se asemeja en cierto modo a la de Wikipedia, algo lógico ya que Wikidata surge como la base de conocimiento (creada y mantenida de forma colectiva) que da soporte a Wikipedia y que, además, funciona como la plataforma central de gestión de datos para tanto Wikipedia como la mayoría de sus proyectos afines (*ibíd*.: 50). La organización de la información parte de las páginas, de modo que todas las cuestiones para las que existan datos estructurados en Wkidata reciben la denominación de *entity* (entidad) y tienen una página asociada. Además, las entidades se subdividen en otros dos tipos, *ítems* (artículos o temas) y *properties* (propiedades) (*ibíd*.: 51-52). Estos términos son bastante similares a los empleados en otras ontologías.

[16] Entendemos dominio como ámbito o campo conceptual.

Así, los *items* son los tipos de entidades que poseen una página en la que se puede consultar su información, de modo que podemos entenderlos como temas o sujetos. Como la información se guarda en diferentes idiomas, cada página tiene asociado un código que la identifica, de forma que cada ítem tiene el mismo independientemente de la lengua en que se consulten los datos. Por ejemplo, si consultamos la página relativa al ítem correspondiente al álbum *Take Care* mencionado en el apartado anterior, encontramos que su identificador es el Q5104794 (Ilustración 6).

Además del identificador, los primeros datos visibles son los denominados *terms* (*ibíd.*), que se corresponden con la etiqueta o nombre del ítem, una breve descripción y una serie de alias. Estos se presentan de manera diferenciada en cada idioma. Debajo de estos encontramos una serie de *statements*, que listan propiedades asociadas al ítem.

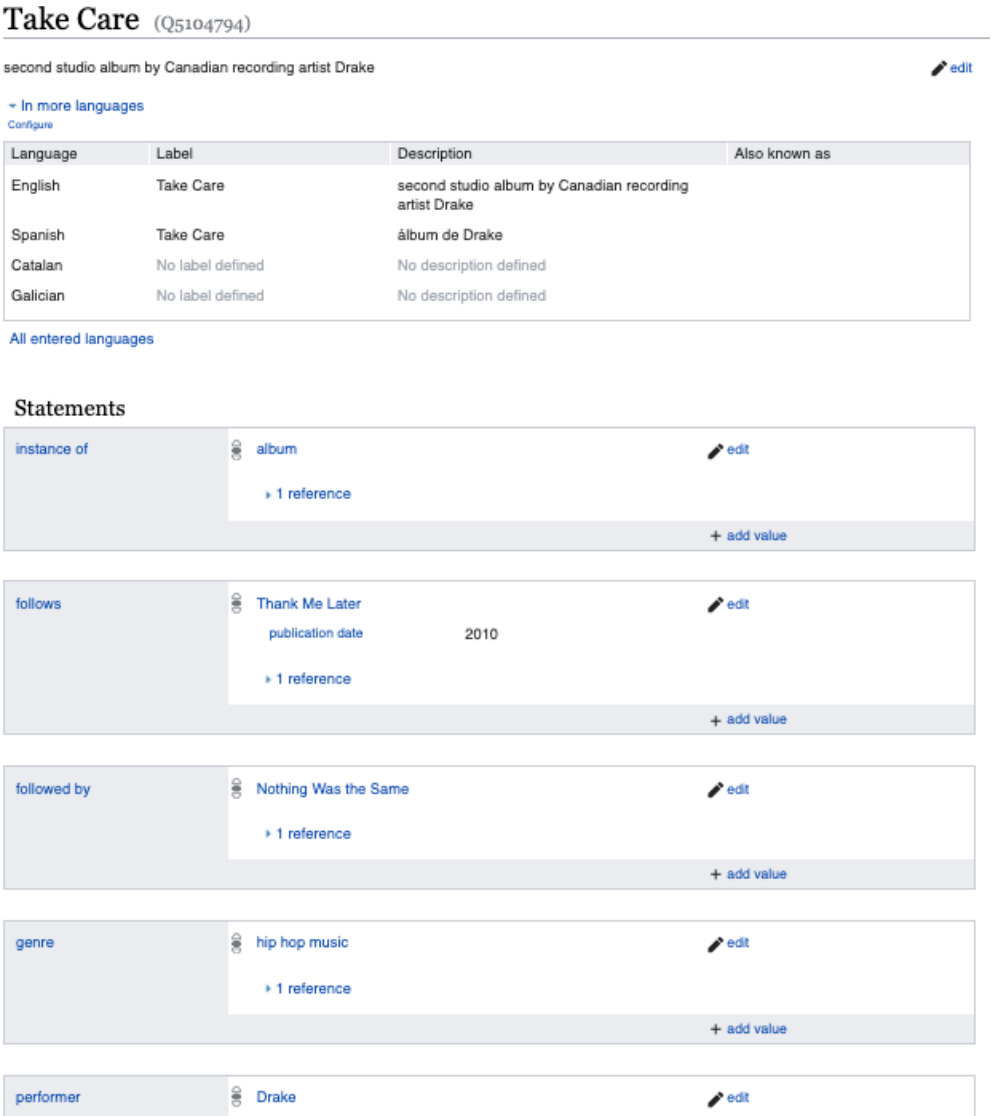


Ilustración 7: Entrada de Wikidata Q5104794

Cada *statement* tiene tres valores: el sujeto (en este caso *Take Care* o Q5104794), la propiedad (como *instance of* que a su vez tiene un identificador, P31, que se puede consultar pinchando en el hipervínculo) y un valor (*album*). Las propiedades, a su vez, tienen un tipo de dato asociado (por ejemplo, *publication date* tiene un *datatype* que es *point in time*), que determina los valores aceptados por ellas. Además, los *statements* pueden ser complementados mediante *qualifiers* (que añaden información adicional necesaria como contexto) y *references* (que corroboran lo alegado).

Podemos visualizar la estructura de la ontología en lo relativo a las diferentes propiedades mediante una herramienta *web* que genera una representación gráfica de manera automática[17]:

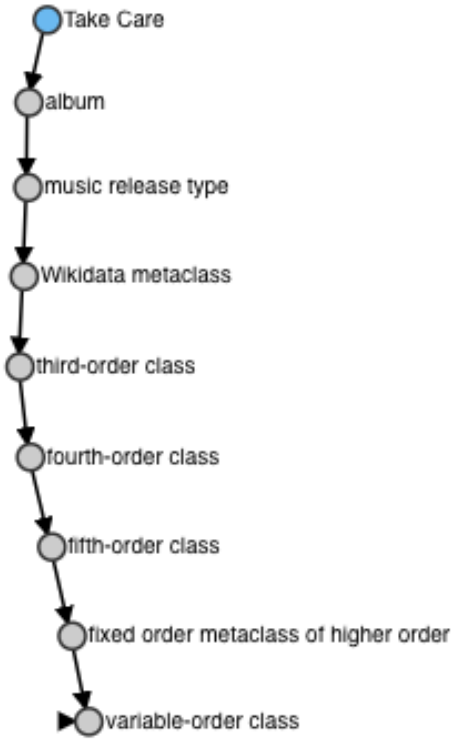


Ilustración 8: Estructura ontológica Wikidata, instance of

Si bien la estructura observada en la Ilustración 7 es sencilla, la arquitectura de Wikidata puede establecer redes de gran complejidad. Como referencia, podemos observar en la Ilustración 8 la organización de *music producer* dentro de la ontología atendiendo a la propiedad *subclass of*:

