Proyecto fin de carrera

Tratamiento del castellano en las herramientas de análisis de textos y su aplicación en las redes sociales

Autor: Carlos Rodríguez Álvarez
Tutor: José Luis Martínez Fernández

Universidad Carlos III de Madrid

INDICE

1 INTRODUCCIÓN

Este documento es la memoria del proyecto 'Tratamiento del castellano en las herramientas de análisis de textos y su aplicación en las redes sociales' concebido como Proyecto Fin de Carrera por el departamento de Informática de la Universidad Carlos III de Madrid.

El propósito es transmitir al lector de una manera clara y formal, a través de las distintas secciones que componen el documento, la causa y el efecto del desarrollo de dicho proyecto en el ámbito de las distintas herramientas de análisis de texto que trabajan en castellano.

1.1 PLANTEAMIENTO DEL PROBLEMA

Hoy en día son cada vez más los productos dedicados a la extracción de información a partir fuentes masivas de datos, normalmente relacionadas con internet (Redes sociales, blogs, correos, foros, etc.).

Dichos productos habitualmente están centrados en la lengua inglesa y presentan grandes carencias en otros idiomas, entre los que se encuentra el castellano.

Este proyecto surge de dicha necesidad con la intención de evaluar las funcionalidades y el rendimiento de dichos productos al trabajar con el castellano.

1.2 OBJETIVOS

Una vez planteada la problemática actual del tratamiento del castellano por las actuales herramientas de análisis de textos se puede definir de forma global que el objetivo principal del presente proyecto es el de **elaborar una relación de empresas y productos dedicados a la extracción de información a partir de textos y evaluar sus funcionalidades y rendimiento al trabajar con el castellano**.

Este objetivo engloba otras cuestiones más específicas que son:

- Introducir y dar una visión general del concepto de procesamiento natural del lenguaje (NPL: *Natural Language Processing*) y sus aplicaciones para la extracción de información de redes sociales e internet y su utilidad para la empresa.

- Elaborar una relación de empresas y productos dedicados a la extracción de información de textos.

- Evaluar dichas herramientas desde distintas perspectivas funcionales:
  - Modelo de negocio.
  - Funcionalidades y prestaciones que ofrece el producto.
  - Idiomas con los que trabaja.
  - Fuentes de datos con las que opera.
  - Facilidades de integración con sistemas de terceros.
  - Acuerdo de nivel de servicio.
  - Documentación para usuarios.

- Recoger, elaborar y formatear una batería de datos en castellano a partir de contenidos de redes sociales e internet con la que evaluar a nivel práctico de las herramientas anteriores.

- Probar las herramientas desde el punto de vista práctico con la colección de datos del punto anterior y evaluar su rendimiento al trabajar con el castellano.

Por otro lado destacar los objetivos personales de la realización de este Proyecto Fin de Carrera:

- Aprendizaje del lenguaje PHP y familiarizarme con el uso de *web services*.
- Diseño y utilización de una base de datos MySQL.
- Creación de un entorno de trabajo que simule un servidor web mediante la herramienta Wampserver.
- Introducción y aprendizaje de distintos conceptos de procesamiento natural del lenguaje.

1.3 HERRAMIENTAS A ESTUDIAR

Para cubrir estos objetivos las herramientas que se han seleccionado y estudiado para este proyecto son:

- Open Amplify [1]
- Semantria [2]
- Alchemy Api [3]
- Textalytics [4]
- Open Calais [5]
- Zemanta [6]

1.4 ESTRUCTURA DEL TRABAJO

En este apartado se pasa a describir a grandes rasgos la estructura del presente documento. Consta de cinco capítulos principales en los que se describe la información asociada al desarrollo del proyecto. Estos capítulos se complementan con una serie de anexos con información adicional relevante sobre la realización del proyecto.

Los contenidos que presentan cada capítulo se enumeran a continuación:

1. **Introducción**: Se realizará una breve presentación del proyecto explicando tanto el planteamiento del problema como los objetivos a alcanzar.

2. **Estado del arte**: Se comentará la situación actual de las tecnologías relacionadas con el proyecto y las aplicaciones para la empresa.

3. **Empresas y productos**: Consistirá en una relación de las empresas y herramientas seleccionadas para el estudio y la comparación de sus distintas características funcionales.

4. **Pruebas técnicas**: Se presentarán los datos con los que se realizarán las pruebas de las herramientas y su tratamiento y formateo dentro del sistema. Finalmente se evaluará el rendimiento técnico de cada herramienta mostrando los resultados obtenidos.

5. **Conclusiones**: Contendrá un resumen del trabajo realizado y se analizarán las conclusiones obtenidas tras el desarrollo y documentación del proyecto.

2 ESTADO DEL ARTE

En este capítulo se presentará de forma más específica el contexto en el que se encuadra el proyecto, realizando una revisión de las tecnologías, herramientas y trabajos realizados para dar una visión general el marco sobre el que se plantea el proyecto.

2.1 ANÁLISIS DE TEXTOS

2.1.1 Introducción

La cantidad de información almacenada en Internet excede nuestra capacidad para reducir y analizar los datos sin la ayuda de computadores con los que aplicar técnicas de análisis automatizadas. Además estas bases de datos siguen creciendo hoy en día de forma exponencial.

La ventaja que tenemos nosotros sobre la máquina es que intuitivamente somos capaces de aplicar una semántica y una lógica al texto. Esta carencia de la máquina es la que vienen a cubrir las distintas técnicas de *Natural Language Processing* (NLP, Procesamiento natural del lenguaje).

El NLP es una importante disciplina de la inteligencia artificial que se ocupa de la investigación y desarrollo de aplicaciones eficaces computacionalmente para la comunicación entre individuos o entre individuos y máquinas por medio de lenguajes naturales.

Una de las primeras aplicaciones del NLP fue la traducción automática, que surgió a finales de la década de los cuarenta, antes incluso de que apareciese el término "inteligencia artificial". Estos primeros intentos de traducir textos por ordenador a finales de los cuarenta y durante la década de los cincuenta fracasaron debido a la escasa potencia de los ordenadores y a la pobre sofistificación lingüística. Fue a partir de la década de los sesenta cuando se empieza a tener cierto éxito debido a que se comienzan a desarrollar interfaces en lenguaje natural para bases de datos y otras aplicaciones informáticas. Es ya en la década de los ochenta y el principio de la de los noventa cuando se perfecciona el terreno de la Traducción Automática.

Algunas de las dificultades lingüísticas con las que los sistemas NLP pueden encontrarse serían las siguientes:

- Ambigüedad en el lenguaje:

  Por ejemplo, una misma palabra puede tener distintos significados y el correcto depende del contexto de la oración. También una oración a menudo no significa lo que literalmente dice, hay que tener en cuenta en la interpretación del mensaje elementos como la ironía o el sarcasmo.

- Detección de separación entre palabras:

  En la lengua hablada no se suelen hacer pausas entre palabra y palabra y estas deben determinarse de tal manera que mantenga un sentido lógico tanto gramatical como contextual. En la lengua escrita, idiomas como el chino mandarín tampoco tienen separaciones entre las palabras.

- Recepción imperfecta de datos:

  En la lengua hablada podemos encontrarnos con distintos acentos extranjeros, regionalismos o dificultades en la producción del habla. En la escrita aparecen errores de transcripción, expresiones no gramaticales o incluso errores en la lectura del texto por parte del receptor.

Las herramientas de NLP hacen uso de distintas técnicas de aprendizaje automático, estadística, reconocimiento de patrones o detección de dependencias para evitar estas dificultades y alcanzar sus objetivos. Estas técnicas podemos dividirlas en dos claros grupos:

- Mediante el uso de técnicas gramaticales y lingüísticas como análisis sintácticos o morfológicos. Estos métodos suelen obtener mejor precisión pero a costa de una menor exhaustividad y requieren mucho trabajo por parte de un gran número de expertos en lingüística computacional.

- Mediante modelos estadísticos y bolsas o grandes grupos de palabras etiquetadas con la información relativa a cada una. Esta opción requiere una gran cantidad de datos anotados manualmente y un mantenimiento constante.

Algunas de las principales ramas de aplicación de los sistemas NLP serían:

- Síntesis del discurso
- Análisis del lenguaje
- Reconocimiento del habla
- Síntesis de voz
- Generación de lenguajes naturales
- Traducción automática
- Respuesta a preguntas
- Extracción de la información

En este proyecto evaluaremos la tarea de extracción de información y dentro de ella tres funcionalidades técnicas muy concretas:

- Reconocimiento y categorización de entidades nombradas.
- Clasificación temática del texto.
- Análisis del sentimiento del texto.

2.1.2 Reconocimiento de entidades nombradas

Named-entity recognition (NER) en inglés. Esta subcategoría de la extracción de la información busca localizar y clasificar los elementos del texto en categorías predefinidas, como nombres de personas, organizaciones, lugares, fechas, cantidades, valores monetarios, porcentajes, etc.

En 1926 las fábricas de Daimler y de Benz, que nunca se conocieron personalmente, se reúnen en una sola, la DaimlerBenz A. G.. Como tantos instrumentos del progreso, el automóvil no fue siempre el vehículo indisputado que ahora nos domina. Hubo fuertes resistencias a su introducción, algunas tan pintorescas como la ley británica que exigía que delante de cada auto debía marchar un hombre con una bandera roja, si era de día, o una luz del mismo color, si era de noche. En los Estados Unidos no había prohibiciones pero sí gran alarma del público por las espantadas que producían en los caballos. En 1904 un tal Henry Hayes de Denver, Colorado, patentó una curiosa solución: un caballo mecánico que llevaba adosadas dos ruedecillas en las patas delanteras y las traseras sujetas al eje delantero del auto. Llevaba también unas lucecillas en los ojos, una bocina y una guantera en los cuartos traseros. Con semejante aditamento hípico, el buen inventor confiaba en tranquilizar a los animales. El éxito lo obtuvo un norteamericano de notable carácter y genio organizador que confirió al automóvil todo el sentido popular y utilitario del siglo XX. Hijo de un campesino acomodado de Dearborn, Michigan, Henry Ford (1863 - 1947) se dedicó desde los dieciséis años a máquinas y motores como aprendiz en los talleres de reparación de Detroit. Cuando tenía veintiséis años se empleó en la compañía de electricidad y devoraba cuanto caía en sus manos sobre invenciones mecánicas. De este modo se enteró en 1892 de que Charles Duryea había diseñado el primer automóvil norteamericano. Ford decidió hacer otro tanto y, después de las diez horas de jornada diaria, se refugiaba en su pequeña caseta a montar piezas y probar motores por la noche. Su primer cuadriciclo fue concluido en 1896.
Persona Organización Lugar Fecha

Tabla1 Ejemplo reconocimiento de entidades

En el ejemplo de la **Tabla1** se observa lo que sería la identificación y categorización de entidades para un texto dado. Los sistemas de reconocimiento de entidades parten de un bloque de texto plano sin categorizar como este y devuelve la información estructurada con las entidades identificadas. Esta respuesta suele venir en formatos como XML, RDF ó JSON.

```
<?xml version="1.0" encoding="UTF-8"?>
      <entidades>
            <entidad>
                  <nombre>Benz</nombre>
                  <tipo>Persona</tipo>
            </entidad>
            <entidad>
                  <nombre>DaimlerBenz A. G.</nombre>
                  <tipo>Organización</tipo>
            </entidad>
            <entidad>
                  <nombre>Estados Unidos</nombre>
                  <tipo>Lugar</tipo>
            </entidad>
      </entidades>
```

Tabla2 Ejemplo posible respuesta reconocimiento de entidades en XML

Actualmente estos sistemas de reconocimiento de entidades nombradas han llegado a conseguir un rendimiento casi humano para el inglés. Por ejemplo en la conferencia MUC-7 [7] el mejor sistema consiguió una medida-F del 93,37% mientras que el reconocimiento por parte del humano obtuvo un 97,60%.

A pesar de estas prometedoras cifras los sistemas NER siguen siendo muy frágiles, es decir, los sistemas desarrollados para un dominio no funcionan tan bien en otros ámbitos siendo necesarios grandes esfuerzos para adaptarlos. Actualmente el trabajo se está enfocando en reducir dicha mano de obra mediante técnicas de aprendizaje semisupervisado y en conseguir mayor robustez entre dominios [8].

2.1.3 Categorización temática

Consiste en asignar categorías predefinidas a un texto plano. En nuestro caso estas categorías vendrán dadas por la temática del contenido del texto y pueden tomar valores como política, entretenimiento, deportes, etc.

```
<?xml version="1.0" encoding="UTF-8"?>
	<tweets>
		<tweet>
			<content>En pleno de convalidación decreto d
			recortes, dejamos propuestas: Pleno monográfico
			sobre empleo, mesa para nuevo modelo
			productivo...</content>
			 <temas>
				<tema>economía</tema>
				<tema>política</tema>
			</temas>
		</tweet>
		<tweet>
			<content>Una tesis sobre las crónicas de boxeo de
			Manuel Alcántara recupera el género
			http://t.co/HlgT7Vvf</content>
			 <temas>
				<tema>literatura</tema>
				<tema>deportes</tema>
			</temas>
		</tweet>
	</tweets>
```

Tabla3 Ejemplo de tweets clasificados temáticamente en XML

Esta clasificación basada en contenido funciona etiquetando y dando un peso a las expresiones y entidades particulares del texto ponderando todas ellas para determinar la temática final. Hay que tener en cuenta que, por ejemplo, al menos un 20% de los elementos etiquetados deben ser de la misma temática para poder considerarla para el texto.

La categorización de textos puede aplicarse a:

- Filtrado de correo electrónico no deseado, permitiéndote discernir por la temática entre el *spam* o el correo legítimo.
- Enrutamiento de correo electrónico, para direcciones de correo generales permite distribuirlos y clasificarlos.
- Resumen y clasificación de noticias o artículos periodísticos según su contenido.
- Evaluación del contenido del texto, si es apto para determinadas edades o grupos sociales.
- Simplificación y sinopsis de textos de más tamaño.

Los sistemas de categorización temática de textos presentan unas dificultades similares a los de reconocimiento de entidades, agravándose aún más cuando tratan de evaluar textos cortos. Mensajes de texto, *twitter*, correos electrónicos, titulares de prensa, etc. son cada vez más comunes y manejan ciertas particularidades del lenguaje que a estas herramientas les cuesta dominar.

2.1.4 Análisis de sentimiento

En términos generales consiste en, a partir de un texto dado, determinar la actitud del interlocutor con respecto al tema tratado. La actitud puede considerarse su juicio o evaluación, su estado emocional al escribir o el efecto emocional que intenta generar al lector.

El método más básico consiste en determinar que un texto puede tener un sentimiento positivo, neutro o negativo. Se extraen todos los conceptos del texto y se les da un valor por ejemplo del -10 al +10 (de más negativo a más positivo) en función de su significado (Negativo: 'Triste', 'enfadado'. Positivo: 'Ganar',' feliz'). Después se ponderan todos los resultados y se determina así la polaridad del texto. Evaluando también el número de expresiones positivas y negativas en el texto podremos determinar también el grado de acuerdo del interlocutor con la idea expuesta, es decir, si existe un gran número de opiniones enfrentadas podremos decir que el autor se encuentra en desacuerdo con lo expuesto.

Para realizar estas acciones los sistemas de análisis de textos utilizan elementos de aprendizaje de máquina, como el análisis semántico latente (LSA), máquinas de vectores de soporte, bolsas de palabras u orientación semántica.

El problema de la mayoría de los algoritmos de los sistemas de análisis de sentimiento es que utilizan términos simples para expresar el sentimiento acerca de un producto o servicio. Sin embargo, los factores culturales, matices lingüísticos y contextos diferentes hacen que sea extremadamente difícil que un texto plano se traduzca en un simple sentimiento pro o contra. El propio ser humano no tiene porque interpretarlo de la misma manera, estos evaluadores aciertan con el sentimiento solo el 79% de las veces, por tanto podemos decir que un sistema automatizado con una precisión del 70% está haciendo un gran trabajo casi tan bueno como el ser humano [9].

En los últimos tiempos el auge de los medios sociales en internet como *blogs*, foros o redes sociales ha impulsado el interés en el análisis del sentimientos. Con la proliferación de opiniones, valoraciones, recomendaciones y otras formas de críticas en la red, la opinión *online* se ha convertido en algo a tener en cuenta para las empresas que buscan comercializar sus productos, identificar nuevas oportunidades o gestionar su reputación. Las empresas buscan como automatizar los procesos de filtrado de ruido, comprensión de las conversaciones o la identificación de los contenidos relevantes y por ello están buscando soluciones en el campo del análisis de sentimientos.

El hecho de que los seres humanos a menudo no estén de acuerdo sobre el sentimiento de un mismo texto ilustra cómo de grande es esta tarea para los ordenadores. Además cuanto más corta sea la cadena de texto, más difícil se hace esta labor. Aún queda mucho por descubrir e investigar en el campo del análisis del sentimiento y el interés en ellos es máximo dado el esplendor que viven las redes sociales y sus aplicaciones para la empresa.

2.1.5 Aplicaciones para la empresa

La mayor parte de la información relevante que pasa por las manos de una empresa se genera de forma no estructurada (Correos, mensajes, foros, redes sociales, transcripciones, reclamaciones,…). Si la empresa solo se queda con el mensaje transaccional que suponen estos medios estaría perdiendo gran parte del potencial real que contiene esta información, pudiendo con él mejorar en gran medida sus estrategias de negocio.

Esta estrategia o inteligencia de negocio fue definida ya en 1958:

"Habilidad de recoger las relaciones entre los hechos presentados para orientar las acciones a la meta deseada".

Hans Peter Luhn
"A Business Intelligence System"
IBM Journal, Octubre 1958

Esta metodología acabó desembocando en la aparición de los estudios de mercado y del marketing, área en la que el uso de herramientas de análisis de textos es una pieza fundamental.

MARKETING

Al mando de una empresa uno puede optar por un gran abanico de estrategias a seguir y a lo que podríamos llamar la ciencia encargada de indicar cuál es la opción más beneficiosa para nuestros objetivos sería el marketing.

Generalizando podemos decir que las principales fuentes de datos de una empresa serían los dejados por sus clientes, competidores y mercados. Huelga decir que esta información se encuentra desordenada y que en muchos casos su dimensión supera la capacidad de analizarla sin el uso de herramientas que automaticen la labor de sintetizar y extraer la información útil.