[17] A través del Wikidata Graph Builder: https://angryloki.github.io/wikidata-graph-builder/

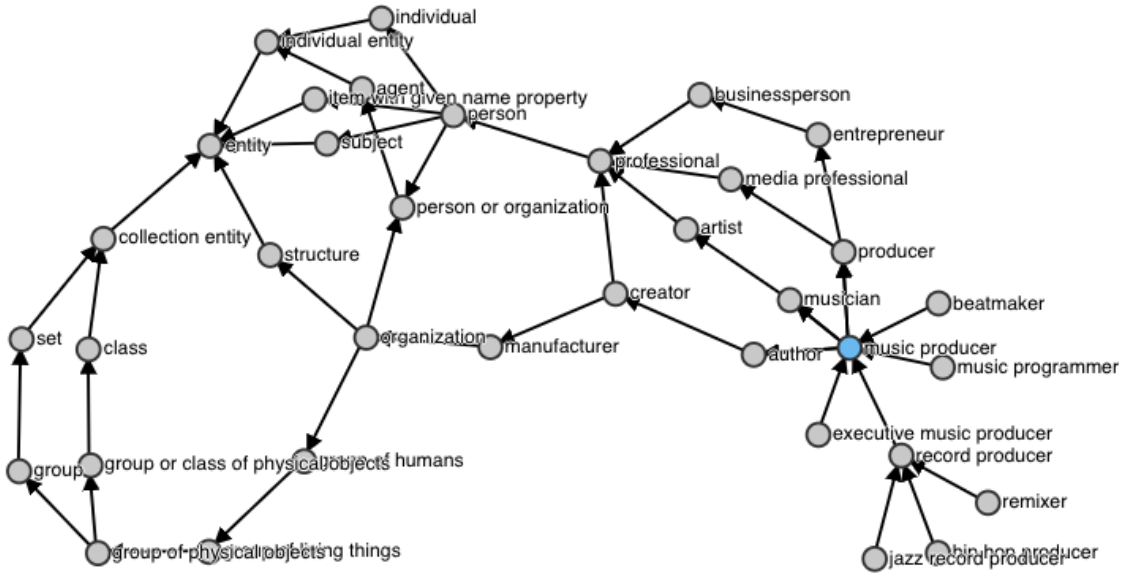


Ilustración 9: Estructura ontológica Wikidata, subclass of

Además de su arquitectura, el hecho de que Wikidata cuente con datos actualizados, es de gran utilidad en nuestra aplicación, puesto que el mundo musical está en continuo cambio y necesitaremos poder analizar opiniones relativas a los acontecimientos más recientes (como novedades discográficas o artistas emergentes). De manera complementaria, podremos contar con los datos extraídos de la API de Spotify, que permite el acceso y la utilización del amplio catálogo de obras y metadata asociado (como artistas, álbumes, canciones, fechas, géneros, etc.) de la plataforma musical. Dado que las entradas de Wikidata relativas a artistas, álbumes o sencillos incluyen una propiedad que las vincula a esta plataforma (como *Spotify album ID*, P2205), existe la posibilidad de utilizar este enlace para recabar información los casos en que no existan en Wikidata datos concretos que pudieran ser de utilidad (como, por ejemplo, referencias a canciones concretas de un álbum).

La decisión de utilizar en el sistema una ontología como Wikidata, que se encuentra en fases avanzadas de su desarrollo y vinculada a múltiples otras plataformas como hemos visto, permite aportar versatilidad al sistema. La mayoría de los sistemas revisados en la bibliografía desarrollan ontologías *ad hoc*, construidas mediante los datos aportados por el corpus de opiniones, utilizando métodos como los mencionados arriba en el estudio de Zhou y Chaovalit (2008), o basándose en ontologías preexistentes como referencia como en Salas Zárate, Valencia García *et al.* (2017). Si bien requieren una fase extra de desarrollo y no tienen la flexibilidad de Wikidata en lo relativo a su actualización y vinculación con plataformas externas, su estructura está planteada desde el inicio para ser integrada en el

sistema de AS, por lo que su contenido está optimizado para la tarea. Mientras, Wikidata abarca conocimiento genérico y no exclusivamente de nuestro dominio, por lo que es mucho más extensa y contiene una cantidad de ruido[18] mayor.

Dado que diseñamos el sistema para atender al dominio de la música, es importante establecer reglas condicionales en la consulta de la ontología para que recupere aquellas entidades relacionadas con este (y que no confunda, por ejemplo, género musical, Q188451, con género literario, Q223393); es decir, para que el uso de la ontología quede limitado al ámbito elegido. De este modo, contaremos con un recurso ontológico específico sin la necesidad de desarrollar una compleja ontología del dominio. Si bien este tipo de restricciones deben determinarse de forma natural, el trabajo de desarrollo conceptual de la ontología ya está realizado. Puesto que la estructuración del conocimiento puede realizarse de muchas maneras, la fase de conceptualización es crítica en el desarrollo ontológico, por lo que la reutilización de ontologías ya implementadas y testadas parece una estrategia mucho más eficiente que su desarrollo de cero (Noy y McGuinness, 2001).

En cuanto al procedimiento de identificación de aspectos, el proceso a seguir es el siguiente:

> This module receives both a corpus of opinions and the domain ontology as input. The sentences in the corpus that contain the classes, individuals, datatype and object properties of the domain ontology are then identified from this input. Once the features have been recognized, they are grouped in accordance with their semantic distance and are then attached to a main concept of the ontology. (Peñalver Martínez *et al.*, 2014: 13)

Por tanto, la labor de identificación de características propiamente dicha se basa en la selección de aquellos términos a los que se alude en los textos de entrada, o lo que es lo mismo: la determinación de los conceptos o entidades sobre los que se está opinando. Cuando estos aspectos o características[19] se han seleccionado, se agrupan y vinculan a los conceptos principales de la ontología según su distancia semántica, asociándose entre sí según las diferentes propiedades definidas en la Wikidata (como las mostradas arriba, *instance of* o *subclass of*). Así, se establece una relación entre los aspectos presentes en el texto utilizando las propiedades expresadas en la ontología.

Tomemos como ejemplo la siguiente reseña:

> "Even as a grandfather, I'm aware of Taylor Swift and her music. I've always thought she had a pleasant voice and her albums were certainly pop radio ready. I didn't listen much

[18] Entendemos aquí ruido como información no relevante para nuestro dominio/aplicación.
[19] Utilizamos ambos términos de manera equivalente en este contexto.

> until her 2014 album '1989'. I bought my first Taylor Swift album. I liked the album but I liked the cover version by Ryan Adams much more."

Se identificarían en este texto aspectos como 'Taylor Swift', 'music', 'voice', 'album', 'pop radio', '1989', 'cover version' y 'Ryan Adams'. De esta manera, 'Taylor Swift' y 'Brian Adams' se asociarían a 'singer' (por la propiedad ontológica *occupation*). Por su parte, 'voice' se asocia a 'singer' (propiedad *used by*). '1989' se reconoce como instancia de 'album', y a su vez con 'Taylor Swift' (como *performer*). 'Album' y 'cover version' ambos son instancias de *musical term*, que se relaciona con 'music' (propiedad *facet of*), y así sucesivamente.