Una vez obtenida esta información se interpreta y analiza dando pie a las posibles estrategias a tomar por la empresa, como por ejemplo:

- Incluir nuevas características a un producto, crear nuevas líneas, nuevas marcas o servicios adicionales como facilidades de pago o garantías.

- Sacar al mercado un productos a precios bajos para garantizar una buena acogida o altos por ser novedoso y crear una sensación de calidad y exclusividad.

- Bajar el precio de un producto para captar clientela y bloquear a la competencia.

- Distribuir el producto por nuevas vías como Internet o visitas a domicilio.

- Promocionar el producto mediante ofertas, patrocinios o publicidad en los medios.

Las herramientas de análisis de textos son especialmente útiles no sólo para predecir qué estrategias de negocio pueden ser más efectivas, sino también para evaluar el éxito que han podido tener en el mercado y planificar futuras decisiones

ÁREAS Y SECTORES DE APLICACIÓN

Aquí se presentan algunos ejemplos de áreas y sectores empresariales concretos en los que las herramientas de análisis de texto y minería de datos pueden jugar un papel importante:

- **Compañías de Seguros y salud privada**:

  La información que pasa por las aseguradoras (Pólizas, volantes, prescripciones médicas, grabaciones de llamadas...) puede servirles para por ejemplo para predecir clientes potenciales que puedan comprar nuevas pólizas, para identificar patrones de comportamiento para clientes con riesgo o para identificar comportamiento fraudulento.

- **Banca**

  Los bancos cuentan con grandes cantidades de información como la relativa a las cuentas, las transacciones y autorizaciones de las tarjetas de crédito, las incidencias y reclamaciones realizadas por los clientes o las operaciones realizadas por los clientes en la propia entidad.

  Estos datos pueden servirles para detectar patrones de uso fraudulento de tarjetas de crédito, identificar clientes leales, determinar gastos en tarjetas de crédito por grupos y comercios o identificar reglas de mercado a partir de históricos.

- **Telecomunicaciones**

  En el sector de las telecomunicaciones se almacena información interesante sobre llamadas telefónicas (Destino, fecha, duración...) o conexiones a internet realizadas por el cliente.

  Además de los numerosos usos comerciales que se le puede dar a esta información también puede usarse para detectar fraudes de Internet, robos de líneas, *hackeos*, etc.

- **Biología y medicina**

Estudios como el del Proyecto Genoma Humano tienden a descubrir cómo funcionan nuestros genes y su influencia en la salud desde un punto de vista físico y funcional.

La bioinformática se encarga de aplicar la tecnología de los computadores a la gestión y análisis de datos biológicos. Esto permite llegar a extraer conocimientos biológicos y médicos a partir de bases de datos experimentales alimentadas mediante técnicas de minería de datos.

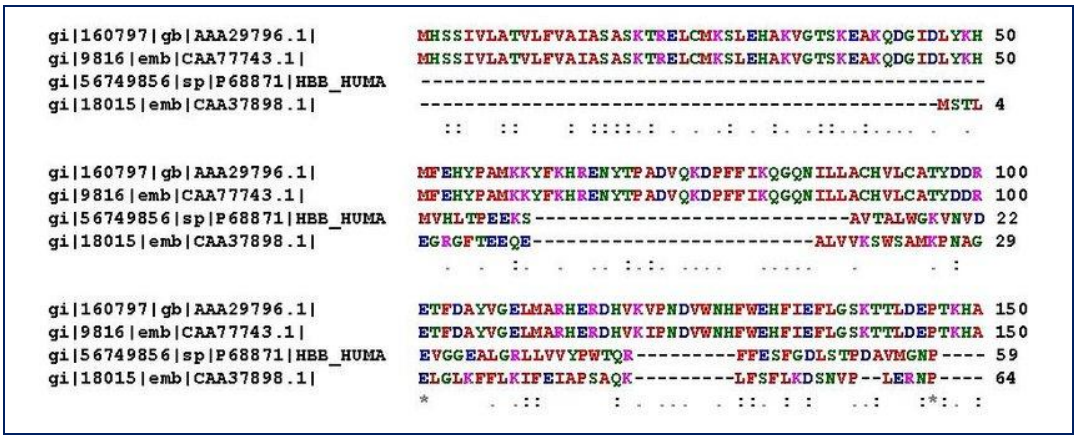


Imagen1 Ejemplo de alineación de diferentes proteínas de hemoglobina realizado por el Instituto Europeo de Bioinformática

Adicionalmente en el campo médico también se almacena información sobre los pacientes como enfermedades pasadas, tratamientos impuestos, pruebas realizadas, evolución...

A partir de dichos datos es posible identificar las terapias médicas más satisfactorias, asociar síntomas y estudios de factores de riesgo a diferentes patologías y facilita la realización de estudios epidemiológicos y análisis de rendimientos de campañas de información o prevención.

2.2 REDES SOCIALES

2.2.1 Introducción

El término red social lleva utilizándose durante cerca de un siglo para referirse a estructuras sociales compuestas por personas conectadas entre sí por distintos tipos de relaciones tales como amistad, parentesco o intereses en común. Estos espacios están normalmente destinados al libre intercambio de conocimientos e ideas entre los individuos que lo componen.

2.2.2 Redes sociales en internet

El fin que ha llevado a la creación de redes sociales en internet es el de diseñar un lugar en el que miles de personas con intereses comunes puedan concentrarse e interactuar entre sí desde sus ordenadores.

Es en 1995 cuando nace el primer sitio web con la intención de cubrir esta necesidad. La web classmates.com fue concebida con el objetivo de que los usuarios registrados pudiesen recuperar el contacto con antiguos compañeros de clase.

A principios del siglo XXI comienzan a surgir sitios web con la idea de crear círculos de amigos en línea, pero no es hasta 2003 con la aparición de Myspace o Xing cuando idea se popularizó y rápidamente se convirtió en un fenómeno de masas. En seguida las grandes compañías de Internet como Google o Yahoo intentaron lanzar sus propias redes sociales en 2004 con Orkut y Yahoo 360° respectivamente aunque no alcanzaron la aceptación y la popularidad que pretendían.

Hoy en día las redes sociales se encuentran en constante crecimiento y evolución y son uno de los principales pilares de las conocidas como Web 2.0 y Web 3.0, donde los usuarios dejan de ser usuarios pasivos para participar y contribuir activamente en el contenido de la red. Dentro de estas tendencias uno de los fenómenos que está en mayor auge son las compras en redes sociales (Shopping 2.0).

TIPOLOGÍA DE LAS REDES SOCIALES EN INTERNET

No existe una tipología de redes sociales clara pero quizá una de las más extendidas sea la de dividirlas en horizontales y verticales al igual que en su día se clasificaron a los portales web.

- **Horizontales**: También llamadas masivas o de propósito general.
  - Facebook.
  - Google+.
  - Tuenti.

- **Verticales:** dirigidas a los usuarios para ofrecer un contenido específico y promover una actividad concreta.
  - Perfiles profesionales: Linkedin, Viadeo.
  - Animales: LoveMascotas.
  - Videos: YouTube, Vimeo.
  - Microbloggin: Twitter.
  - Compras: Buyvip, Privalia, Letbonus.
  - Fotografía: Flickr, Fotolog, Instagram.
  - Música: Goear, Spotify.
  - Literatura: Libros.com.
  - Cine: Moviehaku.
  - Citas: Meetic.
  - Turismo: Tripadvisor, Minube, Toprural.
  - Gastronomía: Kukers, Vinogusto.
  - Deportes: Comunio, Supermanager.

Imagen2 Logos de algunas redes sociales más populares de internet

EJEMPLOS DE REDES SOCIALES

- **MySpace**: Primera gran red social con mayor auge entre 2005 y 2008. De libre acceso ofrece al usuario un espacio para personalizar con fotos, videos, música o entradas de blog.

- **Facebook**: Comenzó como la red social de los estudiantes de la universidad de Harvard pero en la actualidad es de libre acceso. Las buenas decisiones de marketing y el ser una plataforma en la que terceros pueden desarrollar aplicaciones para beneficiarse la han aupado a ser la red social generalista líder.

- **Twitter**: La red permite publicar mensajes de texto plano de un máximo de 140 caracteres (*tweets*). Los usuarios pueden suscribirse a los tweets de otros usuarios.

- **Instagram**: Permite editar fotos, compartirlas y seguir las publicaciones de otras personas. Ha tenido una gran aceptación debido a su compatibilidad con otras redes sociales como Facebook ó Twitter.

- **Vine**: Desarrollada por Twitter permite grabar y compartir pequeños videos de un máximo de 6 segundos de duración que dan como resultado una especie de gifs animados.

- **Amazon Buyvip**: Espacio en el que los usuarios subscritos reciben avisos de campañas comerciales con descuentos, comparten sus opiniones y pueden comparar productos de importantes marcas.

- **Tuenti**: Una red social de características similares a las de Facebook de origen español y solo accesible por invitación. Actualmente está empezando a expandirse por otros países de Europa.

- **Google+**: Intento de Google para competir con Facebook que aún no cuenta con la aceptación necesaria para hacerle sombra, aunque gracias a otras herramientas como Youtube o Gmail está ganando adeptos durante el 2013.

- **Linkedin**: Concebido para que usuarios y empresas compartan sus perfiles profesionales.

ANÁLISIS DE REDES SOCIALES DE INTERNET

Con la creciente importancia de las redes sociales en las interacciones humanas diversos campos, desde la sociología hasta la gestión de conocimiento empresarial, las han tomado como objeto de estudio y fuentes de información. Para ello se han de estudiar tanto la información y el conocimiento que pueden aportar los nodos como las asociaciones o nexos entre los mismos y su estructura.

Hay que tener en cuenta que no es lo mismo analizar unas redes sociales que otras ya que cada una tiene sus particularidades y características. Por ejemplo en una red estructurada como Tripadvisor la opinión de todos los usuarios vale lo mismo a la hora de evaluar un establecimiento. En otras redes como Twitter la temática es mucho más general, no todos los nodos o usuarios tienen el mismo peso y repercusión, por lo general las opiniones no son objetivas y la información está fluyendo en tiempo real.

Es por ello que una parte importante de este proyecto está dedicada al estudio de tweets, ya que pueden convertirse en estudios de mercado inmediatos y fiables para medir la aceptación que pueden tener programas de televisión, películas, decisiones políticas, etc.

ESTADÍSTICAS

A continuación se muestran algunos gráficos con las estadísticas de las redes sociales más utilizadas a nivel global durante el 2013 según un estudio de GlobalWebIndex[10].

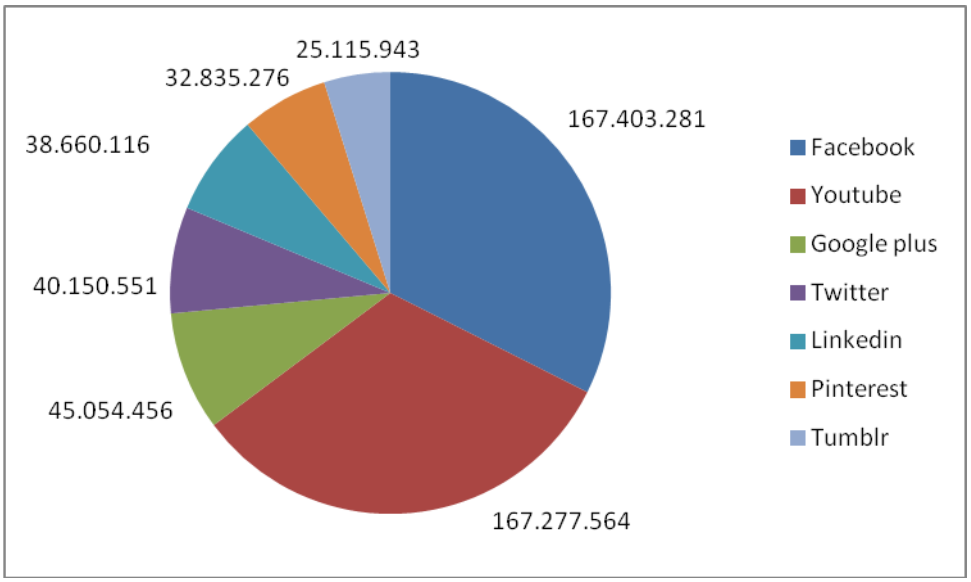


Gráfico1 Número de visitas únicas por mes en 2013

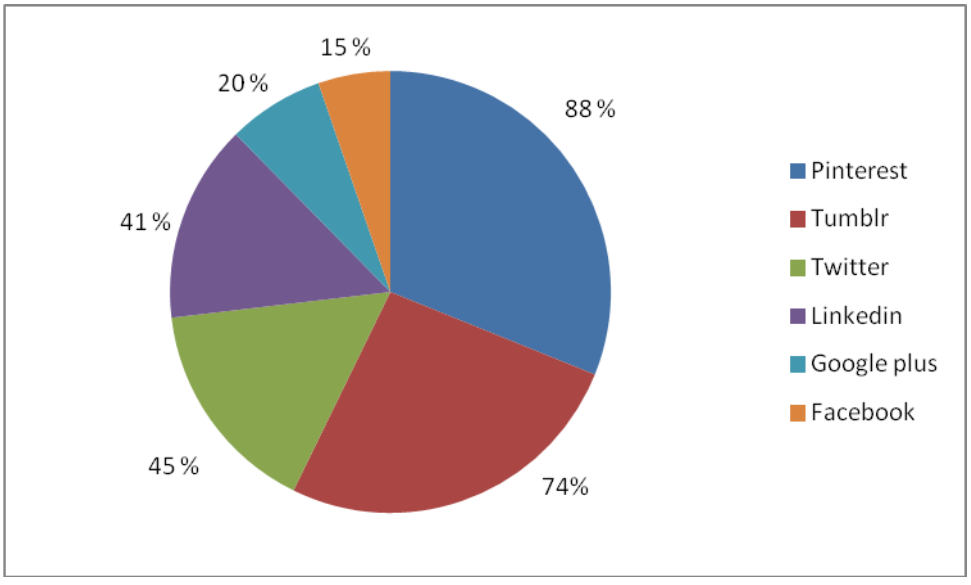


Gráfico2 Porcentaje de aumento de usuarios activos durante el 2013

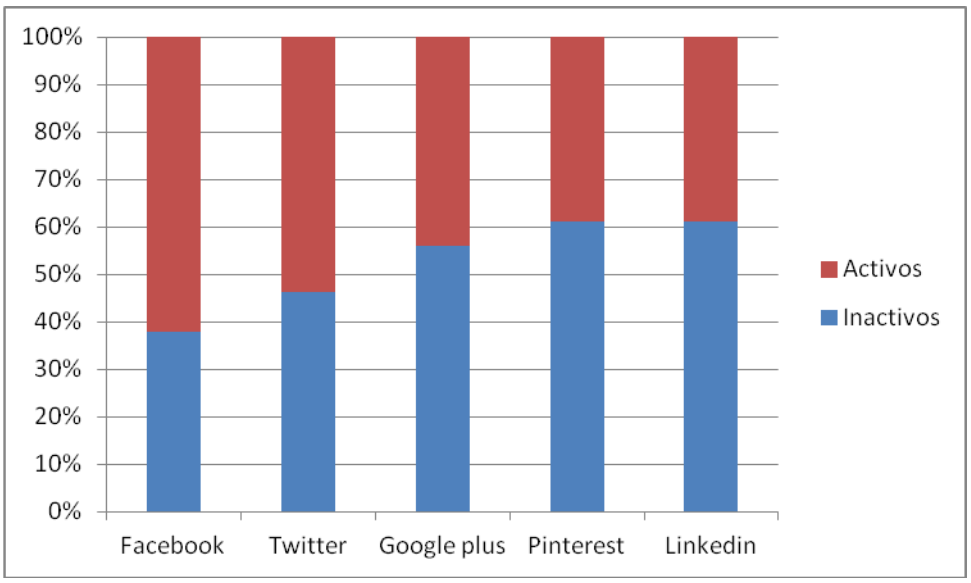


Gráfico3 Porcentaje de usuarios activos e inactivos del total de usuarios durante el 2013

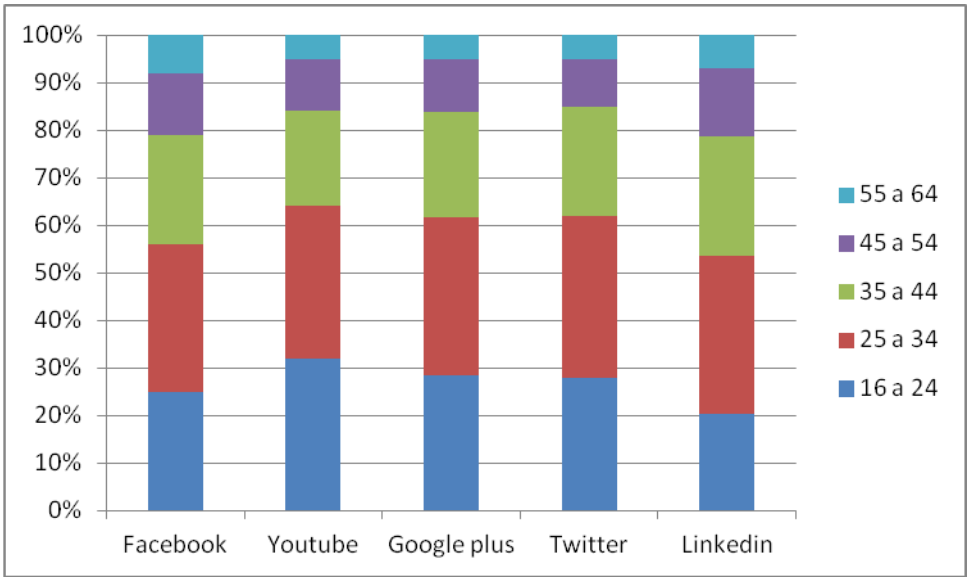


Gráfico4 Porcentaje de usuarios por rango de edad

2.2.3 Información en las redes sociales

Facebook cuenta con cerca de 900 millones de usuarios en todo el mundo. Cierto que un porcentaje de estos usuarios estará inactivo y que otros tantos se tratarán de cuentas falsas pero eso no quita la cantidad de información almacenada que puede existir en sus bases de datos:

- Listas de amigos.
- Grupos y páginas de aficiones.
- Biografías de perfiles.
- Fotos, videos, etc.

Además sin ser datos públicos Facebook también recopila información del tipo:

- Cada vez que un usuario ingresa y sale de la página.
- Información eliminada por el usuario (Amistades denegadas, etiquetas, mensajes o chats privados, etc.)
- Ubicaciones geográficas e IPs con las que accedes a la aplicación.
- Números telefónicos con los que te registras.