Por lo tanto, vemos que la arquitectura de la ontología permite que las entradas se organicen jerárquicamente y el sistema las relacione entre sí.

Una vez el sistema reconoce y relaciona todas las características de una opinión, creemos que es importante también considerar la relevancia relativa de cada una de ellas. Como se advierte en Peñalver-Martínez *et al*. (2011, 2014) los primeros trabajos de identificación de características asignan la misma importancia a todas, lo cual es un error, pues su peso relativo dentro de la opinión varía en función del número de repeticiones en los comentarios, más importante cuantas más veces aparezca; o su posición a lo largo del texto, pues generalmente los inicios y, sobre todo, los finales de un texto sintetizan más claramente la posición del autor. Para valorar estos aspectos, empleamos aquí también la fórmula propuesta por Peñalver Martínez *et al*. (2014: 15-16):

$$score(f, userop_i) = z_1 * |O_1| + z_2 * |O_2| + z_3 * |O_3|$$

Esta permite asignar una valoración numérica a una característica específica (*f*) dentro de una opinión concreta ($useorp_i$), valorando su posición relativa en el texto (z_i, según se sitúe al principio, en el medio o al final de este) y el número de repeticiones de la característica en cada una de las partes (O_j).

Siguiendo el ejemplo de Peñalver *et al*. (*ibíd.*), si asignamos 80/50/100 respectivamente a los valores de cada posición, en la opinión de un usuario sobre la característica "*song*" ("película" en el original), que aparece un total de cuatro veces a lo largo del texto (distribuidas de la forma 1/1/2), hallamos que su puntuación es:

$$score(song, usuario_x) = 80 * 1 + 50 * 1 + 100 * 2 = 330$$

Este baremo, no obstante, tiene sentido solo para las opiniones más extensas. En comentarios propios de redes de *microblogging* como Twitter, es difícil estructurar el texto en partes

diferenciadas, por lo que establecemos los 280 caracteres de Twitter como el umbral sobre el que empezaremos a contemplar esta separación en partes.

Como nota adicional, cabe señalar que la fórmula se utiliza de la misma manera para los textos españoles y los ingleses. Puesto que la ontología utilizada es la misma, puede esperarse un funcionamiento equivalente en ambos idiomas.

Por último, cabe indicar que esta puntuación tan solo muestra el peso de esta característica. En el módulo que revisamos a continuación obtenemos el siguiente componente necesario para realizar nuestro análisis de opinión.

4.4 Módulo de identificación de polaridad

Aquí se busca calcular la polaridad de las características encontradas previamente. Para ello empleamos la técnica desarrollada en Peñalver-Martínez *et al*. (2014), también utilizamos los recursos de SentiWordNet 3.0 (Baccianella *et al*., 2010) para el inglés y CRiSOL (Molina González, Martínez Cámara y Martín Valdivia, 2015) para el español (que adapta al español los contenidos de SentiWordNet), con la ayuda de la herramienta Babelfy (Moro, Cecconi, y Navigli, 2014) para realizar un proceso de desambiguación.

SentiWordNet (SWN) se define como un recurso léxico diseñado para dar soporte a tareas de clasificación sentimental o aplicaciones de minería de opinión (Baccianella *et al*., 2010: 2200). Para ello utiliza los conceptos (*synsets*, conjuntos de sinónimos) del lexicón WordNet y les asigna valores distintos de polaridad (positiva, negativa y neutra), con los que podemos construir vectores, que nos permiten operar en la etapa posterior para extraer los resultados globales de análisis sentimental. Como explicamos más abajo, utilizaremos también una herramienta de desambiguación para facilitar la evaluación, dado que:

> Each synset *s* is associated to three numerical scores *Pos(s)*, *Neg(s)*, and *Obj(s)* which indicate how positive, negative, and 'objective' (i.e., neutral) the terms contained in the synset are. Different senses of the same term may thus have different opinion-related properties (*ibíd*.: 2200)

Veamos unas muestras de los datos que podemos encontrar en la herramienta, partiendo de la oración *"Feels So Good is my least favorite song of this album. It is very boring"*. Dada esta opinión, hallamos en SentiWordNet (SWN) los siguientes valores para *favorite*:

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
n	05790758	0.25	0	favourite#3 favorite#1	something regarded with special favor or liking; "that book is one of my favorites"

Cada una de las columnas expresa, en orden: su parte del discurso (*POS*), el código identificativo del *synset* dentro de SWN (*ID*), sus valores de positividad y negatividad

(*PosScore* y *NegScore*), los términos sinónimos recogidos por este *synset* (*SynsetTerms*) y una pequeña glosa que muestra su significado y ejemplos de su uso (*Gloss*).

De este modo, se formaría para *favorite* el vector (0.25, 0, 0.75), considerando que el valor neutro se obtiene mediante la fórmula *ValorNeutro = 1 – ValorPositivo – ValorNegativo*.

Buscando además los valores de *boring* vemos lo siguiente:

> a 01345307 0 0.25 wearisome#1 tiresome#1 tedious#1 slow#5 irksome#1 ho-hum#1 dull#4 deadening#1 boring#1 so lacking in interest as to cause mental weariness; "a boring evening with uninteresting people"; "the deadening effect of some routine tasks"; "a dull play"; "his competent but dull performance"; "a ho-hum speaker who couldn't capture their attention"; "what an irksome task the writing of long letters is"- Edmund Burke; "tedious days on the train"; "the tiresome chirping of a cricket"- Mark Twain; "other people's dreams are dreadfully wearisome"

Si bien este *synset* abarca más términos sinónimos y ejemplos de uso, su estructura es la misma. Aquí el vector obtenido sería *boring* (0, 0.25, 0.75). Con estos vectores, como explicamos más abajo, es posible realizar un cálculo de la polaridad de cada aspecto.

Para el funcionamiento del sistema en el ámbito español, implementamos la herramienta léxica CRiSOL (Molina González, Martínez Cámara y Martín Valdivia, 2015). El punto de partida en el desarrollo de esta es el lexicón de palabras de opinión en español SOL (Spanish Opinion Lexicon) (Martínez Cámara *et al*., 2013), una adaptación del lexicón en inglés BLEL (Bing Liu English Lexicon) (Hu y Liu, 2004), que contiene palabras de opinión, positivas y negativas, incluyendo algunas de ellas con erratas habituales en la red (un total de 6.789 términos). SOL se obtuvo mediante la traducción automática al español de BLEL. Por esta razón, se realizó una segunda fase de adaptación manual, mediante la que se obtuvo iSOL (improved Spanish Opinion Lexicon), un lexicón equivalente, pero que se adecuaba a las características de la lengua española que no se tomaban en cuenta en el proceso de traducción automática:

> Por un lado, debido a la inflexión morfológica española, se tiene que mientras un adjetivo inglés, por lo general, no posee ni género ni número, y es representado por un solo término, al adjetivo español le corresponde hasta cuatro posibles palabras traducidas del inglés, dos para el género (masculino o femenino) y dos para el número (singular o plural). Por otra parte, siguiendo la filosofía de Bing Liu se introdujo en las listas algunas palabras mal escritas o inexistentes en el Diccionario de la Real Academia Española (DRAE) ya que aparecen con mucha frecuencia en el contenido de los medios de comunicación social (Molina González, Martínez Cámara y Martín Valdivia, 2015: 145)

Gracias a esta adaptación, iSOL cuenta con un total de 8.135 términos. Esta expansión viene determinada por la adaptación manual a la morfología española. De modo que un adjetivo

como *boring*, puede ser traducido como *aburrido*, *aburrida*, *aburridos* o *aburridas*, en función del contexto lingüístico en que aparezca.

Mediante la combinación de iSOL y los datos de SWN, vinculados a partir del Multilingual Central Repository (MCR) (*ibíd*.) se creó CRiSOL:

> El proceder habitual en el uso de una base de conocimiento léxica basada en la estructura de WordNet, como es el caso de MCR y de SentiWordNet, se corresponde con el uso del identificador de los conceptos (ILI) para recuperar la información asociada al concepto. En este caso no se cuenta con ILIs, sino con formas lingüísticas de una lista de palabras de opinión. MCR asocia a cada lema un ILI, lo cual identifica inequívocamente uno de los posibles conceptos del lema. Tomando ese ILI, ya sí es posible acudir a SentiWordNet y obtener las puntuaciones de polaridad asociadas a dicho concepto. Por tanto, el proceso de generación de CRiSOL comenzó con la obtención de los lemas de las palabras de iSOL. Una vez obtenidos los lemas, el siguiente paso fue encontrar el ILI asociado al lema en MCR. (*ibíd*.: 146)

De este modo, se pueden vincular los lemas y posteriormente asignarles los valores de polaridad equivalentes en SWN. El resultante es CRiSOL, que, si bien es bastante más limitado en extensión que SWN, al haberse desarrollado con especial atención a las palabras de opinión, es una herramienta adecuada para nuestra tarea.

Para la palabra *precioso*, CRiSOL muestra lo siguiente:

```
<is:term xmlns:is="http://sinai.ujaen.es/isol_polar">
      <is:id>ID-1864</is:id>
      <is:text>precioso</is:text>
      <is:polarity xmlns:is="http://sinai.ujaen.es/isol_polar" source="SINAI">
         <is:label>positive</is:label>
      </is:polarity>
      <is:polarity     xmlns:is="http://sinai.ujaen.es/isol_polar"     source="SentiWordNet"
pos="a">
         <is:positive>0.625</is:positive>
         <is:objective>0.375</is:objective>
         <is:negative>0.</is:negative>
      </is:polarity>
   </is:term>
```

El lenguaje de marcado (en este caso XML) es distinto al empleado en SWN, pero vemos que se indican, en este caso de manera directa, los tres valores de polaridad, por lo que se construiría el vector *precioso* (0.625, 0, 0.375).

Dado que el tamaño de CRiSOL es mucho menor que el de SWN (8.135 frente a 117.000), muchos términos españoles no tendrán asociados valores de polaridad por lo que la tarea de análisis se complicaría. Aunque el recurso cubre específicamente palabras de opinión y, por tanto, aquellas más importantes para el análisis, nos parece necesario implementar un recurso adicional para cubrir el resto. Peinado y Maestre (2013) desarrollaron una herramienta de Python que vincula los términos de SWN en inglés con los correspondientes en español

(aunque no todos ellos), Sentiwordnet-BC, de modo que podemos utilizar este recurso como capa adicional para las palabras no recogidas en CRiSOL. A diferencia del SWN original, Sentiwordnet-BC[20] muestra también los valores de objetividad/neutralidad como parte de sus entradas, por ejemplo:

pos	word_en	word_sp	positive	negative	objective
n	performance	interpretación	+0.125	-0	0.875

Si bien hemos encontrado diferentes ejemplos de implementación de este recurso, no hay mucha información sobre su desarrollo. Por lo tanto, y ya que no todos los términos en inglés tienen su contraparte en español, asumimos que no se ha obtenido mediante traducción automática, sino que posiblemente se haya recogido mediante la vinculación de los términos utilizando un procedimiento similar al descrito arriba en lo relativo a CRiSOL, utilizando un repositorio central para asociar las entradas automáticamente. Aunque la implementación parece sólida, ante las dudas en su origen decidimos utilizarlo solo en un segundo plano, en los casos en que CRiSOL no sea suficiente.

En Peñalver-Martínez *et al*. (2014), el valor de polaridad para cada palabra se obtenía a partir de la media de los diferentes sentidos encontrados en SWM, es decir, de cada uno de los *synsets* donde se encuentre la palabra; no obstante, esta técnica es poco precisa lingüísticamente y puede ser mejorada fácilmente. Dado que una palabra puede utilizarse en sentidos muy diferentes (incluso opuestos), utilizar la media de los sentidos existentes en el lexicón de polaridad no aporta necesariamente información específica sobre su uso, sino un valor aproximado en el mejor de los casos. Nos parece necesario desambiguar el sentido de la palabra para poder considerar de forma más adecuada su polaridad en el cálculo.

Si bien la desambiguación (*word sense disambiguation*, WSD) es una de las tareas de PLN más complejas, hoy en día existen herramientas que permiten realizar este proceso automáticamente. Por ello, seguimos a Salas Zárate, Medina Moreira *et al.* (2017) en lo relativo a la implementación del proceso de desambiguación en su sistema, utilizando la herramienta Babelfy (Moro, Cecconi, y Navigli, 2014):

> Each entry in SWN has multiple senses; for example, the word "better" has sixteen senses in SWN (four that belong to the category "adjective," seven that belong to the category "noun," two that belong to the category "adverb," and three that belong to the category "verb"). In order to address this issue, we have used Babelfy. (Salas Zárate, Medina Moreira *et al.*, 2017: 5)

[20] Disponible en GitHub: https://github.com/rmaestre/Sentiwordnet-BC

Esta misma herramienta resulta apropiada para resolver esta cuestión tanto en inglés como en español, ya que está desarrollada para un funcionamiento multilingüe en tareas de desambiguación y vinculación de entidades (*entity linking*, EL):

> Babelfy is the first approach which explicitly aims at performing both multi-lingual WSD and EL at the same time. The approach is knowledge-based and exploits semantic relations between word meanings and named entities from BabelNet, a multilingual semantic network which provides lexicalizations and glosses for more than 9 million concepts and named entities in 50 languages. (Moro, Cecconi, y Navigli, 2014: 1)

Su funcionamiento está basado en la red semántica BabelNet (Navigli y Ponzetto, 2012), una gran base de conocimiento multilingüe que se construye a partir de la integración automática de la información de plataformas como Wikipedia, WikiData, OmegaWiki, WordNet, Open Multilingual WordNet y Wiktionary (Moro, Cecconi, y Navigli, 2014: 2). De este modo, se puede realizar el proceso de desambiguación vinculando los términos siguiendo el marco de WordNet, gracias a sus complejos procesos heurísticos mediante los que se desarrolla la interpretación semántica de un texto que permite elegir el mejor de los sentidos para el fragmento entre los candidatos existentes:

> Our state-of-the-art approach, Babelfy, is based on a loose identification of candidate meanings (substring matching instead of exact matching) coupled with a densest subgraph heuristic which selects high-coherence semantic interpretations. Here we briefly describe its three main steps:
> 1. Each vertex, i.e., either concept or named entity, is automatically associated with a semantic signature, that is, a set of related vertices by means of random walks with restart on the BabelNet network.
> 2. Then, given an input text, all the linkable fragments, i.e., pieces of text being equal to or substring of at least one lexicalization contained in BabelNet, are selected and, for each of them, the possible meanings are listed according to the semantic network.
> 3. A graph-based semantic interpretation of the whole text is produced by linking the candidate meanings of the selected fragments using the previously computed semantic signatures. Then a densest subgraph heuristic is used to extract the most coherent interpretation and finally the fragments are disambiguated by using a centrality measure within this graph. (*ibíd.*: 2)

Tras la desambiguación, se obtiene un sentido único para la palabra que permitirá hallar sus valores correspondientes de SWN que, en el ejemplo previo de Salas Zárate, Medina Moreira *et al.* (2017) respecto a '*better*', produciría lo siguiente:

> The word "better" is identified as "(comparative of "good") superior to another (of the same class or set or kind) in excellence or quality or desirability or suitability; more highly skilled than another" in Babelfy, which corresponded in SWN to the adjective ID "00230335", with the following positive, negative, and neutral scores: ScorePosSwn = 0.875, ScoreNegSwn = 0, and ScoreNeuSwn = 0.125. (*ibíd.*: 5)

Para identificar la polaridad de las características reconocidas en el módulo anterior, seguimos la formula propuesta en Peñalver-Martínez *et al*. (2014), basada en un sencillo cálculo vectorial:

Vector(f , $userop_x$)= score(f , $userop_x$)(ScorePos, ScoreNeg, ScoreNeu)*

El vector 'V', representa el valor obtenido para la polaridad del aspecto *f* (feature), dentro del texto (identificado como $userop_x$). Para su obtención se utiliza además el resultado del módulo anterior (el peso de la característica o *score*) y un vector de polaridad que se corresponde a su vez con la media de los vectores de polaridad de las palabras en el entorno próximo de este aspecto (pero no del aspecto en sí). Como mencionamos anteriormente, los métodos de selección de las palabras próximas más efectivos experimentalmente en Peñalver-Martínez *et al*. (2014: 29) y Salas-Zárate, Valencia García *et al*. (2017: 19) son N_GRAM *Before* y N_GRAM *Around*, respectivamente, donde parámetro N_GRAM indica el número de palabras de separación a considerar. Sin embargo, en nuestro caso, el dominio de aplicación es distinto del suyo y, además, vamos a implementar también un segundo idioma, por lo que nos parece más idóneo testar primero la efectividad de todos los métodos:

> - N_GRAM Before: this method obtains the N_GRAM words before the linguistic expression of the feature in the user's opinion.
> - N_GRAM After: this method obtains the N_GRAM words after the linguistic expression of the feature in the user's opinion.
> - N_GRAM Around: this method obtains the N_GRAM words before the linguistic expression of the feature in the user's opinion and the N_GRAM words after the linguistic expression of the feature in the user's opinion.
> - All Phrase: this method obtains all the words in the same sentence as the linguistic expression of the feature in the user's opinion. Peñalver-Martínez *et al*. (2014: 16)

En Salas Zárate, Medina Moreira *et al.* (2017) eligen un valor de N_GRAM de entre 2 y 6, para el AS de opiniones procedentes de Twitter, que parece apropiado debido al carácter fundamentalmente breve de los textos de esta y otras redes sociales.

Por último, también es necesario observar y considerar una serie de 'reglas lingüísticas' como las señaladas en Zhao y Li (2009):

-Ante la presencia de expresiones de negación en el texto (como *no* o *nada*), los valores de positividad y negatividad del vector correspondiente se invertirán. Esto es útil para analizar correctamente expresiones como "el disco no era tan bueno como esperaba", donde se indica una opinión negativa mediante la negación de palabras positivas. La valoración de las negaciones también ha sido estudiada en el contexto del análisis sentimental en lengua española en Jiménez Zafra *et al.* (2019), donde los resultados también apuntaban a una mejora de la precisión del sistema tras el tratamiento de las negaciones. Siguiendo el esquema marcado en este estudio, establecemos una pauta para delimitar el alcance que las expresiones de negación tienen en el texto y, consecuentemente, qué valores

de polaridad afectan (*ibíd.*: 4-5). Así, dado que Stanza realiza previamente un análisis oracional, las expresiones de negación afectarán a las expresiones con las que se relacionen en su árbol de dependencias correspondiente. En el ejemplo anterior, se niega el verbo ser, de modo que la negación afecta al atributo, es decir a 'tan bueno'. Estas técnicas limitan el campo de acción de la regla, impidiendo que tenga consecuencias no deseadas.

-En el caso de encontrar palabras de polaridad desconocida, se tratará de determinar sus valores como equivalentes (u opuestos) a los de sus sinónimos (o antónimos), de acuerdo con la propuesta de Ding y Liu (2007). Para ello se utiliza como referencia el lexicón WordNet (Fellbaum, 1998), que como introdujimos brevemente al hablar de SentiWordNet se organiza mediante relaciones semánticas entre las que destaca la sinonimia.

-Cuando se hallen aun así palabras de polaridad desconocida, esta se podrá asignar como equivalente por defecto a la de otras palabras a las que vayan unidas mediante conjunciones.

-En los experimentos de Elfajr y Sarno (2018) se comprobó que la precisión de su sistema mejoraba al considerar que los emoticonos tienen un mayor peso emocional en las opiniones que las palabras, por lo que decidimos aplicar también su técnica de duplicar el valor de polaridad expresada por los emoticonos y *emojis* detectados.

4.5 Módulo de minería de opinión y clasificación

Finalmente, la polaridad global para una opinión se calcula considerando todos los vectores obtenidos dentro del texto para cada una de las características consideradas, siguiendo el método explicado arriba. Al realizar el sumatorio de todos ellos, se obtiene un nuevo vector de tres coordenadas. La polaridad del documento en su totalidad viene simplemente determinada por la coordenada cuyo valor sea mayor que las demás, siguiendo el procedimiento de Peñalver-Martínez *et al*. (2014). Este es un baremo sencillo pero que utiliza un mecanismo novedoso y cuyos resultados prácticos han sido prometedores (*ibíd.*):

$$Polaridad(userop_x) = \sum_{i=1}^{n} V(f_i, userop_x)$$

El valor vectorial de polaridad de una opinión a nivel documento viene determinada por la suma por coordenadas de los vectores de polaridad baremados (según su peso relativo) de cada uno de los aspectos valorados en la opinión. El resultado de esta suma es, por tanto, un nuevo vector de tres coordenadas. Estos valores, en caso de ser distintos de cero, deben ser necesariamente positivos, ya que no existen valores negativos en el sistema de puntuación

planteado en SWN o en los mecanismos de cálculo planteados en este sistema. De esta manera, al comparar las tres coordenadas entre sí, la que tenga un valor superior marcará el valor de polaridad final del documento.