Toda esta información correctamente analizada ofrece a Facebook un gran conocimiento y poder, ya no solo del individuo, sino de las tendencias de los grupos sociales de cualquier país que haga uso de la red social hasta el punto de que son muchas las cuestiones éticas y morales que se plantean.

En cualquier caso en la actualidad el gran titán de este tipo de información a pesar de no contar con una red social tan masificada como Facebook es Google. Estos cuentan con información ya no solo de usuarios registrados si no de cualquiera que haga uso de sus navegadores, buscadores o dispositivos.

	Cuenta	Nombre Direcciones de correo Aplicaciones y sitios asociados Ubicaciones Navegadores Plataformas
	Gmail	Mensajes enviados y destinatarios Mensajes recibidos y procedencia Conversaciones y chats
	Historial de ubicaciones	Ubicación actual Historial de ubicaciones Distancia recorrida Sitios visitados Viajes recientes Porcentaje de tiempo que pasas en casa o en el trabajo semanalmente
	Calendar	Calendarios Eventos recientes
	Cloud Print	Impresoras conectadas
	Contactos	Todos los contactos y sus datos personales
	Docs	Documentos abiertos Documentos compartidos
	Picasa	Mis imágenes e imágenes consultadas
	Play Music	Mis canciones y listas de reproducción
	Play Store	Aplicaciones instaladas
	Wallet	Compras Información de los medios de pago
	Youtube	Mis videos y videos reproducidos
	Historial web	Búsquedas Consultas principales Tipos de búsquedas (Imágenes, noticias, etc.)
	Android	Dispositivos asociados Copias de seguridad IMEI

Tabla4 Ejemplo de parte de la información que puedes consultar desde el panel de control de una cuenta de google

3 EMPRESAS Y PRODUCTOS

Todas las empresas seleccionadas basan parte de su negocio en proveer licencias de uso a un servicio web (SaaS API) de análisis de texto. Este, a partir de un texto plano a analizar, devuelve en un formato interpretable la información extraída.

Gracias a que todas las APIs se basan en la llamada a un servicio web hacen que su integración a sistemas de terceros sea sencilla, pudiendo ser usada mediante, por ejemplo, SOAP ó los métodos GET y POST del protocolo HTTP.

La cartera de clientes a la que están enfocadas estas herramientas es reducida aunque en su mayoría dicha cartera está compuesta por grandes empresas.

3.1 OPEN AMPLIFY

Modelo de negocio

Se trata de una empresa especializada en el procesamiento del lenguaje natural y el análisis de texto. Su producto es uno de los líderes en el sector desde hace 7 años.

Además del acceso a su servicio web la empresa oferta otra serie de productos relacionados con el análisis de textos enfocados a internet y a las redes sociales, como por ejemplo

- SocialView: Herramienta visual de análisis de textos y reportes de audiencia
- Tribes CRM: Software para la administración de la relación con los clientes.
- Twittervs: Aplicación para AppStore que compara en tiempo real famosos, películas o equipos deportivos y te devuelve cuál está mejor valorado en Twitter.

Algunas de las principales empresas que utilizan su software son Ebay, Microsoft Dynamics, Ford o Unilever.

Funcionalidades

Se trata de un servicio web capaz de, además de analizar y extraer entidades y relaciones, devolver una representación estructurada del significado, el estilo o la intención. Por tanto las funcionalidades que oferta su API son las siguientes:

- Análisis de *Topics*: Devuelve los temas, entidades, dominios y localizaciones de los que trata el texto junto con su connotación e importancia y las relaciones existentes entre sí.
- Análisis de acciones: Devuelve las acciones detectadas en el texto y su nivel de decisión, orientación o temporalidad.
- Análisis de estilo: Devuelve el estilo que el autor le ha dado al texto en medida por ejemplo de lo florido de la escritura o de si se trata de algo más coloquial.

- Análisis demográfico: Devuelve la edad, el género o el nivel de educación que probablemente tenga el autor del texto y/o la audiencia a la que va destinado el mismo.

- Análisis de búsqueda: Devuelve los topics y las acciones asociadas a palabras clave introducidas en la búsqueda.

Idiomas con los que trabaja

Actualmente la herramienta solo está preparada para procesar textos en inglés.

Fuentes de datos con los que opera

Al servicio pueden pasársele como parámetros de entrada tanto un texto plano como una dirección web con el texto a analizar.

La respuesta del servicio es también parametrizable y puede venir en formatos tales como XML, RDF, RDFA, CSV, DART, OAS, JSON ó con versiones más visuales para en páginas HTML interpretables por el navegador.

Acuerdo de nivel de servicio

Cualquier usuario registrado puede utilizar una demo con hasta 1000 transacciones al día de forma gratuita. En caso de querer superarlas la aplicación será de pago.

Documentación para usuarios

Además de manuales en Inglés y códigos de ejemplo para distintos idiomas de programación cuenta con una serie de recursos como blogs o foros donde los usuarios pueden dar su opinión, pedir soporte y estar informados de las últimas novedades de la herramienta y del sector.

3.2 SEMANTRIA

Modelo de negocio

Fundada en 2011 Semantria como tal solo comercializa la licencia y el acceso a su API. Para los no desarrolladores es posible descargar un complemento para Excel que permite configurar la herramienta, analizar textos e interpretar sus resultados desde la hoja de cálculo.

Son otras empresas como Lexalytics o Zapier las que hacen las veces de *front-end* de Semantria y comercializan su propio software de análisis de información y monitorización de redes sociales.

Funcionalidades

Su API ofrece las siguientes posibilidades:

- Puntuación del sentimiento del texto.
- Síntesis de las ideas del documento.
- Extracción tema del documento.
- Extracción de entidades.
- Tipificación de entidades extraídas.
- Extracción de las relaciones entre las entidades. *
- Categorización consulta basada
- Extracción de opiniones. *
- Clasificación según consultas.
- Reconocimiento de facetas y atributos entre varios textos.
- Reconocimiento de temas en común entre varios textos.
- Reconocimiento de entidades en común entre varios textos.
- Clasificación según consultas entre varios textos.

*Solo para textos en Inglés.

Idiomas con los que trabaja

Semantria está preparado para operar con textos en Inglés, Francés, Alemán, Portugués, Chino y Castellano aunque como se ha visto en el punto anterior es con el idioma Inglés con el que puede ofrecer el máximo de sus posibilidades .

Fuentes de datos con los que opera

El servicio puede trabajar con entradas tanto de textos sueltos como de colecciones de textos.

El formato de respuesta del servicio es también parametrizable y puede venir en XML ó JSOP.

Acuerdo de nivel de servicio

Semantria permite hasta un total de 10000 llamadas gratuitas a su servicio a cualquier usuario registrado que quiera probar su herramienta. Si se desea superar ese crédito se ofertan varios paquetes de llamadas en función de las necesidades de cada cliente.

Documentación para usuarios

Cuenta con multitud de manuales en inglés y numerosos ejemplos. Además tiene a disposición de cualquiera una serie de SDKs para configurar y utilizar su herramienta en las siguientes plataformas: C++, Java, .Net, PHP, Python, Ruby y JavaScript.

3.3 ALCHEMY API

Modelo de negocio

AlchemyApi fue fundada en 2005 y, al igual que la anterior, solamente comercializa los accesos a su servicio web de análisis de textos, aunque sí pone a disposición de sus usuarios unas herramientas visuales muy simples para realizar estas llamadas. Actualmente soporta un tráfico de 3500 millones de transacciones mensuales, con más de un 90% de las mismas de pago.

Algunas de las empresas más importantes que utilizan el software de AlchemyAPI son Outbrain, LiveFyre, YouGov, Jive Software, Cision, and Shutterstock.

Funcionalidades

Su API de análisis de texto ofrece las siguientes posibilidades:

- Extracción de entidades.
- Análisis de sentimientos. *
- Extracción de palabras clave.
- Identificación de conceptos. *
- Extracción de relaciones. *
- Categorización del texto.
- Identificación del autor del texto.
- Identificación del idioma (entre más de 97).
- Síntesis del texto.
- Identificación de fuentes o canales web (ATOM ó RSS *feeds*)

*Solo para textos en Inglés.

Idiomas con los que trabaja

El total de las funcionalidades solo puede aprovecharse para la lengua inglesa pero también es capaz de trabajar en Francés, Alemán, Italiano, Portugués, Ruso, Sueco y Castellano.

Fuentes de datos con los que opera

En la llamada al servicio se puede pasar como parámetro un texto o una dirección web a analizar.

El formato de la respuesta es parametrizable y puede venir en XML, JSON, RDF y otros microformatos.

Acuerdo de nivel de servicio

Al registrase en su web AlchemyAPI permite gratuitamente hasta 30000 llamadas diarias a su servicio para usos no comerciales y 1000 para usos comerciales. De ser necesarias más llamadas el producto pasa a ser de pago.

Documentación para usuarios

Cuenta en inglés con un blog y con documentación detallada sobre la llamada a su servicio y la utilización de su API. Además tiene a disposición de cualquier programador una serie de SDKs en las siguientes plataformas: Python, PHP, Node.js, Ruby, Android OS, Perl, Java, C/C++ y .NET.

3.4 TEXTALYTICS

Modelo de negocio

Fundada en 1998, Daedalus es una empresa pionera en el sector de procesamiento y análisis de información en España. Además del licenciamiento de su servicio web de análisis de textos Textalytics, Daedalus comercializa otros productos como:

- K-Site**:** Es una completa familia de módulos de tecnología lingüística, semántica y de gestión de contenidos multimedia, pensada para ofrecer soluciones personalizadas a clientes corporativos.

- Stilus**:** Es herramienta de revisión de textos multiidioma. Realiza no solo correcciones ortográficas y gramaticales, sino también una revisión de estilo con sugerencias y explicaciones didácticas.

- Sentimentalytics**:** Es un plug-in para navegador que analiza "al vuelo" y etiqueta semánticamente los timelines que aparecen en redes y herramientas sociales.

Algunas de sus principales cliente son Telefónica, grupo PRISA, Hispasat, Unidad Editorial, Iberdrola, Vocento, Amper, RTVE, Eutelsat, Hibü, Instituto Cervantes ó Digimind.

Funcionalidades

Textalytics se divide en una serie de APIs, cada una con una funcionalidad:

- Topics Extraction API: Identifica conceptos como personas, lugares, empresas, fechas, direcciones, citas y relaciones entre ellos.

- Text Classification API: Clasifica textos de acuerdo a categorías existentes (ej.: IPTC) o a clases definidas por el usuario.

- Sentiment Analysis API: Identifica si un texto es irónico o si expresa una opinión positiva o negativa, según el contexto.

- Language Identification API: Averigua automáticamente el idioma de un texto. En más de 20 idiomas.

- Spell, Grammar and Style Proofreading API: Detecta errores ortográficos, gramaticales y de estilo en varios idiomas.

- Semantic Linked Data Viewer API: Consulta distintas fuentes Linked Data a partir de un identificador.

- User Demographics API: Averigua si un usuario es una empresa o una persona, su edad y su género.

- Lemmatization, POS and Parsing API: Analiza morfosintácticamente un texto, proporcionando lemas y etiquetas de discurso. En varios idiomas.

- Speech Recognition and Speaker Diarization API: Capacidades de reconocimiento de voz y locutor para cualquier aplicación. Aplica la tecnología lingüística de Textalytics sobre contenidos de audio y vídeo.

Idiomas con los que trabaja

Textalytics acepta textos tanto en Castellano como en Inglés, aunque se está empezando a trabajar en otras lenguas.

Fuentes de datos con los que opera

El texto en plano a analizar se pasa como parámetro de entrada en la llamada al servicio. Se permiten también documentos en formato JSON o XML.

El servicio puede responder con formato JSON o XML.

Acuerdo de nivel de servicio

Al registrase en su web de Textalytics se conceden hasta 500000 créditos mensuales de forma gratuita a utilizar en sus distintos servicios. Si son necesarios más accesos la herramienta pasa a ser de pago.

Documentación para usuarios

La web cuenta con documentación tanto en castellano como en inglés en la que se detalla la utilización del servicio y la interpretación de su respuesta. Adicionalmente cuentan con SDKs descargables en Java, PHP y Python para integrar rápidamente la herramienta en cualquier plataforma en estos idiomas.

3.5 CALAIS

Modelo de negocio

Calais lleva en el sector desde 2008 y su negocio se basa exclusivamente en la venta de licencias para la utilización de su servicio web de análisis de textos OpenCalais (gratuito) ó ProfessionalCalais (de pago).

Presenta muchas menos funcionalidades que las anteriores empresas pero fomenta más la libre utilización de su software promocionando herramientas propias, como Marmoset ó Tagaroo, o de terceros como Sulava ó WikiDo que utilizan la versión gratuita de su servicio.

Funcionalidades

Su API de análisis de texto ofrece las siguientes posibilidades:

- Extracción y categorización de entidades.
- Extracción de temas de la entidad. *
- Extracción de relaciones entre entidades. *
- Evaluación de relevancia de las entidades.
- Categorización del texto. *

*Solo para textos en Inglés

Idiomas con los que trabaja

OpenCalais trabaja en Inglés, Francés y Castellano, aunque es con el Inglés con el que ofrecen el máximo de las funcionalidades.

Fuentes de datos con los que opera

En la llamada al servicio se puede pasar como parámetro un texto o una dirección web a analizar.

El formato de la respuesta es parametrizable y puede venir en XML, JSON, RDF, N3 y otros microformatos.

Acuerdo de nivel de servicio

Al registrarse en la página de Calais se obtiene libre acceso al servicio OpenCalais, con un máximo de 50000 llamadas diarias. En caso de necesitar superarlas se deberá contratar el servicio ProfessionalCalais de pago.

Documentación para usuarios

Cuenta en inglés con un blog y con documentación detallada sobre la llamada a su servicio y la utilización de su API. Además reúnen y ponen a disposición de cualquier usuario herramientas, aplicaciones pluggins ó SDKs desarrollados por ellos mismos o por terceros que utilizan como back-end su servicio y que pueden resultar de utilidad.

3.6 ZEMANTA

Modelo de negocio

Zemanta lleva en funcionamiento desde 2007 y su herramienta es un *plugin* enfocado a editores, anunciantes o *bloggers* y que realiza sugerencias de contenido a partir de la información publicada.

Algunas de los principales clientes que han integrado Zemanta en sus webs son Forbes, ReadWriteWeb, Wall street Journal, The Blaze ó WordPress.

Funcionalidades

Podemos resumir las funciones de su API SaaS en las siguientes:

- Reconocimiento y clasificación de las entidades nombradas. A partir de esta información propone las siguientes sugerencias.
- Relevancia de las entidades.
- Categorización de textos.
- Reconoce nombres de sitios web para que con un sólo clic se incorpore un hipervínculo en la primera aparición del nombre.
- A partir de las entidades reconocidas, su clasificación y de la categorización del texto:
    - Sugiere fotografías públicas y también fotos de tu usuario de Flickr para ilustrar el artículo.
    - Propone noticias y artículos de distintos medios para que se incluyan al final de la entrada a modo de artículos relacionados
    - Propone palabras clave (*tags*) para las entradas.

Idiomas con los que trabaja

Actualmente Zemanta solamente soporta textos en Inglés y no tiene planes de implementar cualquier otro idioma a corto plazo. De todos modos al tratarse de una herramienta que en su mayoría identifica marcas, personas, localizaciones u otros términos universales puede ser una herramienta útil para otros idiomas.

Fuentes de datos con los que opera

El servicio web de Zemanta acepta como parámetro de entrada el texto en plano o como HTML.

El formato de la respuesta es parametrizable y puede venir en XML, JSON, WNJSON ó RDFXML.

Acuerdo de nivel de servicio

Al registrase en su web Zemanta permite gratuitamente hasta 1000 llamadas diarias a su servicio. En caso de querer superarlas habría que ponerse en contacto con ellos para acordar un plan de pago.

Documentación para usuarios

Cuenta en inglés con documentación y con ejemplos de código para realiza la llamada a su servicio, aunque hay que decir que es la más escasa de todas.

4 PRUEBAS TÉCNICAS

4.1 PREPARACIÓN DE DATOS

4.1.1 Objetivos del análisis

Para este estudio, como ya hemos comentado con anterioridad, vamos a centrarnos en evaluar tres funcionalidades concretas de las que nos ofrecen las distintas herramientas con las que trabajaremos.

- Reconocimiento y categorización de entidades nombradas.
- Clasificación temática del texto.
- Análisis del sentimiento y el tono emocional del texto.

En el siguiente cuadro se detalla la relación entre las funcionalidades anteriores y los productos a estudiar.

	ALCHEMY_API	TEXTALYTICS	OPENCALAIS	OPENAMPLIFY	SEMANTRIA	ZEMANTA
ENTIDAD	X	X	X	X	X	X
CATEGORIZACIÓN	X	X		X	X	
SENTIMIENTO		X		X	X	

Tabla5 Relación de herramientas estudiadas y funcionalidades a evaluar

Una vez obtenidos los resultados evaluaremos el rendimiento de cada herramienta mediante los siguientes parámetros estadísticos:

- **Precisión:** fracción de instancias recuperadas que son relevantes.

- **Recall o exhaustividad**: fracción de instancias relevantes que han sido recuperadas.

- **Valor-F:** Media armónica de las dos métricas anteriores para ponderar como de lejanas se encuentran ambas.

4.1.2 Muestra de datos

Para el estudio se utilizarán dos baterías de textos, una primera (ANCORA) de artículos periodísticos de distintos medios en la que ya vienen identificadas y clasificadas las entidades y una segunda (TASS) de tweets de los que conocemos la temática de la que tratan y el sentimiento.

Corpus AnCora[11]

Es un corpus en castellano y catalán con los siguientes niveles de anotación:

- Lema y categoría morfológica.
- Constituyentes y funciones sintácticas.
- Estructura argumental y papeles temáticos.
- Clase semántica verbal.
- Tipo denotativo de los nombres deverbales.
- Sentidos de WordNet nominales.
- **Entidades nombradas y su clasificación.**
- Relaciones de correferencia.

Contiene 500000 palabras y se compone fundamentalmente de textos periodísticos. Fue desarrollado por el *Centre de Llenguatge i Computació* (Universidad de Barcelona, España) y el Grupo de Procesamiento del Lenguaje Natural (*Universitat Politècnica de Catalynya,* España).

CORPUS ANCORA		
Formato entrada XML		
279 artículos		
500000 palabras		
2584 entidades		
Tipos de entidades	*Person*	39,74 %
	Location	22,6 %
	Organization	23,64%
	Other	14 %

Tabla6 Cuadro resumen de la información de entidades nombradas del corpus Ancora

Corpus TASS 2012[12]

Este Corpus está compuesto por 7218 tweets escritos en castellano de 153 usuarios del mundo de la política, la comunicación o la cultura publicados entre Diciembre del 2011 y Abril del 2012. Las entidades que han participado en su creación son Daedalus (España), Grupo de Sistemas inteligentes (GSI, Universidad Politécnica de Madrid) y Sistemas Inteligentes de Acceso a la Información (SINAI, Universidad de Jaén).

Por cada tweet el corpus contiene la siguiente información:

- Entidades nombradas.
- Clasificación temática del texto.
- Sentimiento del texto entre positivo, negativo o neutral.
- Grado de acuerdo o desacuerdo del interlocutor con el sentimiento expuesto.

Este concepto se explicará mejor en los siguientes ejemplos del corpus.

```
<twit>
	<twitid>142379173120442368</twitid>
	<user>LosadaPescador</user>
	<content>Gonzalo Altozano tras la presentación de su libro 101
	españoles y Dios. Divertido, emocionante y
	brillante.http://t.co/4BdljMhB</content>
	<date>2011-12-02T00:06:55</date>
	<lang>es</lang>
	<sentiments>
		<polarity>
			<value>P</value>
			<type>AGREEMENT</type>
		</polarity></sentiments>
	<topics>
		<topic>literatura</topic>
	</topics>
</twit>
```

Tabla7 Ejemplo 1 del corpus TASS 2012

En este primer ejemplo, el contenido del tweet es totalmente positivo. Y por tanto su polaridad sería P (Positiva) con AGREEMENT (Acuerdo).

```
<twit>
	twitid>142387445416210433</twitid>
	<user>CarmendelRiego</user>
	<content>Habia prometido responder a todos, pero me ha sido
	imposible. Y hoy no doy para mas. MUCHAS GRACIAS A
	TODOS</content><date>2011-12-02T00:39:47</date>
	<lang>es</lang>
	<sentiments>
		<polarity>
			<value>NONE</value>
			<type>DISAGREEMENT</type>
		</polarity>
	</sentiments>
	<topics>
		<topic>otros</topic>
	</topics>
</twit>
```

Tabla8 Ejemplo 2 del corpus TASS 2012

Sin embargo, en este segundo ejemplo, en un mismo tweet se expresan ideas tanto positivas ('había prometido responder a todos' o 'muchas gracias a todos') como negativas ('pero me ha sido imposible' o 'hoy no doy para más').
Al final la polaridad del tweet queda como NEU y para indicar esa contraposición de ideas se define como DISAGREEMENT.

<table>
<tr><th colspan="5">CORPUS TASS</th></tr>
<tr><td colspan="5">Formato entrada XML</td></tr>
<tr><td colspan="5">7218 tweets</td></tr>
<tr><td colspan="5">5341 entidades</td></tr>
<tr><td colspan="5">9568 clasificaciones temáticas</td></tr>
<tr><td colspan="2" rowspan="10">Tipos de temáticas</td><td colspan="2">Entretenimiento</td><td>17,54%</td></tr>
<tr><td colspan="2">Política</td><td>32,61 %</td></tr>
<tr><td colspan="2">Economía</td><td>9,85 %</td></tr>
<tr><td colspan="2">Literatura</td><td>1,03 %</td></tr>
<tr><td colspan="2">Tecnología</td><td>2,27 %</td></tr>
<tr><td colspan="2">Deportes</td><td>1,18 %</td></tr>
<tr><td colspan="2">Fútbol</td><td>2,63 %</td></tr>
<tr><td colspan="2">Música</td><td>5,92 %</td></tr>
<tr><td colspan="2">Cine</td><td>2,56 %</td></tr>
<tr><td colspan="2">Otros</td><td>24,41 %</td></tr>
<tr><td rowspan="3">Valores polaridad sentimiento</td><td>NONE</td><td>32,01 %</td><td rowspan="3">Grado de acuerdo con el sentimiento</td><td rowspan="2">AGREEMENT</td></tr>
<tr><td rowspan="2">P: Positivo</td><td rowspan="2">38,56 %</td></tr>
<tr><td>N: Negativo</td><td>29,43 %</td><td>DISAGREEMENT</td></tr>
</table>

Tabla9 Cuadro resumen de la información del corpus TASS 2012

4.1.3 Formateo de datos

Dichos corpus de datos vienen en un formato XML que formatearemos para quedarnos con la información útil y se insertará en una base de datos para facilitarnos su tratamiento. Toda la parte técnica se ha desarrollado en una plataforma PHP y MySQL.

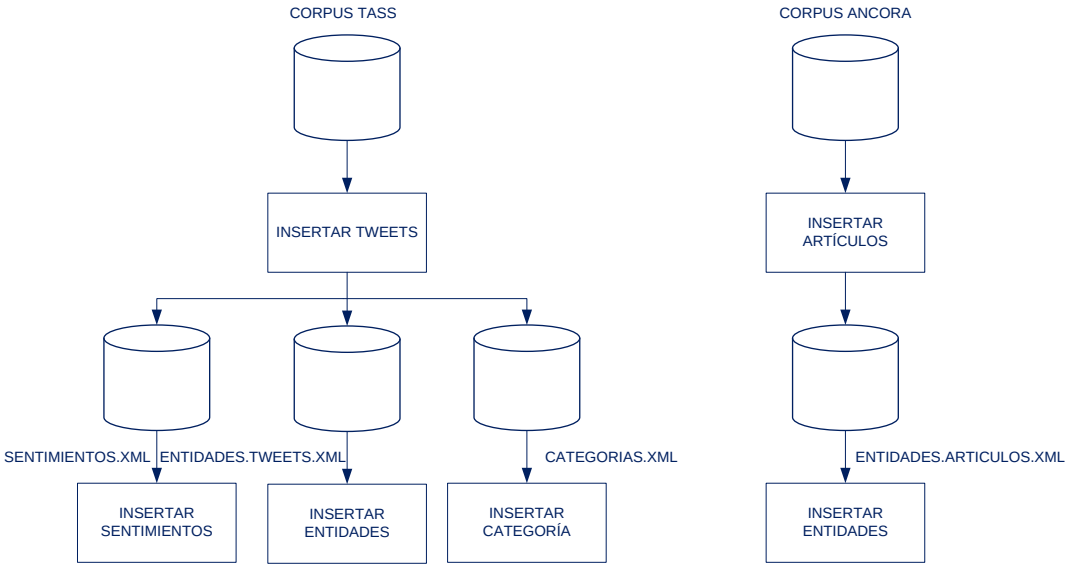


Imagen3 Diagrama de flujo de formateo de datos de entrada y su inserción en la BD

CATEGORIA.XML

```
<?xml version="1.0" encoding="UTF-8"?>
	<tweet>
		<id>142493511634259968</id>
		<texto>Hoy asisitiré en Madrid a un seminario sobre la Estrategia
		Española de Seguridad organizado por FAES.</texto>
		<topics>
			<topic>política</topic>
		</topics>
	</tweet>
```

SENTIMIENTOS.XML

```
<?xml version="1.0" encoding="UTF-8"?>
	<tweet>
		<id>144033311872925696</id>
		<texto>Rajoy rompe su silencio en la fiesta del Congreso plagada
		de ausentes http://t.co/C52Ex8M0</texto>
		<sentimiento>
			<valorpol>N</valorpol>
			<tipopol>AGREEMENT</tipopol>
		</sentimiento>
	</tweet>
```

ENTIDADES.TWEETS.XML

```
<?xml version="1.0" encoding="UTF-8"?>
	<tweet>
		<id>160816903575707649</id>
		<texto>Que rrrriiiicaaaaa RT @naranzu: @AlejandroSanz y de
		regalo...una Leche Frita desde Cadiz para que (cont)
		http://t.co/2gsWw7Gw</texto>
		<entidad>@AlejandroSanz</entidad>
		<entidad>Cadiz</entidad>
	</tweet>
```

ENTIDADES.ARTICULOS.XML

```
<?xml version="1.0" encoding="UTF-8"?>
	<articulo>
		<id>4</id>
		<texto> Si ahora hay gente que te odia no es porque seas el
		número uno, sino porque has matado el romanticismo y porque
		formas ya parte de los nuevos ricos. Toda tu cultura se reduce a
		esperar al rival, a morderle, a sujetarle y a extorsionarle.
		Después, o en medio, le das un mordisco traicionero y te llevas
		los puntos a la guarida donde preparas tu próximo asalto. Oye,
		haznos caso. Párate un momento, mira hacia atrás y reconócete.
		Verás qué jugoso era todo aquello y cuánta gente te quería.
		Volverás a oír los aplausos enemigos y las admiraciones de casa.
		Seguirás arriba, no te preocupes, pero habrás rescatado un trozo
		de ilusión para los domingos. Un Madrid algo peor que el de la
		Liga y un Atleti bastante mejor se encontraron, en la Copa, en un
		punto intermedio del recorrido. El empate resultó correcto
		geográficamente, pero futbolísticamente ofreció sólo una
		apariencia de justicia. </texto>
		<entidad>
			<nombre>Madrid</nombre>
			<tipo>location</tipo>
		</entidad>
		<entidad>
			<nombre>Liga</nombre>
			<tipo>other</tipo>
		</entidad>
		<entidad>
			<nombre>Atleti</nombre>
			<tipo>organization</tipo>
		</entidad>
		<entidad>
			<nombre>Copa</nombre>
			<tipo>other</tipo>
		</entidad>
	</articulo>
```

Tabla10 Ejemplo de los ficheros formateados que se manejarán

Como se ha comentado con anterioridad, debido al gran tamaño de estos ficheros, se ha diseñado una base de datos sobre la que trabajar con mayor facilidad.

4.1.4 Base de datos

A continuación se expone el diseño de la base de datos que utilizaremos para almacenar los resultados del análisis realizado por las distintas herramientas. Posteriormente se partirá de dicha información para obtener las estadísticas de cada herramienta y determinar su capacidad para trabajar en castellano. En el apartado de anexos se detalla la información almacenada en cada tabla.

ESQUEMA DE LA BASE DE DATOS

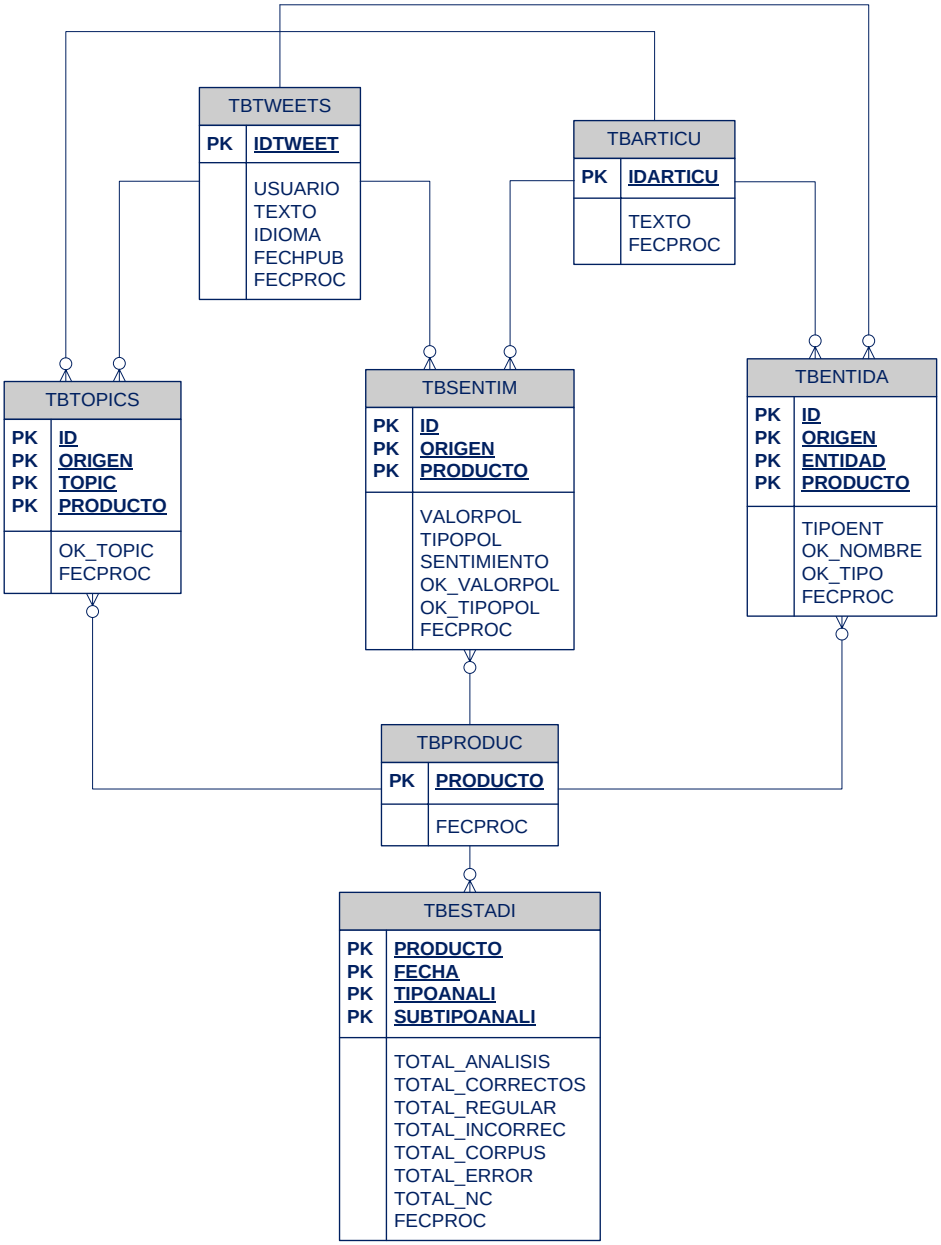


Imagen4 Modelo relacional de la base de datos utilizada en el estudio

4.2 PRUEBA DE HERRAMIENTAS

Estas pruebas se han realizado durante Agosto del 2014. Como se ha comentado con anterioridad todos los productos seleccionados funcionan como una API SaaS.

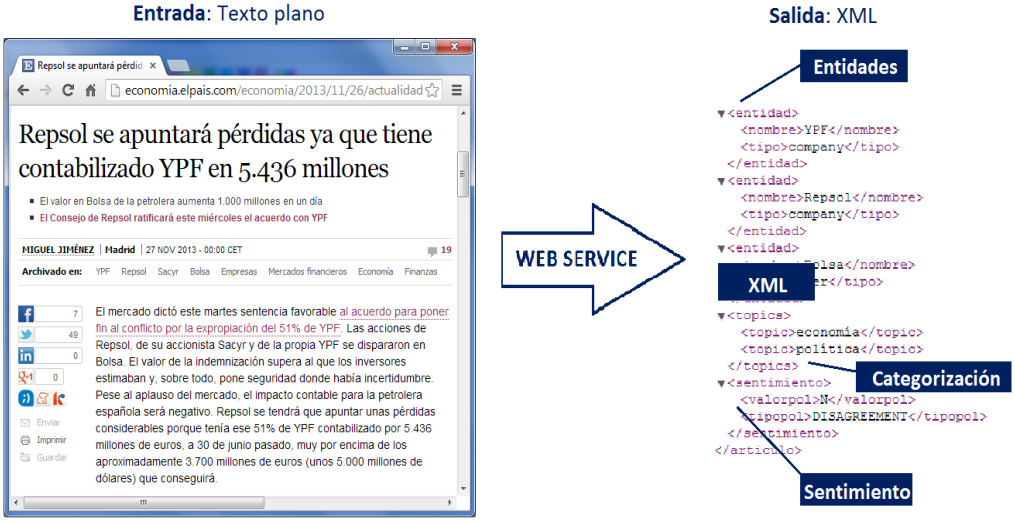


Imagen5 Ejemplo de funcionamiento de cualquiera de las APIs a estudiar

Para integrar estas llamadas en nuestro sistema en PHP se ha hecho uso de la librería cURL que nos permite realizar peticiones mediante el método POST de HTTP entre otros muchos protocolos.

```
// Crear un nuevo recurso cURL
$ch = curl_init();
// Se establece la URL y otras opciones de la llamada
curl_setopt($ch, CURLOPT_URL, $url);
curl_setopt($ch, CURLOPT_POST, 1);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
// Se pasan los parámetros del servicio
curl_setopt($ch, CURLOPT_POSTFIELDS, $data);
// Se captura la respuesta
$response = curl_exec ($ch);
// Cerrar el recurso cURL y liberar recursos del sistema
curl_close ($ch);
// Se crea un objeto XML para interpretar la respuesta
$respuesta = new SimpleXMLElement($response);
```

Tabla11 Ejemplo de llamada a cualquiera de los servicio web a estudiar

A continuación se detallará como se interpreta la respuesta de cada herramienta y se integra en el sistema para cada una de las funcionalidades a estudiar.

4.2.1 Reconocimiento y categorización de entidades nombradas

RECONOCIMIENTO NOMBRE ENTIDAD

Lo primero a evaluar será que la entidad reconocida se encuentra en el CORPUS. Por tanto se insertará un registro en la tabla de entidades TBENTIDA por cada entidad reconocida por la herramienta y se actualizará el campo OK_NOMBRE de la siguiente manera.

SI	El nombre identificado es exactamente el mismo que el del CORPUS
NO	El nombre identificado no se parece en nada a alguno del CORPUS
RG	El nombre identificado contiene una parte igual que el del CORPUS (Ej: La herramienta identifica 'Juan' y en el CORPUS existe 'Juan Pérez')
NA	Si la herramienta identifica un tipo de entidad que no existe en el CORPUS no aplicará a nuestro estudio ya que no se podrá contrastar si es correcto o no
ER	Cuando la herramienta devuelve un error controlado en la llamada

Tabla12 Posibles valores que puede tomar el campo OK_NOMBRE en la tabla TBENTIDA

Por tanto a la hora de evaluar el rendimiento de esta funcionalidad se tendrán en cuenta los registros con valor 'SI' contra el resto de entidades devueltas.

CATEGORIZACIÓN DE ENTIDAD

En este caso se evalúa tanto que el nombre haya sido reconocido correctamente como que la categoría asignada a la entidad sea también correcta. Dado que cada herramienta identifica sus propios tipos de entidad se ha creado la siguiente correspondencia para determinar el correcto análisis.