Propongamos un nuevo ejemplo para visualizar esto más fácilmente:

Partamos de una opinión ficticia, como "me pareció que el álbum tenía una sonoridad muy interesante, pero el cantante es tan malo que no fui capaz de escucharlo entero", donde se han valorado los aspectos "*cantante*" y "*álbum*" cuyos vectores asociados son[21]:

$V_1(cantante, userop_x) = (135, 265, 98)$

$V_2(álbum, userop_x) = (240, 188, 200)$

El vector resultante por la suma de estos dos es:

$Polaridad(userop_x) = (375, 453, 298)$.

Como se observa que *ScoreNeg* > *ScorePos* > *ScoreNeu*, diremos que la opinión del usuario x es negativa.

Puesto que en un estudio de opinión este resultado global es solo parcialmente indicativo de las opiniones de usuarios, nuestro sistema recopilará también los resultados de cada uno de los aspectos valorados en la opinión. De este modo, además de saber cuál es la opinión general sobre algo, se podrán hacer estudios sobre cuáles son los factores que determinan dicho resultado. De este modo, en la opinión que acabamos de revisar, si bien su resultado general es negativo, se valora favorablemente el aspecto *álbum*, a pesar de que el aspecto *cantante* haya sido más determinante en el resultado final.

Si se cuentan con múltiples opiniones relativas a la misma entidad, como reseñas de un disco específico, se puede estudiar, por ejemplo, que aspectos han sido más o menos determinantes en su valoración. Esto hará que la usabilidad del sistema sea mucho mayor, ya que facilitará realizar análisis bastante más detallados. Si bien es imposible valorar de manera manual un gran número de opiniones de forma conjunta, encontramos que un solo valor de polaridad puede no ser suficiente en la toma de decisiones, por lo que, aunque útil, preferimos incluir la opción de mostrar los resultados de los aspectos considerados para añadir un nivel adicional de detalle cuando se considere necesario.

---

[21] Los valores numéricos son también inventados, con el fin de mantener la explicación simple, dada la dificultad de realizar manualmente el proceso completo de cálculo de cada uno de los apartados anteriores.

5. Conclusiones

La minería de opinión es una disciplina de gran actualidad y no exenta de complejidad. Hemos propuesto en este trabajo un sistema que sigue una de las líneas experimentales más prometedoras que buscan dar peso a los aspectos semánticos del problema del AS. El método basado en aspectos y con soporte ontológico nos permite trabajar con Wikidata, que abarca una extensa base de conocimiento sobre el dominio vinculada a plataformas de relevancia para el ámbito musical. Ante el problema de abarcar la gran variedad expresiva existente en la lengua usada online hoy en día, el continuo desarrollo y mantenimiento de los recursos lingüísticos es una de las tareas clave. La elaboración y actualización de una ontología del dominio es una labor costosa, de modo que la posibilidad de utilizar una plataforma robusta y actualizada constantemente gracias a la labor colectiva facilita esta cuestión.

La metodología empleada en nuestra propuesta constituye una línea de trabajo de gran interés e implementada en diversas ocasiones en los últimos tiempos, pero no es la única en el ámbito del AS. Muchas otras investigaciones centradas en técnicas de *machine learning* proponen enfoques interesantes y prometedores que, no obstante, quedan fuera del alcance del presente trabajo, por su complejidad técnica. Nuestro planteamiento busca poner el foco en cuestiones lingüísticas y formular nuevas aproximaciones a partir de las propuestas realizadas hasta el momento. Sin embargo, el tratamiento computacional del lenguaje es una tarea delicada y que todavía no ha sido resuelta de manera ideal. Aspectos como la resolución de problemas derivados de la ambigüedad en el lenguaje todavía no se han perfeccionado, algo incluso más relevante en las opiniones online, donde el estilo está menos marcado que en los medios tradicionales y abundan el uso del humor y el sarcasmo, expresiones agramaticales, faltas de ortografía, abreviaciones, *emojis* o imágenes y *memes*, que sin duda definen el lenguaje del medio y su manera de comunicar opiniones. Si bien hemos buscado establecer técnicas para el tratamiento de algunos de estos fenómenos, como las erratas, abreviaturas o emoticonos, gran parte de estas cuestiones deberán afrontarse en futuras investigaciones para perfeccionar los sistemas de minería de opinión.

En cualquier caso, nos parece necesario discutir una serie de puntos relativos al tratamiento del lenguaje en nuestro sistema:

-La solución propuesta busca permitir una funcionalidad multilingüe (en inglés y español), en la que el procesamiento de cada lengua es independiente. Sin embargo, por este medio no

es posible analizar correctamente textos mixtos o en *spanglish*, fenómeno con relativa frecuencia en las redes sociales.

-En la implantación de reglas lingüísticas para una mejor valoración de la polaridad de un texto, la orientación principal es establecer guías que permitan mejorar el procesamiento del texto y la eficacia en el cálculo, sin complicar los requisitos de procesamiento del sistema de manera desproporcionada. Precisamente por ello, el esquema de tratamiento lingüístico de este tipo de sistemas está muy lejos de verdaderamente entender las opiniones, por lo que estas reglas no están exentas de defectos. Este es un terreno en el que trabajos futuros podrían incidir para buscar una adecuación del sistema al verdadero uso del lenguaje, más allá de reglas básicas para el cálculo.

-Por el tipo de contenido de cada plataforma, no todas las opiniones podrán ser garantizadas con las mejores garantías. En el análisis de opiniones de carácter más breve o informal, como es habitual en redes sociales como Twitter, es probable que el sistema se encuentre con comentarios que no pueda procesar correctamente, ya que en estas plataformas abundan las conversaciones casuales, con un gran peso de la ironía o referencias humorísticas, que, como sugerimos, son algunos de los aspectos sin resolver del PLN. De este modo, sería interesante examinar en la práctica qué plataformas son las que tendrían una mayor eficacia en el análisis. Nuestro enfoque abierto a diferentes tipos de implementación busca precisamente permitir este tipo de comparativas, pues las opiniones en internet tienen caracteres muy variados, por lo que deberá estudiarse de qué manera analizar cada tipo.

De cara a la implementación de soluciones para estos problemas, sería interesante cuantificar su ocurrencia relativa, para poder priorizar cuáles serían las medidas que más influencia tendrían en la precisión del sistema.

Como se ha apuntado en diferentes momentos, el desarrollo de recursos es una etapa clave para el desarrollo de un sistema de MO de este tipo, especialmente en lo relativo a la ontología del dominio y el lexicón de polaridad. Por ello, cabe dedicar un tiempo a valorar las decisiones tomadas en nuestro diseño.

En primer lugar, hemos dado suficientes razones para el uso de Wikidata como ontología en lugar de una desarrollada ad hoc o la reutilización de una ya existente como The Music Ontology[22]. Si bien esta segunda opción, como ya advertimos, añadiría concreción y

[22] Disponible en http://musicontology.com/

especificidad, no sería tan versátil como la alternativa propuesta. En el caso de desarrollar una ontología a partir de un corpus de opiniones limitado, limitaría la operación del sistema a dicho corpus, mientras que buscamos, idealmente, poder aplicarlo a cualquier tipo de opiniones recogidas de internet. The Music Ontology, por su parte, si bien parece bien desarrollada y documentada, no tiene el nivel de actualización de Wikidata[23] ni da soporte a lenguas diferentes de la inglesa. Nuestra apuesta por Wikidata en el diseño busca expandir el tipo de operaciones realizadas con este sistema, abarcando un campo conceptual más amplio y al día.