<table>
<tr><th colspan="7">TIPOS DE ENTIDAD</th></tr>
<tr><th>CORPUS</th><th>ALCHEMY API</th><th>TEXTALYTICS</th><th>OPENAMPLIFY</th><th>OPENCALAIS</th><th>SEMANTRIA</th><th>ZEMANTA</th></tr>
<tr><td rowspan="3">Person</td><td rowspan="3">Person</td><td rowspan="3">PERSON</td><td rowspan="3">Person</td><td rowspan="3">Person</td><td rowspan="3">Person</td><td>/people/person</td></tr>
<tr><td>/music/artist</td></tr>
<tr><td>/soccer/football_player</td></tr>
<tr><td rowspan="6">Location</td><td>Region</td><td rowspan="6">LOCATION</td><td rowspan="6">Location</td><td>City</td><td rowspan="6">Place</td><td rowspan="2">/location/location</td></tr>
<tr><td>Country</td><td>ProvinceOrState</td></tr>
<tr><td>GeographicFeature</td><td>Country</td><td rowspan="2">/location/continent</td></tr>
<tr><td>City</td><td>Continent</td></tr>
<tr><td>StateOrCounty</td><td>Region</td><td rowspan="2">/location/country</td></tr>
<tr><td>Continent</td><td>NaturalFeature</td></tr>
<tr><td rowspan="4">Organization</td><td rowspan="2">Organization</td><td rowspan="4">ORGANIZATION</td><td rowspan="4">Organization</td><td rowspan="2">Organization</td><td rowspan="4">Company</td><td>/organization/organization</td></tr>
<tr><td>/sports/sports_team</td></tr>
<tr><td rowspan="2">Company</td><td rowspan="2">Company</td><td>/business/company</td></tr>
<tr><td>/government/government</td></tr>
<tr><td>Other</td><td>Other</td><td>Other</td><td>Other</td><td>Other</td><td>Other</td><td>/government/politica</td></tr>
<tr><td>Other</td><td>Other</td><td>Other</td><td>Other</td><td>Other</td><td>Other</td><td>Other</td></tr>
</table>

Tabla13 Correspondencia de tipo de entidad entre los posibles valores devueltos por cada herramienta y los del CORPUS

En caso de que la correspondencia entre el valor del CORPUS y el devuelto por la herramienta sea correcta se actualizará a 'SI' el campo OK_TIPO de la tabla TBENTIDA y a 'NO' en caso contrario. Será a partir de estos valores con los que se evaluará el rendimiento de esta funcionalidad.

4.2.2 Categorización temática del texto

Al igual que en el caso anterior los resultados devueltos varían en función de cada herramienta y por tanto se ha diseñado también una tabla de correspondencias entre las distintas posibilidades.

TOPICS				
CORPUS	**ALCHEMY_API**	**TEXTALYTICS**	**OPENAMPLIFY**	**SEMANTRIA**
política	culture_politics	policía y justicia	Politics	Politics
	law_crime	política		Elections
		disturbios, conflictos y guerra		
entretenimiento	arts_entertainment	arte, cultura y espectáculos	Entertainment	Social Media
literatura			Literature	
cine	recreation	estilo de vida y tiempo libre	Arts	Art
música				
economía	business	economía, negocios y finanzas	Business	Banking
				Economics
				Business
tecnología	science_technology	ciencia y tecnología	Science	Software and Interne
				Technology
	computer_internet		Computers	Hardware
				Mobile Devices
			Internet	Investing
				Robotics
	gaming			Biotechnology
			Consumer electronics	Video Games
				Science
deportes	sports	deporte	Sports	Sports
fútbol			Health and Fitness	
otros	otros	otros	otros	otros

Tabla14 Correspondencia de categorías del texto entre los posibles valores devueltos por cada herramienta y los del CORPUS

En caso de que la correspondencia entre el valor del CORPUS y el devuelto por la herramienta sea correcta se actualizará a 'SI' el campo OK_TOPIC de la tabla TBTOPICS y a 'NO' en caso contrario. Será a partir de estos valores con los que se evaluará el rendimiento de esta funcionalidad.

4.2.3 Análisis de sentimiento del texto

POLARIDAD DEL SENTIMIENTO

En este primer punto se evaluará que el sentimiento general del texto sea positivo, negativo o nulo (P, N ó NONE). Cada herramienta devuelve y cuantifica el sentimiento de distinta forma y para evaluarlo formatearemos la respuesta de cada herramienta como se indica en la siguiente tabla.

	PARÁMETRO	VALORPOL
OPENAMPLIFY	Mean = 0	NONE
	Mean > 0	P
	Mean < 0	N
SEMANTRIA	Sentiment = 0	NONE
	Sentiment > 0	P
	Sentiment < 0	N
TEXTALYTICS	Scoretag =NEU	NONE
	Scoretag = P ó P+	P
	Scoretag = N ó N+	N

Tabla15 Campo de la respuesta XML de cada herramienta que se utilizará para determinar la polaridad del sentimiento

En caso de que la correspondencia entre el valor del CORPUS y el calculado a partir de la respuesta de la herramienta sea correcta se actualizará a 'SI' el campo OK_VALORPOL de la tabla TBSENTIM y a 'NO' en caso contrario. Será a partir de estos valores con los que se evaluará el rendimiento de esta funcionalidad.

GRADO DE ACUERDO

Como se ha explicado anteriormente se trata de evaluar el grado de acuerdo del autor del texto con lo expuesto y se cuantifica en función de la cantidad de los sentimientos enfrentados (positivos y negativos) que se encuentran en el texto. Los posibles valores pueden ser 'AGREEMENT' en caso de que todo el texto siga una misma línea de sentimiento o 'DISAGREEMENT' en caso de que exista un enfrentamiento.

Para determinar el grado de acuerdo o desacuerdo a partir de lo devuelto por cada herramienta se enfrentarán las expresiones positivas y negativas recuperadas y en caso de que el sumatorio no se corresponda con el sentimiento general del texto se considerará que se está en desacuerdo con lo expuesto y en caso contrario de acuerdo.

Solo se evaluará el grado de acuerdo de los tweets cuyo sentimiento general haya sido correcto. En caso de que la correspondencia entre el valor del CORPUS y el calculado a partir de la respuesta de la herramienta sea correcta se actualizará a 'SI' el campo OK_TIPOPOL de la tabla TBSENTIM y a 'NO' en caso contrario. Será a partir de estos valores con los que se evaluará el rendimiento de esta funcionalidad.

4.3 RESULTADOS DEL ANÁLISIS

Como se ha explicado anteriormente evaluaremos el rendimiento de cada herramienta calculando la precisión, exhaustividad y su valor-F. Las fórmulas para calcularlos serían las siguientes.

PRECISION	$P = \frac{\lvert\{\text{documento s relevantes}\} \cap \{\text{documento s recuperado s}\}\rvert}{\lvert\{\text{documento s recuperado s}\}\rvert}$
EXHAUSTIVIDAD	$E = \frac{\lvert\{\text{documento s relevantes}\} \cap \{\text{documento s recuperado s}\}\rvert}{\lvert\{\text{documento s relevantes}\}\rvert}$
VALOR-F	$F = 2 \cdot \frac{\text{Precisión} \cdot \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}}$

Tabla16 Fórmulas para calcular las métricas de evaluación del rendimiento de las herramientas

Después de haber realizado las pruebas y para evaluar los resultados lo primero que tenemos que tener en cuenta es la diferencia entre ambos corpus. Para las pruebas de extracción de entidades nombradas se han utilizado artículos periodísticos, es decir, textos relativamente largos con información sobre lo que están tratando y escritos en un lenguaje formal. Para el resto de pruebas se han utilizado *tweets*, de pocos caracteres, sin un contexto claro y en ocasiones con un lenguaje poco formal.

Por tanto se tendrá en cuenta que cualquier herramienta de análisis de texto va a trabajar con mayor facilidad y obtendrá mejores resultados con los artículos periodísticos que con los *tweets*. Ambas facetas son interesantes de estudiar ya que la segunda, como se ha explicado, aunque mucho más compleja es el formato en que se encuentran gran parte de la información actual en internet y los medios sociales.

Teniendo en cuenta estas premisas pasamos a evaluar los resultados.

4.3.1 Evaluación reconocimiento y clasificación de entidades nombradas

HERRAMIENTAS	TOTAL CORPUS	TOTAL RECUPERADO	TOTAL RELEVANTES	P	E	F
OPENAMPLIFY	2584	813	206	25.34%	7.97%	12.13%
SEMANTRIA	2584	999	914	91.49%	35.37%	51.02%
ALCHEMY API	2584	1361	1092	80.24%	42.26%	55.36%
TEXTALYTICS	2584	2344	2125	90.66%	82.24%	86.24%
OPENCALAIS	2584	1151	996	86.53%	38.54%	53.33%
ZEMANTA	2584	842	643	76.37%	24.88%	37.54%

Tabla17 Resultados reconocimiento de entidades nombradas en artículos

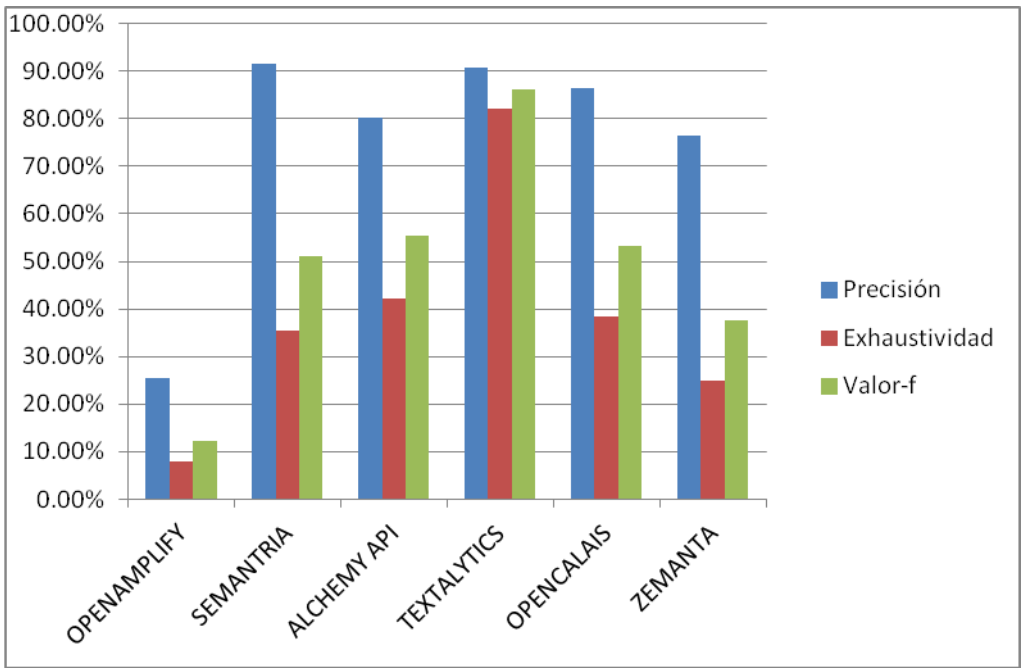


Gráfico5 Representación gráfica de los resultados del reconocimiento de entidades nombradas en artículos

En este primer apartado destaca la herramienta de Textalytics sobre las demás que es capaz de detectar cerca del 82 % de las entidades nombradas con una precisión superior al 91%. En segunda línea Semantria, Alchemi api y Opencalais obtienen también un buen rendimiento en este apartado. Detrás quedan Zemanta y sobre todo Openamplify, la cual solo llega a detectar el 7,97 % de las entidades

.

HERRAMIENTAS	TIPO ENTIDAD	TOTAL CORPUS	TOTAL RECUPERA	TOTAL RELEVANT	P	E	F
OPENAMPLIFY	Todos	2584	813	172	21.16%	6.66%	10.13%
	Personas	1027	350	107	30.57%	10.42%	15.54%
	Localizac.	584	326	55	16.87%	9.42%	12.09%
	Organiz.	611	31	7	22.58%	1.15%	2.18%
	Otros	362	106	3	2.83%	0.83%	1.28%
SEMANTRIA	Todos	2584	999	804	80.48%	31.11%	44.88%
	Personas	1027	518	452	87.26%	44.01%	58.51%
	Localizac.	584	437	328	75.06%	56.16%	64.25%
	Organiz.	611	27	22	81.48%	3.60%	6.90%
	Otros	362	17	2	11.76%	0.55%	1.06%
ALCHEMY API	Todos	2584	1361	977	71.79%	37.81%	49.53%
	Personas	1027	721	570	79.06%	55.50%	65.22%
	Localizac.	584	379	289	76.25%	49.49%	60.02%
	Organiz.	611	211	104	49.29%	17.02%	25.30%
	Otros	362	50	14	28.00%	3.87%	6.80%
TEXTALYTICS	Todos	2584	2344	1584	67.58%	61.30%	64.29%
	Personas	1027	815	732	89.82%	71.28%	79.48%
	Localizac.	584	532	417	78.38%	71.40%	74.73%
	Organiz.	611	367	292	79.56%	47.79%	59.71%
	Otros	362	630	143	22.70%	39.50%	28.83%
OPENCALAIS	Todos	2584	1151	924	80.28%	35.76%	49.48%
	Personas	1027	472	442	93.64%	43.04%	58.97%
	Localizac.	584	439	332	75.63%	56.85%	64.91%
	Organiz.	611	240	150	62.50%	24.55%	35.25%
	Otros	362	0	0	-	0.00%	-
ZEMANTA	Todos	2584	842	522	62.00%	20.20%	30.47%
	Personas	1027	280	230	82.14%	22.40%	35.20%
	Localizac.	584	300	210	70.00%	35.96%	47.51%
	Organiz.	611	119	62	52.10%	10.15%	16.99%
	Otros	362	143	20	13.99%	5.52%	7.92%

Tabla18 Resultados reconocimiento y categorización de entidades nombradas en artículos

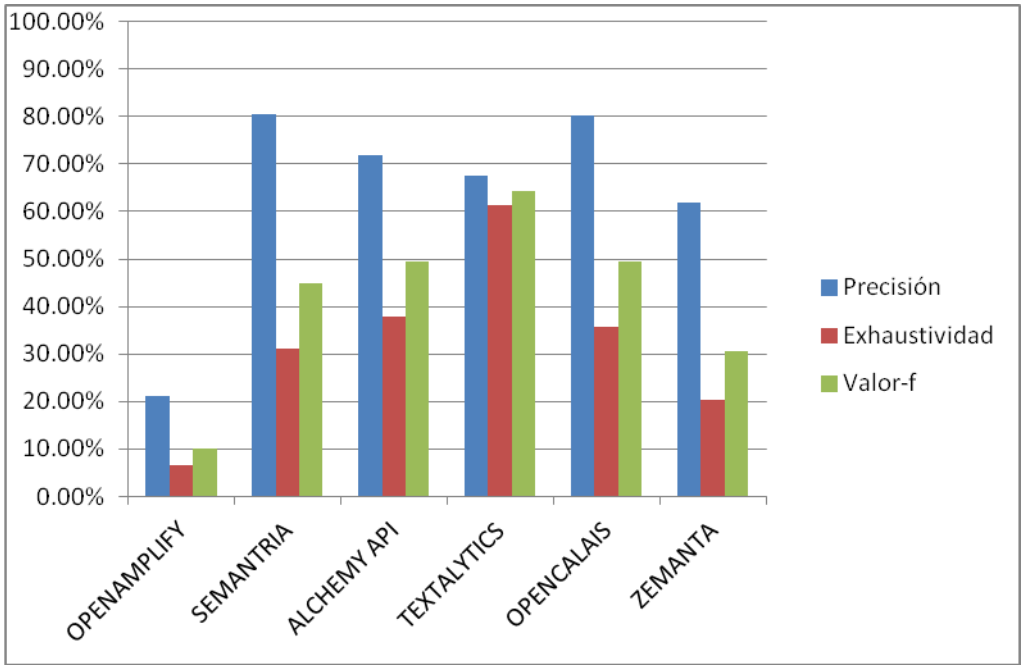


Gráfico6 Representación gráfica de los resultados de categorización de todas las entidades en artículos

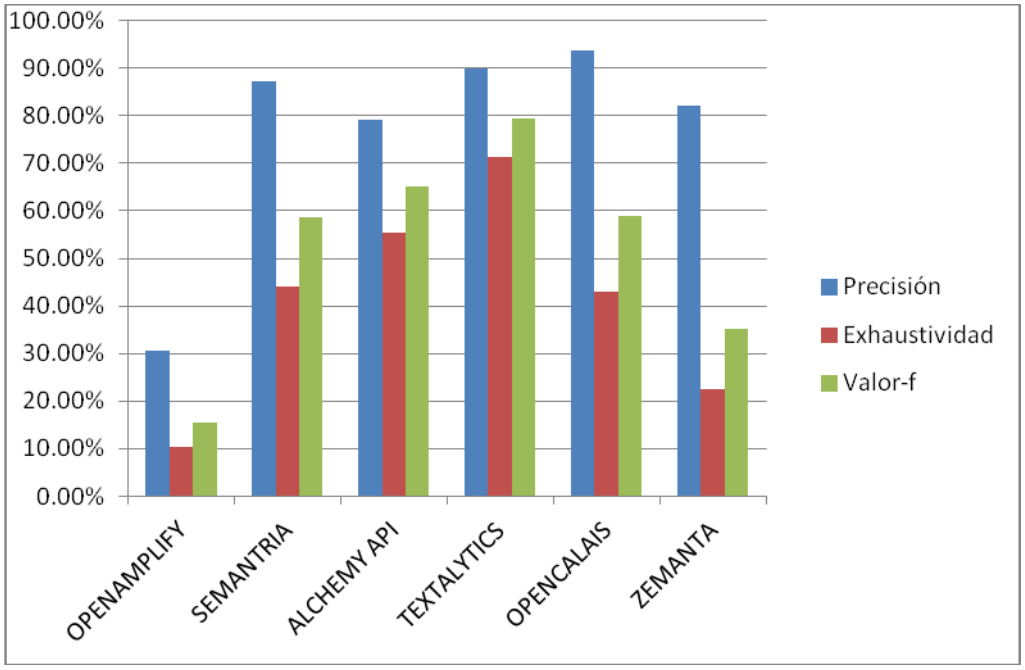


Gráfico7 Representación gráfica de los resultados de categorización de entidades en artículos para personas