En segundo lugar, nuestra elección de SentiWordNet como lexicón de polaridad viene en gran medida marcada por su posición de estándar en el sector, ya que se ha implementado en múltiples sistemas de orientaciones muy distintas, de modo que ha pasado a convertirse el modelo a seguir en cuanto a la determinación de polaridades. Una elección más complicada fue la del tipo de recursos a implementar para el funcionamiento en español. Dado que uno de los criterios marcados en el proyecto es el tratamiento específico de cada lengua, hemos buscado aplicar recursos desarrollados de manera dedicada. Así, frente a la opción de implementar un único sistema que traduzca automáticamente los términos de SentiWordNet, que en nuestro criterio no iría en la línea de nuestro planteamiento, hemos elegido CRiSOL como recurso básico para el tratamiento de términos sentimentales, con el apoyo de una librería que vincula los términos de SWN en inglés con los del español. Sin embargo, dado que la suma de estos dos recursos no cubre el mismo número de palabras totales que SWN, aunque sí las más habituales o relevantes, la solución no es definitiva.

Ante este problema, cabría probar la implementación de alguna alternativa menos extendida, como ML-Senticon (Cruz *et al*., 2014) o las propuestas de Amores, Arco y Borroto (2016), que presentan bajo los nombres de SentiWordNet 4.0 y SpanishSentiWordNet. Sin embargo, ambos proyectos son mucho más recientes y menos explorados, lejos del reconocimiento de SWN y, además, surgen como una supuesta mejora de SWN. Consideramos que proyectos de este tipo todavía deben evaluarse y madurar más. Además de que, en el caso del segundo, ni siquiera está disponible para su utilización, por lo que la aplicabilidad en el tiempo presente no es posible. El testado de diferentes lexicones de polaridad posteriores a la

[23] De hecho, su repositorio de Github lleva sin actualizarse desde 2012.

implementación del sistema podría traer grandes frutos, pero en esta fase del diseño decidimos elegir una alternativa sencilla y sólida.

Precisamente porque este proyecto se presenta tan solo como la primera fase de diseño y prototipado del sistema, hemos tratado de ofrecer una configuración clara pero que cubra un amplio espectro de usos. Buscamos ofrecer un sistema de AS para el ámbito musical que pueda introducir una serie de análisis no practicados hasta el momento pero que pueden añadir una nueva dimensión al tipo de métricas obtenidas a partir de la distribución de la música en internet. Como diseño inicial, hemos buscado un planteamiento genérico, pero en una hipotética puesta en marcha, podría adaptarse a situaciones reales de uso que modificaran su configuración. En este sentido, nuestras elecciones facilitan una posible expansión del sistema. Tan solo necesitaríamos un lexicón de polaridad plenamente multilingüe para que el sistema pudiera operar en muchos más idiomas. Dada la versatilidad de Wikidata, se podría expandir a otros dominios distintos con relativa facilidad.

En cualquier caso, con la vista en una posible implementación futura que pudiera ser distribuida y utilizada, sería necesario revisar el diseño elegido para verificar que no existieran incompatibilidades entre los recursos y librerías empleadas con el uso comercial. Así, por ejemplo, SWN no permite en su licencia este tipo de uso, por lo que sería una razón extra para el replanteamiento del tipo de sistema de valoración de polaridad. Sin embargo, por el momento, en esta primera fase de diseño, la prioridad es establecer un sistema que construya sobre modelos ya existentes, implementando pequeñas utilidades y mejoras, con el objetivo de establecer nuevas direcciones a las que las investigaciones sobre minería de opinión puedan dirigirse en el futuro, con un énfasis importante en el tratamiento del lenguaje y no solo de las capacidades y técnicas de computación.

En conclusión, este proyecto busca revitalizar las investigaciones en análisis sentimental con base aspectual y guiado por ontologías mediante una implementación multilingüe dedicada y su aplicación al dominio musical. Ambas líneas se encuentran en sintonía con la dirección en que se dirige nuestro mundo: la comunicación es global y existe una continua mezcla de culturas y lenguas, contexto en el cual la música tiene un importante papel, como uno de los motores culturales y artísticos más fuertes. Al mismo tiempo, en una época marcada por los *big data*, podemos afirmar que esta disciplina va a seguir estando vigente y recibiendo atención, pues toda la información disponible en la red tiene poca importancia si no podemos interpretarla y darle un sentido.

6. Bibliografía

Ali, F., Kwak, K.-S. y Kim, Y.-G. (2016). Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification. *Applied Soft Computing, 47*, 235–250. https://doi.org/10.1016/j.asoc.2016.06.003

Amores, M., Arco, L., y Borroto, C. (2016). Unsupervised Opinion Polarity Detection based on New Lexical Resources. *Computación y Sistemas, 20*(2), 263-277. https://doi.org/10.13053/cys-20-2-2318

Anta, A. F., Chiroque, L. N., Morere, P., & Santos, A. (2013). Sentiment analysis and topic detection of Spanish tweets: A comparative study of of NLP techniques. *Procesamiento del lenguaje natural, 50*, 45-52. http://hdl.handle.net/10045/27863

Baccianella, S., Esuli, A., y Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *Lrec, 10*, No. 2010, pp. 2200-2204). http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

Balahur A., y Turchi M. (2012). Multilingual sentiment analysis using machine translation? En *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis (WASSA '12)*. Association for Computational Linguistics, USA, 52–60. https://dl.acm.org/doi/10.5555/2392963.2392976

Balahur, A., y Turchi, M. (2013). Improving sentiment analysis in Twitter using multilingual machine translated data. En *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013* (pp. 49-55). https://www.aclweb.org/anthology/R13-1007

Balahur, A., y Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language, 28*(1), 56-75. https://doi.org/10.1016/j.csl.2013.03.004

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 34-43.

Cambria, E., Schuller, B., Liu, B., Wang, H., Havasi, C. (2013). Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems 28* (2) : 12-14. ScholarBank@NUS Repository. https://doi.org/10.1109/MIS.2013.45

Carpenter, B. (s. f.). *LingPipe: Language Identification Tutorial*. Alias-I. Recuperado 29 de mayo de 2021, de http://www.alias-i.com/lingpipe/demos/tutorial/langid/read-me.html

Carpenter, B. (2005). Scaling high-order character language models to gigabytes. En *Proceedings of Workshop on Software* (pp. 86-99). https://www.aclweb.org/anthology/W05-1107

Cruz, F., Troyano, J., Pontes, B., y Ortega, F. (2014). ML-SentiCon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas. *Procesamiento Del Lenguaje Natural, 53,* 113-120. Recuperado de http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5041

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., & Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation, 8*(4), 757-771. https://doi.org/10.1007/s12559-016-9415-7

Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. En *2008 IEEE 24th international conference on data engineering workshop* (pp. 507-512). IEEE. https://doi.org/10.1109/ICDEW.2008.4498370