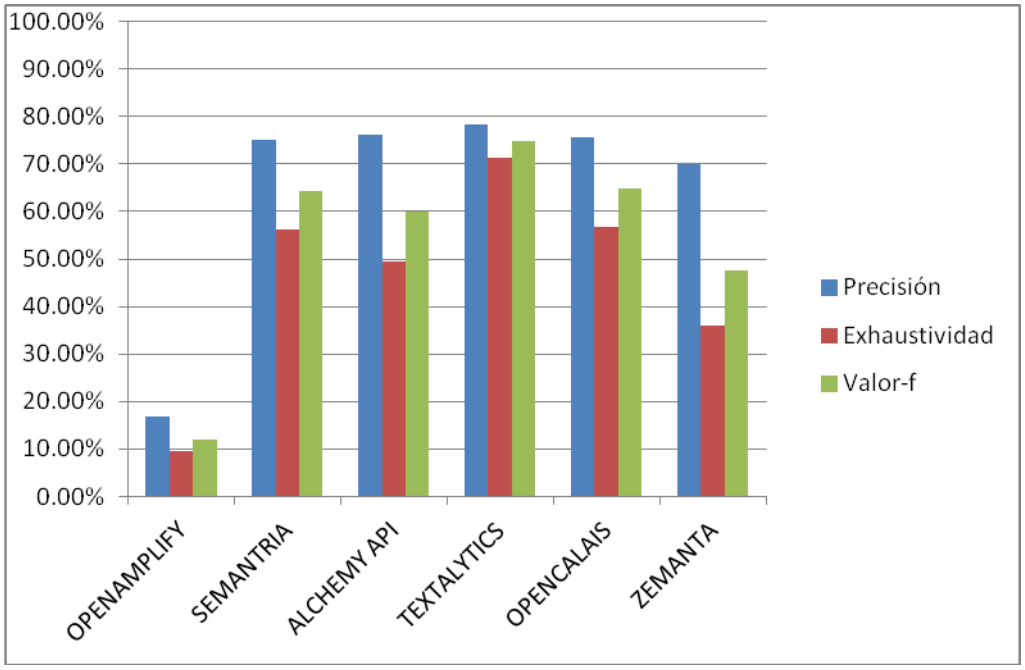


Gráfico8 Representación gráfica de los resultados de categorización de entidades en artículos para localizaciones

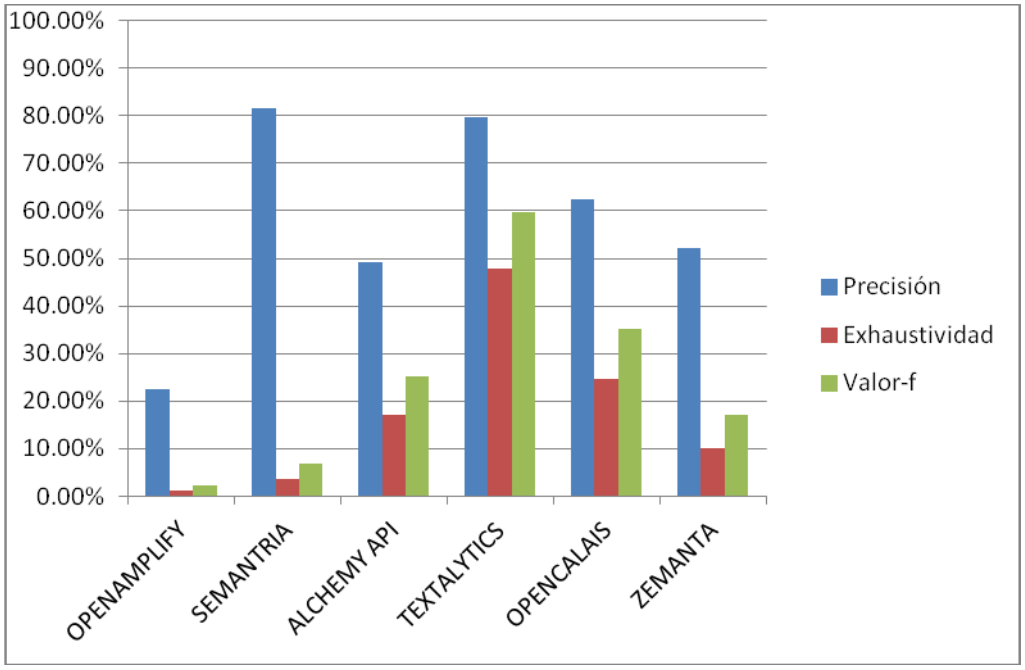


Gráfico9 Representación gráfica de los resultados de categorización de entidades en artículos para organizaciones

En esta segunda parte en la que se evalúa la categorización de las entidades anteriormente recuperadas la herramienta Textalytics sigue destacando sobre las demás seguida de Semantria, Alchemy api y Opencalais. Openamplify vuelve a mostrar el peor rendimiento.

Cabe destacar que en general las cifras son mejores al detectar personas y localizaciones. Esto es debido a que su identificación no depende completamente del idioma.

HERRAMIENTAS	TOTAL CORPUS	TOTAL RECUPERADO	TOTAL RELEVANTES	P	E	F
OPENAMPLIFY	5341	5083	673	13.24%	12.60%	12.91%
SEMANTRIA	5341	2013	908	45.11%	17.00%	24.69%
ALCHEMY API	5341	4342	1090	25.10%	20.41%	22.51%
TEXTALYTICS	5341	11542	4178	36.20%	78.23%	49.49%
OPENCALAIS	5341	2124	846	39.83%	15.84%	22.67%
ZEMANTA	5341	3806	1261	33.13%	23.61%	27.57%

Tabla19 Resultados reconocimiento de entidades nombradas en twitter

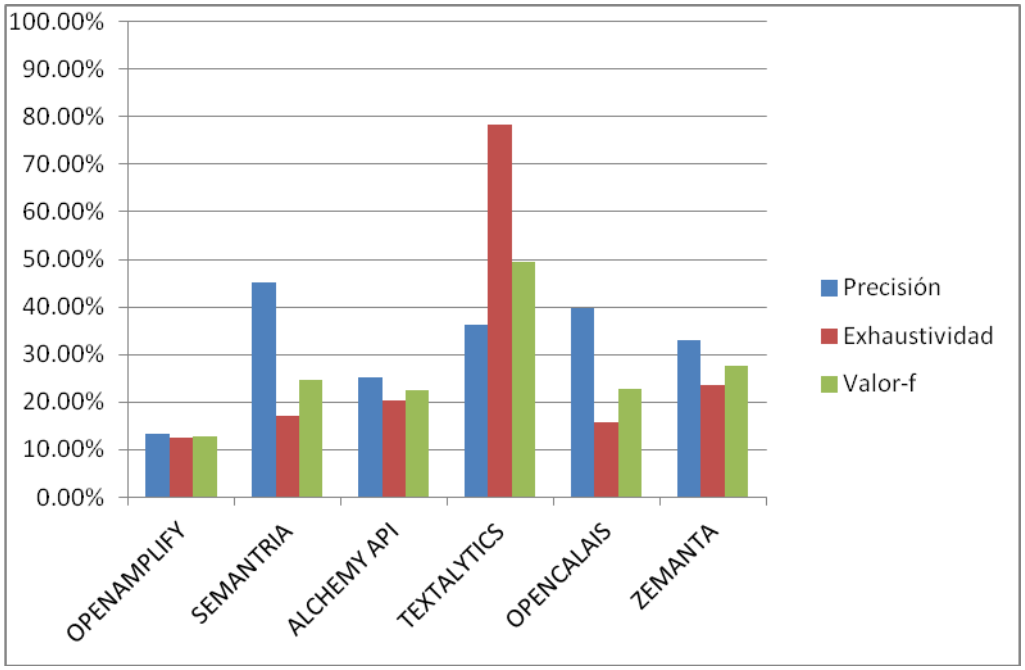


Gráfico10 Representación gráfica de los resultados del reconocimiento de entidades nombradas en twitter

Al evaluar el reconocimiento de entidades en tweets se observa un drástico empeoramiento de los resultados obtenidos respecto a los resultados obtenidos con

los artículos. Esto es lógico debido a la escasez de información y las particularidades del lenguaje de Twitter.

Aún así podemos observar como Textalytics destaca de nuevo, esta vez con una ventaja mucho mayor. Es Zemanta la segunda mejor herramienta en este apartado con resultados muy similares a los de Semantria, Alchemy api y Opencalais. Openamplify de nuevo vuelve a obtener unos resultados sensiblemente peores a los demás productos.

4.3.2 Evaluación categorización temática del texto

HERRAMIENTAS	TOTAL CORPUS	TOTAL RECUPERADO	TOTAL RELEVANTES	P	E	F
OPENAMPLIFY	9568	7276	2433	33.44%	25.43%	28.89%
SEMANTRIA	9568	7330	2458	33.53%	25.69%	29.09%
ALCHEMY API	9568	7182	3028	42.16%	31.65%	36.16%
TEXTALYTICS	9568	8396	3924	46.74%	41.01%	43.69%

Tabla20 Categorización temática del texto

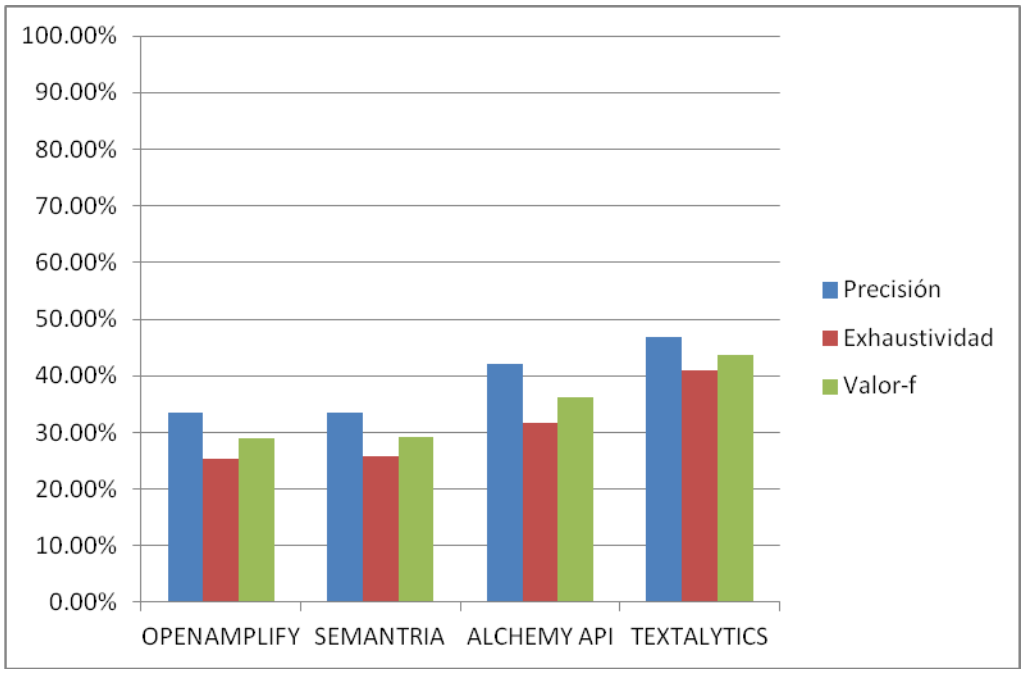


Gráfico11 Representación gráfica de los resultados de la categorización temática del texto

Textalytics sigue obteniendo los mejores resultados respecto a sus competidoras. Alchemy api estaría en segundo lugar. Cabe destacar que tanto Openamplify y Semantria no han sido capaces de clasificar un 85 % de los tweets y el grueso de sus aciertos se corresponde con los tweets sin una temática clara. Por tanto el resultado positivo que han podido tener estas herramientas es engañoso.

4.3.3 Evaluación análisis del sentimiento del texto

HERRAMIENTAS	TIPO SENTIMIEN	TOTAL CORPUS	TOTAL RECUPERA	TOTAL RELEVANT	P	E	F
OPENAMPLIFY	Todos	7218	7213	2454	34.02%	34.00%	34.01%
	Positivo	2783	246	127	51.63%	4.56%	8.39%
	Negativo	2124	926	342	36.93%	16.10%	22.43%
	Neutro	2311	6041	1985	32.86%	85.89%	47.53%
SEMANTRIA	Todos	7218	7218	3934	54.50%	54.50%	54.50%
	Positivo	2783	2000	1295	64.75%	46.53%	54.15%
	Negativo	2124	1393	946	67.91%	44.54%	53.80%
	Neutro	2311	3825	1693	44.26%	73.26%	55.18%
TEXTALYTICS	Todos	7218	7218	4637	64.24%	64.24%	64.24%
	Positivo	2783	2829	2000	70.70%	71.86%	71.28%
	Negativo	2124	1495	1091	72.98%	51.37%	60.29%
	Neutro	2311	2894	1546	53.42%	66.90%	59.40%

Tabla21 Resultados sentimiento general del texto

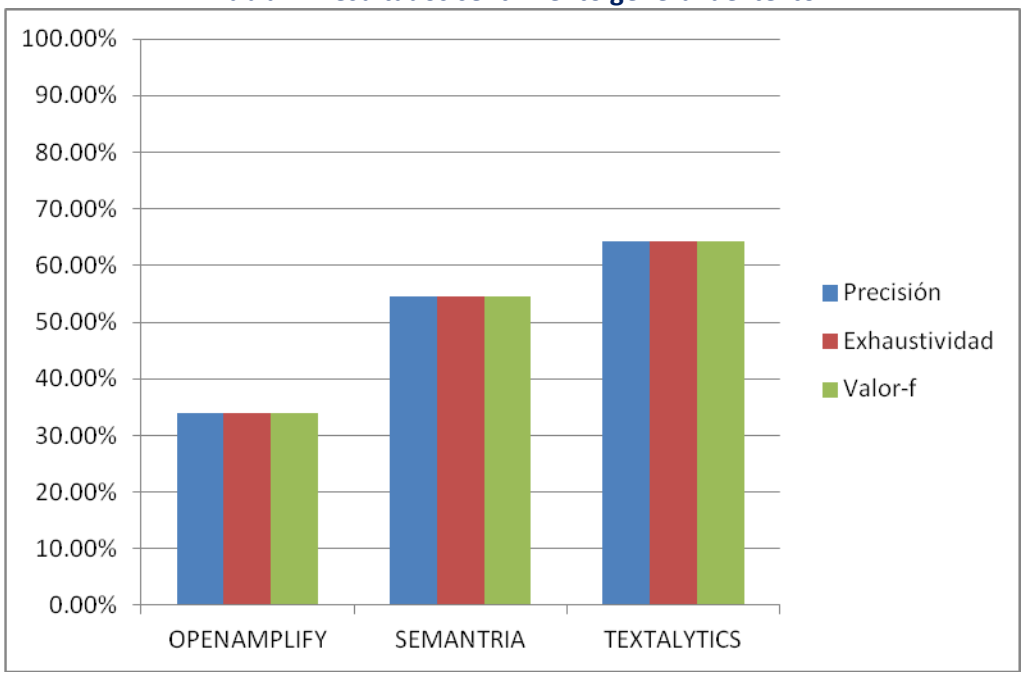


Gráfico12 Representación gráfica de los resultados del análisis del sentimiento del texto

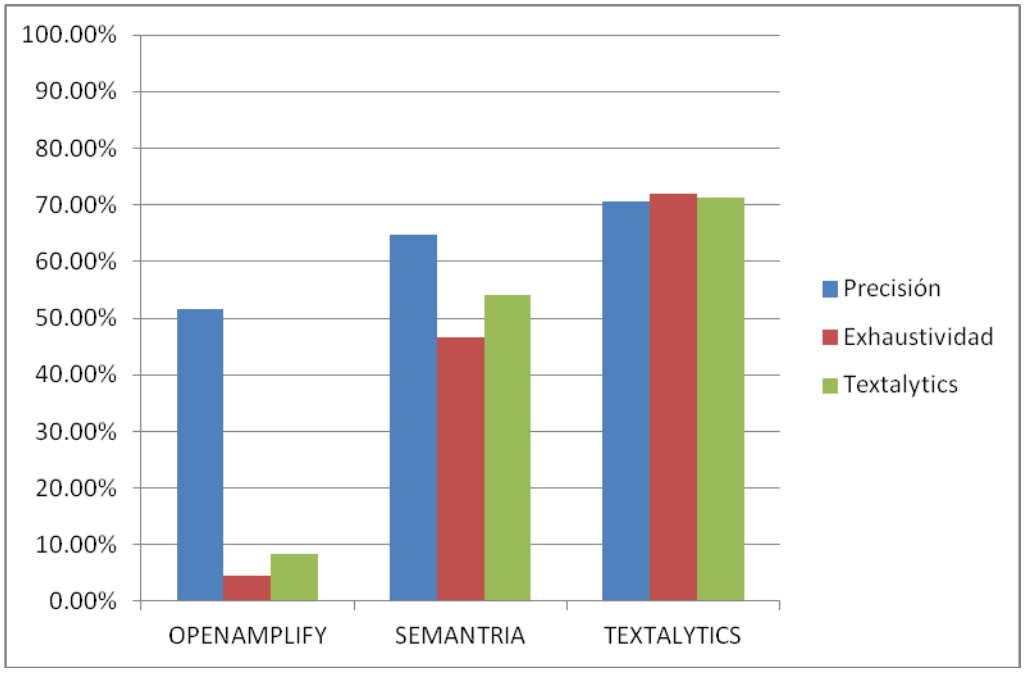


Gráfico13 Representación gráfica de los resultados del análisis del sentimiento positivo del texto

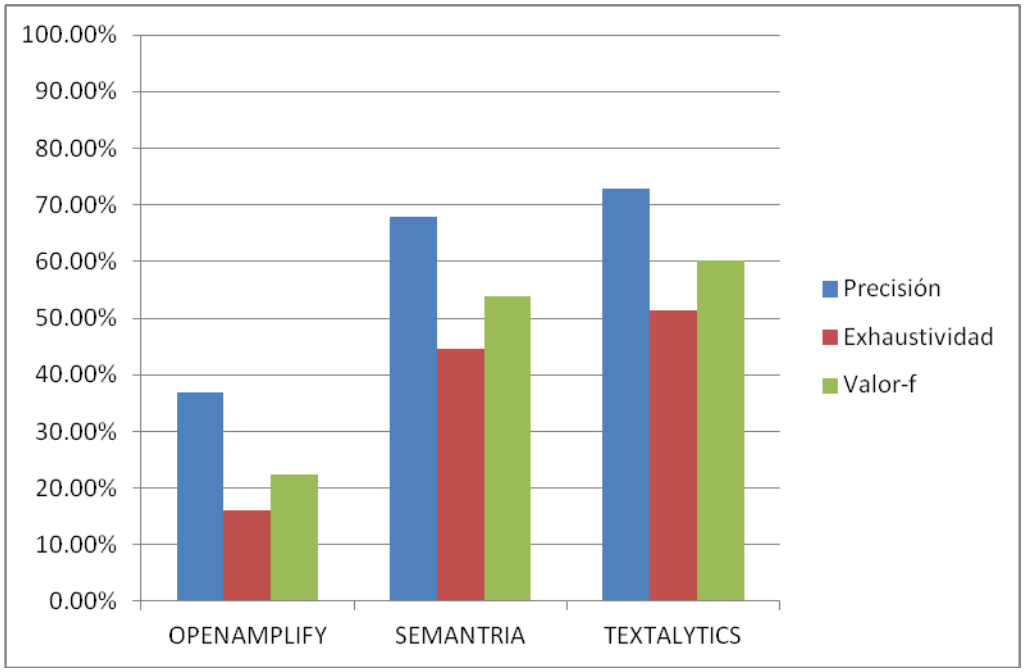


Gráfico14 Representación gráfica de los resultados del análisis del sentimiento negativo del texto

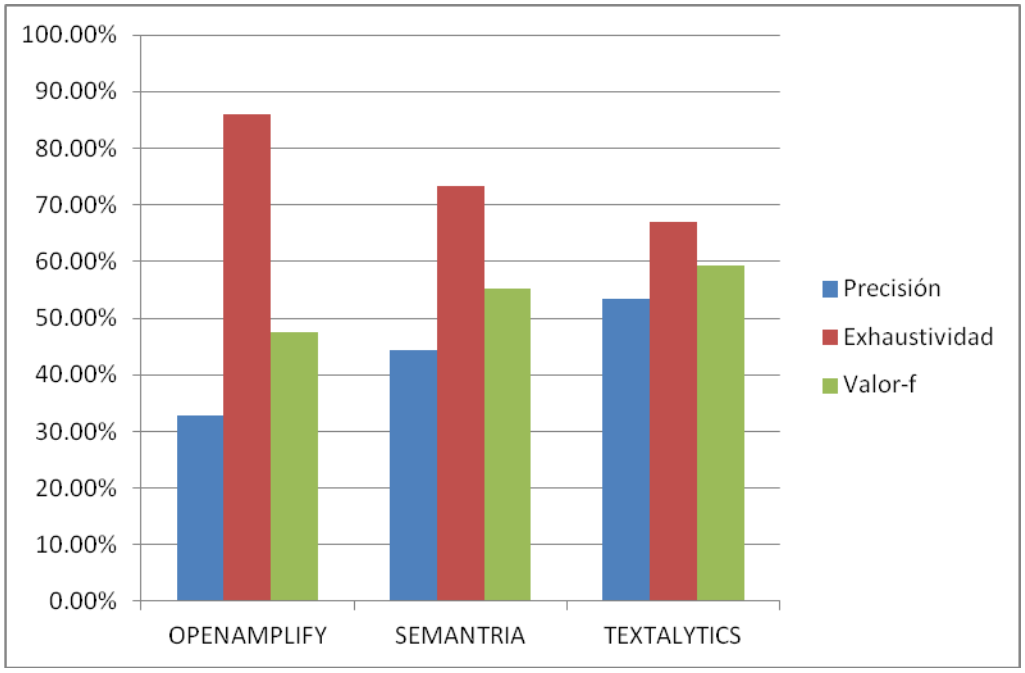


Gráfico15 Representación gráfica de los resultados del análisis del sentimiento neutral del texto

Textalytics vuelve a obtener los mejores resultados con cerca del 60 % de aciertos. Semantria obtiene también unos buenos resultados y es de nuevo Openamplify la que queda en último lugar.

HERRAMIENTAS	ACUERDO	TOTAL CORPUS	TOTAL RECUPERA	TOTAL RELEVANT	P	E	F
OPENAMPLIFY	7218	7218	7213	2160	29.95%	29.93%	29.94%
	6614	6614	7207	2158	29.95%	32.66%	31.24%
	604	604	6	2	33.33%	0.33%	0.66%
SEMANTRIA	7218	7218	7218	3666	50.79%	50.79%	50.79%
	6614	6614	7139	3659	51.25%	55.32%	53.21%
	604	604	79	7	8.86%	1.16%	2.05%
TEXTALYTICS	7218	7218	7218	4307	59.67%	59.67%	59.67%
	6614	6614	6411	4184	65.26%	63.26%	64.25%
	604	604	807	123	15.24%	20.36%	17.43%

Tabla22 Resultados grado de acuerdo del autor con lo expuesto

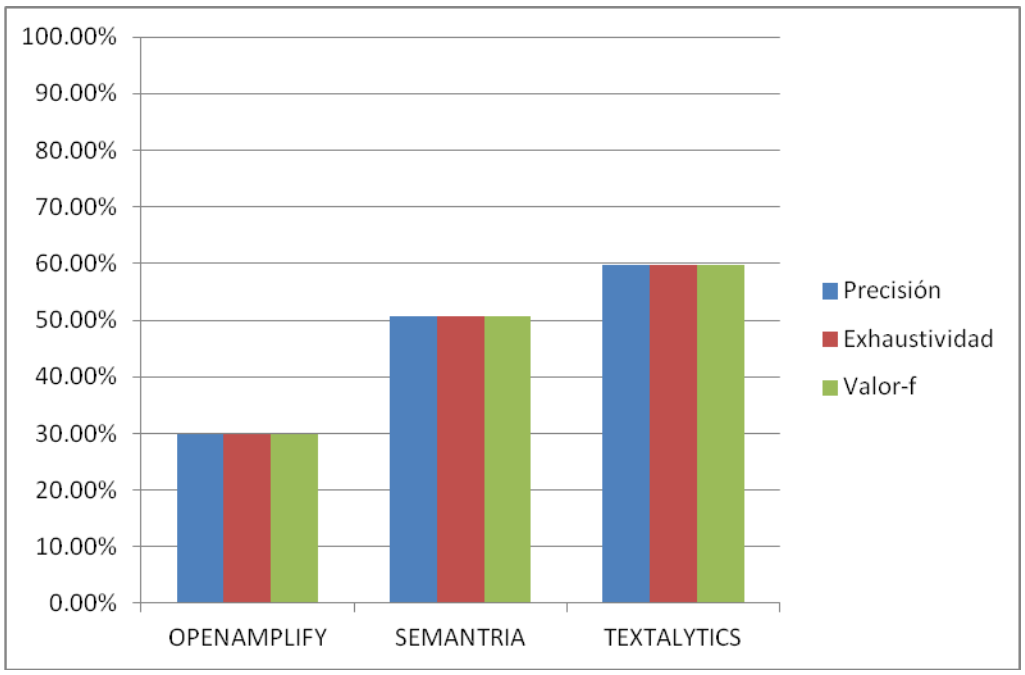


Gráfico16 Representación gráfica de los resultados del análisis de la polaridad del sentimiento del texto

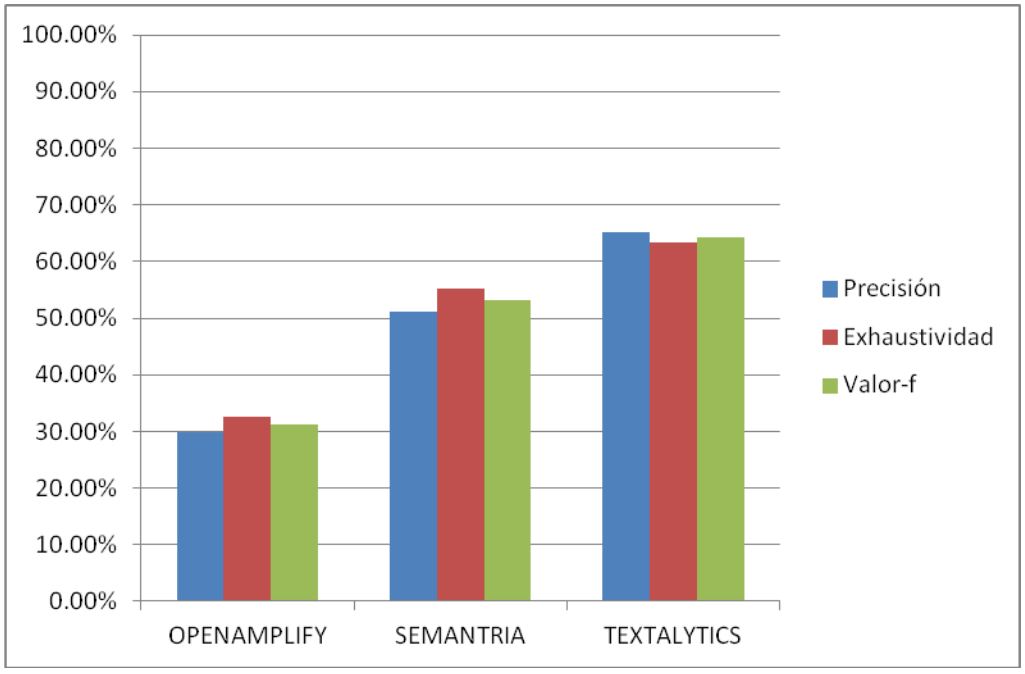


Gráfico17 Representación gráfica de los resultados del análisis de la polaridad en acuerdo con el sentimiento del texto

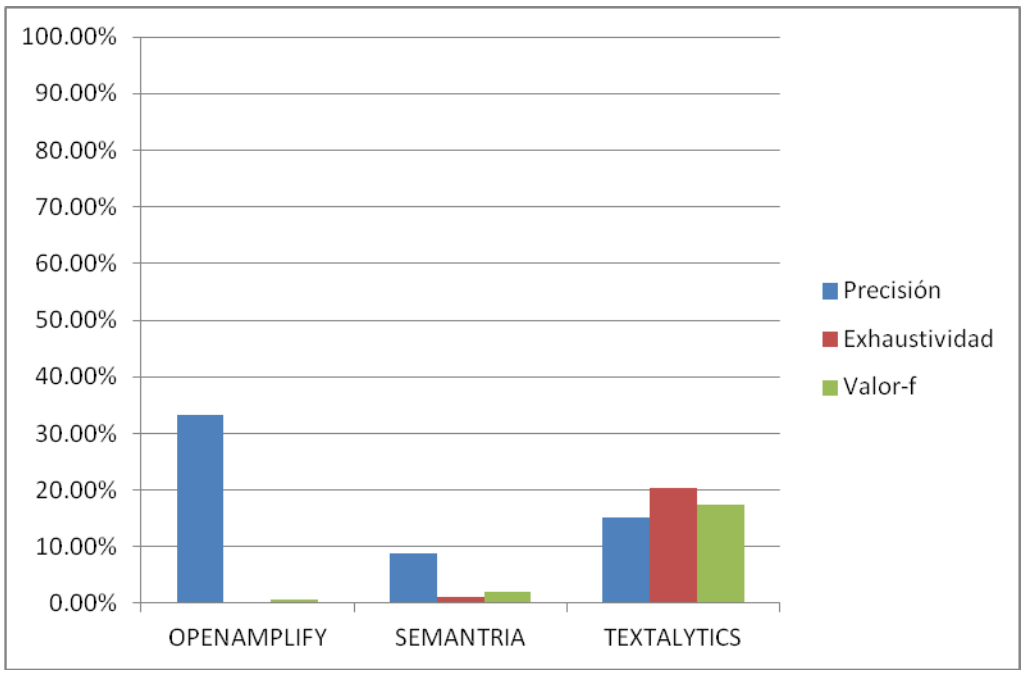


Gráfico18 Representación gráfica de los resultados del análisis de la polaridad en desacuerdo con el sentimiento del texto

Textalytics vuelve a destacar en este último apartado, en el que ni Openamplify ni Semantria son capaces de obtener resultados representativos tal y como podemos observar al ver los malos resultados obtenidos al identificar el grado de desacuerdo.

5 CONCLUSIONES

El primer punto a evaluar es en este apartado es la elaboración de la relación de empresas y productos dedicadas al análisis de textos desde un punto de vista funcional. En la siguiente página se incluye una tabla en la que se resumen las distintas perspectivas evaluadas.

En cuanto a las pruebas técnicas observando los resultados obtenidos podemos llegar a las siguientes conclusiones:

- La herramienta de análisis de texto que sin duda mejor trabaja con el castellano es Textalytics. No solo es la que mejores cifras ha conseguido en el análisis con una amplia diferencia con la siguiente sino que también es la que más funcionalidades ofrece para este idioma. Cabe destacar que es la única empresa española de las estudiadas.

  De las empresas extranjeras la que más funcionalidades oferta obteniendo mejores resultados sería Semantria, estando pese a ello bastante por detrás de la anterior.

  Otros productos como Alchemy Api y OpenCalais que afirman trabajar perfectamente en castellano también han obtenido buenos resultados, sobretodo en el reconocimiento de personas y localizaciones pero bajando el rendimiento en otras facetas en las que es necesario aplicar otras técnicas gramaticales y lingüísticas al texto.

  En un último lugar quedarían Zemanta y Openamplify. Ninguna de ellas está preparada para trabajar con el castellano. Aún así Zemanta puede resultar útil en el reconocimiento de entidades ya que ha demostrado ser capaz de obtener resultados correctos, sobretodo en twitter.

- Otra conclusión interesante a la que llegamos es que a pesar de que muchas herramientas de análisis de texto han obtenido resultados significativos estos no se han podido replicar con tanto éxito en textos cortos. Las particularidades del lenguaje que tienen este tipo de textos que hoy en día cada vez son más normales como *tweets*, correos electrónicos o mensajes de texto hacen que incluso el ser humano no sea capaz de interpretar y extraer su significado real.

	Modelo de Negocio	Funcionalidades de su webservice	Idiomas con los que trabaja	Fuentes de datos con los que opera	Acuerdo de nivel de servicio	Documentación para usuarios
OPENAMPLIFY	Además de su web service de análisis de texto cuentan con otras herramientas comerciales para empresas o dispositivos móviles	• Análisis de topics. • Análisis de acciones. • Análisis de estilo • Análisis demográfico • Análisis de búsqueda	Inglés	XML, RDF, RDFA, CSV, DART, OAS, JSON, HTML	1000 transacciones diarias gratuitas	Manuales y ejemplos de código en inglés. Cunta también con recursos como blogs y foros donde los usuarios pueden compartir opiniones, pedir soporte y estar informados sobre las novedades del sector
SEMANTRIA	Se centra en la comercialización de los accesos a su web service. Para los no desarrolladores proponen un complemento para Excel que permite configurar la herramienta, analizar textos e interpretar los resultados desde una hoja de cálculo	• Puntuación del sentimiento del texto. • Síntesis de las ideas del documento. • Extracción tema del documento. • Extracción de entidades. • Tipificación de entidades extraídas. • Extracción de las relaciones entre las entidades. • Categorización consulta basada • Extracción de opiniones. • Clasificación según consultas. • Reconocimiento de facetas y atributos entre varios textos. • Reconocimiento de temas en común entre varios textos. • Reconocimiento de entidades en común entre varios textos. • Clasificación según consultas entre varios textos.	• Inglés (máximo de funcionalidades) • Francés • Alemán • Portugués • Chino • Castellano	XML, JSOP	Un total de 10000 transacciones gratuitas por usuario	Manuales y ejemplos en inglés. Cuenta también con una serie de SDKs para configurar y utilizar la herramienta en las siguientes plataformas: C++, Java, .Net, PHP, Python, Ruby y Javascript
ALCHEMY API	Se centra en la comercialización de los accesos a su web service. Pone a disposición de sus usuarios algunas herrmientas visuales muy sencillas para realizar estas llamadas.	• Extracción de entidades. • Análisis de sentimientos. • Extracción de palabras clave. • Identificación de conceptos. • Extracción de relaciones. • Categorización del texto. • Identificación del autor del texto. • Identificación del idioma (entre más de 97). • Síntesis del texto. • Identificación de fuentes o canales web (ATOM ó RSS feeds)	• Inglés (máximo de funcionalidades) • Francés • Alemán • Italiano • Portugués • Ruso • Sueco • Castellano	XML, JSON, RDF	30000 llamadas diarias gratuitas para usos no comerciales y 1000 para usos comerciales	Cuenta con un blog y manuales en inglés. También pone a disposición de los desarrolladores una serie de SDKs en las siguientes plataformas: Python, PHP, Node.js, Ruby, Android OS, Perl, Java, C/C++ y .NET

Tabla23 Relación de empresas estudiadas parte 1

	Modelo de Negocio	Funcionalidades de su webservice	Idiomas con los que trabaja	Fuentes de datos con los que opera	Acuerdo de nivel de servicio	Documentación para usuarios
DAEDALUS	Además del licenciamiento de su web service ofertan otras herramientas comerciales de procesamiento y análisis de textos	• Identifica conceptos como personas, lugares, empresas, fechas, direcciones, citas y relaciones entre ellos. • Clasifica textos de acuerdo a categorías existentes (ej.: IPTC) o a clases definidas por el usuario. • Identifica si un texto es irónico o si expresa una opinión positiva o negativa, según el contexto. • Averigua automáticamente el idioma de un texto. En más de 20 idiomas. • Detecta errores ortográficos, gramaticales y de estilo en varios idiomas. • Consulta distintas fuentes Linked Data a partir de un identificador. • Averigua si un usuario es una empresa o una persona, su edad y su género. • Analiza morfosintácticamente un texto, proporcionando lemas y etiquetas de discurso. En varios idiomas. • Capacidades de reconocimiento de voz y locutor para cualquier aplicación. Aplica la tecnología lingüística de Textalytics sobre contenidos de audio y vídeo.	• Inglés • Castellano	XML, JSON	500000 créditos mensuales gratuitas	Documentación tanto en Castellano como en Inglés. Oferta SDKs para desarrolladores en PHP, Java y Python.
CALAIS	Solo comercializa el acceso a su herramienta. Cuenta con menos funcionalidades que el resto pero fomenta la libre utilización de su software promocionando herramientas que se basan en la versión gratuita de su servicio	• Extracción y categorización de entidades • Extracción de temas de la entidad • Extracción de relaciones entre entidades • Evaluación de relevancia de las entidades • Categorización del texto	• Inglés (máximo de funcionalidades) • Francés • Castellano	XML, JSON, RDF, N3	50000 transacciones diarias gratuitas	Cuenta con un blog y manuales en inglés. Además reunen y ponen a disposición de cualquier usuario herramientas, aplicaciones, pluggins o SDKs desarrollados por ellos mismos o por terceros.
ZEMANTA	Comercializa el acceso a su web service, el cual está enfocado principalmente a editores, anunciantes o bloggers para realizar sugerencias de contenido a partir de la información publicada	• Reconocimiento y clasificación de las entidades nombradas. A partir de esta información propone las siguientes sugerencias. • Relevancia de las entidades. • Categorización del textos. • Reconoce nombres de sitios web para que con un sólo clic se incorpore un hipervínculo en la primera aparición del nombre. • A partir de las entidades reconocidas, su clasificación y de la categorización del texto: - Sugiere fotografías públicas y también fotos de tu usuario de Flickr para ilustrar el artículo. - Propone noticias y artículos de distintos medios para que se incluyan al final de la entrada a modo de artículos relacionados - Propone palabras clave (tags) para las entradas.	Inglés	XML, JSON, WNJSON, RDFXML	1000 transacciones diarias gratuitas	Manuales y ejemplos de código en inglés, aunque es la más escasa de todas las estudiadas

Tabla24 Relación de empresas estudiadas parte 2

6 GLOSARIO DE TÉRMINOS

Acuerdo de nivel de servicio, *Service Level Agreement*: Contrato escrito entre un proveedor de servicio y su cliente con objeto de fijar el nivel acordado para la calidad de dicho servicio.