Ding, X., y Liu, B. (2007). The utility of linguistic rules in opinion mining. En *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 811-812). https://doi.org/10.1145/1277741.1277921

Duhan, D. F., Johnson, S. D., Wilcox, J. B., y Harrell, G. D. (1997). Influences on consumer use of word-of-mouth recommendation sources. *Journal of the academy of marketing science, 25*(4), 283-295. https://doi.org/10.1177/0092070397254001

East, R., Hammond, K., Lomax, W., y Robinson, H. (2005). *Marketing Review, 5*(2), 145-157. https://doi.org/10.1362/1469347054426186

Elfajr, N. M., y Sarno, R. (2018). Sentiment analysis using weighted emoticons and SentiWordNet for Indonesian language. En *2018 International Seminar on Application for Technology of Information and Communication* (pp. 234-238). IEEE. http://doi.org/10.1109/ISEMANTIC.2018.8549703

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., y Vrandečić, D. (2014). Introducing Wikidata to the linked data web. En *International semantic web conference* (pp. 50-65). Springer, Cham. https://doi.org/10.1007/978-3-319-11964-9_4

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gebesmair, A. (2017). *Global repertoires: Popular music within and beyond the transnational music industry*. Routledge. https://doi.org/10.4324/9781315093543

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition, 5*(2), 199-220.

Hu, M., y Liu, B. (2004). Mining and summarizing customer reviews. En *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Seattle, Washington, USA, 2004; pp. 168-177. https://doi.org/10.1145/1014052.1014073

Jiménez Zafra, S.M., Martín Valdivia, M.T., Martínez Cámara, E., y Ureña López, L.A. (2019). Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter. *IEEE Transactions on Affective Computing, 10*, 129-141. http://doi.org/10.1109/TAFFC.2017.2693968

Kennedy, H. (2012). Perspectives on sentiment analysis. *Journal of Broadcasting & Electronic Media, 56*(4), 435-450. http://dx.doi.org/10.1080/08838151.2012.732141

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of Emojis. *PLOS ONE, 10*(12), e0144296. https://doi.org/10.1371/journal.pone.0144296

Lau, Raymond Y.K.; Lai, Chapmann C.L.; Ma, Jian; y Li, Yuefeng. (2009). Automatic Domain Ontology Extraction for Context-Sensitive Opinion Mining. *ICIS 2009 Proceedings*. *Paper 35*. http://aisel.aisnet.org/icis2009/35

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1-167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, B., y Zhang, L. (2012). A survey of opinion mining and sentiment analysis. En *Mining text data* (pp. 415-463). Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_13

Lo, S. L., Cambria, E., Chiong, R., y Cornforth, D. (2017). Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review, 48*(4), 499-527. https://doi.org/10.1007/s10462-016-9508-4

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., y McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *En Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60). https://doi.org/10.3115/v1/P14-5010

Martínez Cámara, E., Martín Valdivia, M. T., Molina González, M. D., y Lopez, L. A. U. (2013). Bilingual experiments on an opinion comparable corpus. En *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 87-93). https://www.aclweb.org/anthology/W13-1612

Molina González, M. D., Martínez Cámara, E., & Martín Valdivia, M. T. (2015). CRiSOL: Base de conocimiento de opiniones para el español. http://hdl.handle.net/10045/49286

Moro, A., Cecconi, F., y Navigli, R. (2014). Multilingual Word Sense Disambiguation and Entity Linking for Everybody. En *ISWC-P and D 2014 - Proceedings of the ISWC 2014 Posters and Demonstrations Track, a track within the 13th International Semantic Web Conference, ISWC 2014* (Vol. 1272, pp. 25-28). CEUR-WS. https://doi.org/10.18653/v1/S15-2049

Moro, A., Raganato, A., y Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics, 2*, 231-244. https://doi.org/10.1162/tacl_a_00179

Navigli, R. y Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence, 193*, 217-250. https://doi.org/10.1016/j.artint.2012.07.001

Németh, L. (s. f.). Hunspell. Recuperado 29 de mayo de 2021, de https://hunspell.github.io/

Noy, N. F., y McGuinness, D. L. (2001). *Ontology development 101: A guide to creating your first ontology*. Knowledge Systems Laboratory, Stanford University, 2001.

O'Reilly, T. (2009). *What is web 2.0*. O'Reilly Media, Inc. Recuperado de https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

Pang B., Lee L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr. 2, 1–2,* 1–135. https://doi.org/10.1561/1500000011

Peinado, E., y Maestre, R. (2013). Sentiwordnet-BC. Recuperado de https://github.com/rmaestre/Sentiwordnet-BC

Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M., Moreno, V., Fraga, A., Sánchez-Cervantes, J.L. (2014). Feature-Based Opinion Mining through ontologies. *Expert Systems with Applications*. http://dx.doi.org/10.1016/j.eswa.2014.03.022

Peñalver-Martínez I., Valencia-García R., García-Sánchez F. (2011). Ontology-Guided Approach to Feature-Based Opinion Mining. In: Muñoz R., Montoyo A., Métais E. (eds) Natural Language Processing and Information Systems. NLDB 2011. *Lecture Notes in Computer Science, vol 6716*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22327-3_20

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., y Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. En *Association for Computational Linguistics (ACL) System Demonstrations*. https://www.doi.org/10.18653/v1/2020.emnlp-demos.14

Ravi, K. y Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*. http://dx.doi.org/10.1016/j.knosys.2015.06.015

Salas-Zárate, M. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á., y Valencia-García, R. (2017). Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. *Computational and mathematical methods in medicine, 2017,* 5140631. https://doi.org/10.1155/2017/5140631

Salas-Zárate, M. P., Valencia-García, R., Ruiz-Martínez, A. y Colomo-Palacios, R. (2017). Feature-based opinion mining in financial news: An ontology-driven approach. *Journal of Information Science, 43*(4), 458–479. https://doi.org/10.1177/0165551516645528

Schouten, K., y Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering, 28*(3), 813-830. https://doi.org/10.1109/TKDE.2015.2485209

Silva, A., Bastos, R. y Rocha, R. (2018). Sentiment Analysis in Brazilian Portuguese Tweets in the Domain of Calamity: Application of the Summarization Method and Semantic Similarity in Polarized Terms. En *Proceedings of the 10th International Joint Conference on Computational Intelligence - IJCCI*, p. 225-231. https://doi.org/10.5220/0006947802250231

Spotify. (2021). Web API | Spotify for Developers. Recuperado 24 de enero de 2021, de https://developer.spotify.com/documentation/web-api/

Toutanova, K., Klein, D., Manning, C.D., y Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. En *Proceedings of HLT-NAACL 2003*, 252-259. https://doi.org/10.3115/1073445.1073478

Vrandecic, D. y Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Commun. ACM, 57*, 78-85. https://doi.org/10.1145/2629489

Zhao L. & Li C. (2009). Ontology Based Opinion Mining for Movie Reviews. In: Karagiannis D., Jin Z. (eds) Knowledge Science, Engineering and Management. KSEM 2009. *Lecture Notes in Computer Science, vol 5914*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-10488-6_22

Zhou, L. & Chaovalit, P. (2008). Ontology-supported polarity mining. *Journal of the American Society for Information Science and Technology, 59*(1), 98–110. https://doi.org/10.1002/asi.20735