API (*Aplication Programming Interface*): Interfaz de programación de aplicaciones. Es un conjunto de funciones y procedimientos que ofrece cierta biblioteca para ser utilizada por otro software como una capa de abstracción.

Aprendizaje semisupervisado: Clase de técnicas de aprendizaje automático que utiliza datos de entrenamiento tanto etiquetados como no etiquetados.

Análisis morfológico: El análisis de las palabras para extraer raíces, rasgos flexivos, unidades léxicas compuestas y otros fenómenos.

Análisis sintáctico: El análisis de la estructura sintáctica de la frase mediante una gramática de la lengua en cuestión.

ATOM: Estándar basado en HTTP y XML diseñado para la redifusión y la actualización de recursos web.

Back-end: Dentro de una aplicación se refiere a la parte del servidor que atiende las peticiones y que es transparente para el cliente.

Front-end: Dentro de una aplicación se refiere a la parte con la que interactúa el cliente.

HTTP (*Hypertext Transfer Protocol*): Protocolo usado en cada transacción de la *World Wide Web*.

Lingüística computacional: Campo multidisciplinar de la lingüística y la informática que utiliza la informática para estudiar y tratar el lenguaje humano.

LSA (*Latent semantic analysis*): Técnica de procesamiento natural del lenguaje que analiza relaciones entre conjuntos de documentos y los términos que contienen.

Máquinas de vectores de soporte: Son un conjunto de algoritmos de aprendizaje supervisado relacionados con problemas de clasificación y regresión.

Marketing: Proceso que comprende la identificación de necesidades y deseos del mercado objetivo, la formulación de objetivos orientados al consumidor, la construcción de estrategias que creen un valor superior, la implantación de relaciones con el consumidor y la retención del valor del consumidor para alcanzar beneficios.

Medios sociales, *social media*: Son plataformas de comunicación en lÍnea donde el contenido es creado por los propios usuarios mediante el uso de las tecnologías de la Web 2.0.

Minería de datos: Campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

MUC (*Message Understanding Conferences*): Conferencias organizadas y financiadas por DARPA para fomentar el desarrollo de nuevos y mejores métodos de extracción de información

NER (*Named-entity recognition*): Reconocimiento de entidades nombradas.

Precisión: Métrica empleada en la medida del rendimiento de sistemas de búsqueda y recuperación de información. La precisión sería la fracción de instancias recuperadas que son relevantes.

Recall: Métrica empleada en la medida del rendimiento de sistemas de búsqueda y recuperación de información. El *recall* o exhaustividad sería la fracción de instancias relevantes que han sido recuperadas.

REST (*Representational State Transfer*): técnica de arquitectura software para sistemas hipermedia.

RSS (*Really Simple Syndication*): Formato XML para sindicar o compartir contenido en la web. Se utiliza para difundir información actualizada frecuentemente a usuarios que se han suscrito a la fuente de contenidos.

SaaS (*Software as a Service*): Modelo de distribución de software donde el soporte lógico y los datos que maneja se alojan en servidores de una compañía de tecnologías de información y comunicación a los que se accede con un navegador web a través de Internet.

SDK (*Software Development Kit*): Kit de desarrollo de software. Son paquetes de software creadas para integrar una API en un lenguaje de programación concreto.

SOAP(*Simple Object Access Protocol*): Protocolo estándar que define cómo dos objetos en diferentes procesos pueden comunicarse por medio de intercambio de datos XML

Valor-F: Métrica empleada en la medida del rendimiento de sistemas de búsqueda y recuperación de información. El valor-f sería la media armónica de la precisión y el *recall* para ponderar como de lejanas se encuentran ambas.

XML (*eXtensible Markup Language*): Es un lenguaje de marcas utilizado para almacenar datos de forma legible.

7 BIBLIOGRAFÍA

7.1 EMPRESAS Y PRODUCTOS ESTUDIADOS

[1] OpenAmplify

http://www.openamplify.com/

[2] Semantria

https://semantria.com/

[3] AlchemyApi

http://www.alchemyapi.com

[4] Textalytics

https://textalytics.com/inicio

[5] OpenCalais

http://www.opencalais.com/

[6] Zemanta

http://www.zemanta.com/

7.2 OTRAS REFERENCIAS ELECTRÓNICAS

[11] http://clic.ub.edu/corpus/es/ancora

[8] http://cogcomp.cs.illinois.edu/papers/RatinovRo09.pdf

http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

[12] http://www.daedalus.es/TASS2012/corpus.php

http://www.datadrivenbiz.com/

[10] https://www.globalwebindex.net/

http://www.google.es/trends/

[7]http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

[9] http://mashable.com/

http://net-savvy.com/executive/

http://www.php.net

http://www.phpya.com.ar/

http://socialmediaanalysis.com/

http://socialmedialab.ca/

http://www.socialnetworking-weblog.com/

http://www.wampserver.com/

http://webknox.com/p/named-entity-definition

http://www.wikipedia.org/

7.3 REFERENCIAS DE ARTÍCULOS Y LIBROS

- Graham Wilcoch, "*Introduction to Linguistic Annotation and Text Analytics (Sythesis Lectures on Human Language Technologies)*", 2009.

- Seth Grimes, "*Text Analytics Overview: Technology, Solutions, Market*", 2011.

- Lev Tarinov, Dan Roth, " *Design Challenges and Misconceptions in Named Entity Recognition*", 2009.

- José Manuel Molina López, Jesús García Herrero, "*Técnicas de análisis de datos (Aplicaciones prácticas utilizando Microsoft Excel y Weka)*", 2006.

- David de Ugarte, "*El poder de las redes*", 2007.

- Alejandro Suárez Sánchez-Ocaña, "*Desnudando a Google*", 2012.

- María V. Rosas, Marcelo L. Errecalde, Paolo Rosso, "*Un análisis comparativo de estrategias para la categorización semántica de textos cortos*", 2010.

- Raquel Toribio, Paloma Martínez, César de Pablo-Sánchez, "*Evaluación de la extracción de entidades nombradas de OpenCalais en castellano*", 2010.

- Juan Diego Gómez Fierros, Azucena Montes Rendón, "*Comparativa entre herramientas para la extracción de entidades espaciales geográficas*".

- Francisco Javier Rufo Mendo, Anselmo Peñas Padilla, "*Clasificación de tweets multilingüe*", 2013.

ANEXO I: WAMPSERVER

Wampserver es el programa que se ha utilizado para la realización de este proyecto montando un servidor local en nuestro PC para probar nuestras páginas Web.

Se trata de un entorno de desarrollo web para Windows que permite crear aplicaciones web con Apache, PHP y bases de datos MySQL database. También incluye los gestores SQL PHPMyAdmin y SQLiteManager para simplificar el manejo y la administración de las bases de datos.

CARACTERÍSTICAS

Provee al desarrollador con los elementos necesarios para un servidor web:

- Un Sistema Operativo (Windows).
- Un manejador de base de datos (MySQL).
- Un software para servidor web (Apache).
- Un software de programación script Web (Generalmente PHP o Python)

Es completamente gratuito e incluye tanto las últimas versiones de Apache, PHP y MySQL como las anteriores.

UTILIDAD

El uso de WAMP permite servir páginas HTML a Internet, además de poder gestionar datos en ellas, al mismo tiempo WAMP proporciona lenguajes de programación para desarrollar aplicaciones Web.

LENGUAJES DE PROGRAMACIÓN

Wampserver soporta los siguientes lenguajes de programación:

- HTML, Javascritp, PHP, ASP, ASP.NET, JSP, Python, RUBY, APACHE.

ANEXO II: DICCIONARIO DE DATOS

DICCIONARIO DE DATOS

Tabla TBPRODUC: Productos y herramientas a estudiar

Descripción

Esta tabla contiene los nombres de las herramientas de análisis que se utilizarán en el estudio.

Estructura

Nombre	Descripción	Tipo	Long.	Dec.
PRODUCTO	Nombre reducido de la herramienta: - **CORPUS** - **OPENAMPLIFY** - **DAEDALUS** - **SEMANTRIA** - **OPENCALAIS** - **ALCHEMY_API** - **ZEMANTA**	Alfanumérico	20	
FECPROC	Fecha de procesamiento en el sistema	Timestamp	26	

Índices

Nombre del índice	Principal	Única	Nombre del campo	Orden	Nulos
PK_TBPRODUC	Sí	Sí	IDUSUARIO	Ascendente	No

Tabla TBTWEETS: Tweets

Descripción

En esta tabla se almacena la información original de los tweets provenientes de Twitter.

Estructura

Nombre	Descripción	Tipo	Long.	Dec.
IDTWEET	Identificador del tweet	Alfanumérico	20	
USUARIO	Identificador del usuario que publicó el tweet	Alfanumérico	20	
TEXTO	Texto del tweet	Alfanumérico	140	
IDIOMA	Idioma en el que está escrito el tweet	Alfanumérico	2	
FECHPUB	Fecha de publicación del tweet	Alfanumérico	20	
FECPROC	Fecha de procesamiento en el sistema	Timestamp	26	

Índices

Nombre del índice	Principal	Única	Nombre del campo	Orden	Nulos
PK_TBTWEETS	Sí	Sí	IDTWEET	Ascendente	No

Tabla TBARTICU: Artículos

Descripción

En esta tabla se almacena la información original de los artículos a analizar

Estructura

Nombre	Descripción	Tipo	Long.	Dec.
IDARTICU	Identificador del artículo	Alfanumérico	20	
TEXTO	Texto del artículo	Alfanumérico	6000	
FECPROC	Fecha de procesamiento en el sistema	Timestamp	26	

Índices

Nombre del índice	Principal	Única	Nombre del campo	Orden	Nulos
PK_TBARTICU	Sí	Sí	IDARTICU	Ascendente	No

Tabla TBSENTIM: Sentimiento

Descripción

Se almacenará de cada texto analizado la connotación o la polaridad resultantes del análisis realizado por cada herramienta.

Estructura

Nombre	Descripción	Tipo	Long.	Dec.
ID	Identificador del texto dentro del sistema	Alfanumérico	20	
ORIGEN	Origen del texto analizado. En el sistema actual los posibles valores serían **ARTICULO** ó **TWITTER**	Alfanumérico	20	
PRODUCTO	Producto o herramienta que ha realizado el análisis de la entidad (Consultar tabla de productos)	Alfanumérico	20	
VALORPOL	Valor de la polaridad del sentimiento - **P** - Positiva - **N** - Negativa - **NONE** - Sin polaridad	Alfanumérico	4	
TIPOPOL	Tipo de polaridad o grado de acuerdo - **AGREEMENT** - Acuerdo - **DISAGREEMENT** - Desacuerdo	Alfanumérico	12	
SENTIMIENTO	Valor numérico devuelto por cada herramienta con el que cuantifica cada herramienta	Numérico	7	4
OK_VALORPOL	Indicador de valor de la polaridad correcto en el análisis	Alfanumérico	2	
OK_TIPOPOL	Indicador de tipo de la polaridad correcto en el análisis	Alfanumérico	2	
FECPROC	Fecha de procesamiento en el sistema	Timestamp	26	

Índices

Nombre del índice	Principal	Única	Nombre del campo	Orden	Nulos
PK_TBSENTIM	Sí	Sí	ID	Ascendente	No
			ORIGEN	Ascendente	No
			PRODUCTO	Ascendente	No

Tabla TBENTIDA: Entidades

Descripción

Se almacenarán las entidades extraídas de cada texto analizado junto con su tipo correspondiente.

Estructura

Nombre	Descripción	Tipo	Long.	Dec.
ID	Identificador del texto analizado dentro del sistema	Alfanumérico	20	
ORIGEN	Origen del texto analizado. En el sistema actual los posibles valores serían ARTICULO ó TWITTER	Alfanumérico	20	
ENTIDAD	Entidad extraída en el análisis	Alfanumérico	50	
PRODUCTO	Producto o herramienta que ha realizado el análisis de la entidad (Consultar tabla de productos)	Alfanumérico	20	
TIPOENT	Tipo de entidad. Los posibles valores vendrán dados por cada una de las herramientas	Alfanumérico	20	
OK_NOMBRE	Indicador de nombre de entidad correcto en el análisis	Alfanumérico	2	
OK_TIPO	Indicador de tipo de entidad correcto en el análisis	Alfanumérico	2	
FECPROC	Fecha de procesamiento en el sistema	Timestamp	26	

Índices

Nombre del índice	Principal	Única	Nombre del campo	Orden	Nulos
PK_TBSENTIM	Sí	Sí	ID	Ascendente	No
			ORIGEN	Ascendente	No
			ENTIDAD	Ascendente	No
			PRODUCTO	Ascendente	No

Tabla TBTOPICS: Topics

Descripción

Contiene la información de la categorización textual detectada por cada herramienta.

Estructura

Nombre	Descripción	Tipo	Long.	Dec.
ID	Identificador del tweet	Alfanumérico	20	
ORIGEN	Origen del texto analizado. En el sistema actual los posibles valores serían **ARTICULO** ó **TWITTER**	Alfanumérico	20	
TOPIC	Topic extraído. Los posibles valores vendrán dados por cada una de las herramientas	Alfanumérico	100	
PRODUCTO	Producto o herramienta que ha realizado el análisis del topic (Consultar tabla de productos)	Alfanumérico	20	
OK_TOPIC	Indicador de topic correcto en el análisis	Alfanumérico	2	
FECPROC	Fecha de procesamiento en el sistema	Timestamp	26	

Índices

Nombre del índice	Principal	Única	Nombre del campo	Orden	Nulos
PK_TBTOPICS	Sí	Sí	ID	Ascendente	No
			ORIGEN	Ascendente	No
			TOPIC	Ascendente	No
			PRODUCTO	Ascendente	No

Tabla TBESTADI: Estadísticas con el resultado de los análisis

Descripción

Esta tabla almacena los resultados del análisis realizado por cada herramienta.

Estructura

Nombre	Descripción	Tipo	Long.	Dec.
PRODUCTO	Producto o herramienta evaluada (Consultar tabla de productos)	Alfanumérico	20	
FECHA	Fecha del análisis	Alfanumérico	10	
TIPOANALI	En función de lo que se esté evaluando tomará los siguientes valores: - **ENTIDAD**: Análisis de las entidades extraídas de cada texto - **TOPIC**: Análisis de los temas de los que trata el texto - **SENTIMIENTO**: Análisis del sentimiento de cada texto	Alfanumérico	15	
SUBTIPOANALI	En función del campo TIPOANALI puede tomar los siguientes valores: - **NOMBRE**: Nombre de la entidad extraída - **TIPOENT**: Tipo de entidad extraída - **TOPIC**: Temas - **VALORPOL**: Valor de la polaridad del sentimiento - **TIPOPOL**: Tipo de la polaridad del sentimiento	Alfanumérico	15	
TOTAL_ANALISIS	Total de resultados devueltos por la herramienta	Numérico	12	
TOTAL_CORRECTOS	Del total de resultados devueltos los que son correctos	Numérico	12	
TOTAL_REGULAR	Del total de resultados devueltos los que no son del todo correctos	Numérico	12	
TOTAL_INCORRE	Del total de resultados devueltos los que no son correctos	Numérico	12	
TOTAL_CORPUS	Total de resultados en el CORPUS	Numérico	12	
TOTAL_ERROR	Total de resultados en los que se ha devuelto error	Numérico	12	
TOTAL_NC	Total de resultados que no son evaluados por el CORPUS	Numérico	12	
FECPROC	Fecha de procesamiento en el sistema	Timestamp	26	

Índices

Nombre del índice	Principal	Única	Nombre del campo	Orden	Nulos
PK_TBUSUARI	Sí	Sí	PRODUCTO	Ascendente	No
			FECHA	Ascendente	No
			TIPOANALI	Ascendente	No
			SUBTIPOANALI	Ascendente	No

EJEMPLO DE TWEET EN LA BASE DE DATOS

TBTWEETS

IDTWEET	USUARIO	TEXTO	IDIOMA	FECPUB
142569992418897922	javiersolana	3.Tercer acto.Acuerdo Bancos centrales para dar un poco de tiempo.	es	2011-12-02T12:45:09

TBTOPICS

ID	ORIGEN	TOPIC	PRODUCTO	OK_TOPIC
142569992418897922	TWITTER	economía	CORPUS	SI
142569992418897922	TWITTER	Banking	SEMANTRIA	SI
142569992418897922	TWITTER	economía, negocios y finanzas - macroeconomía - bancos centrales	DAEDALUS	SI
142569992418897922	TWITTER	unknown	ALCHEMY_API	NO
142569992418897922	TWITTER	unknown	OPENAMPLIFY	NO

TBSENTIM

ID	ORIGEN	VALORPOL	TIPOPOL	PRODUCTO	OK_VALORPOL	OK_TIPOPOL
142569992418897922	TWITTER	P	AGREEMENT	CORPUS	SI	SI
142569992418897922	TWITTER	P	AGREEMENT	DAEDALUS	SI	SI
142569992418897922	TWITTER	NONE	AGREEMENT	OPENAMPLIFY	NO	NO
142569992418897922	TWITTER	NONE	AGREEMENT	SEMANTRIA	NO	NO

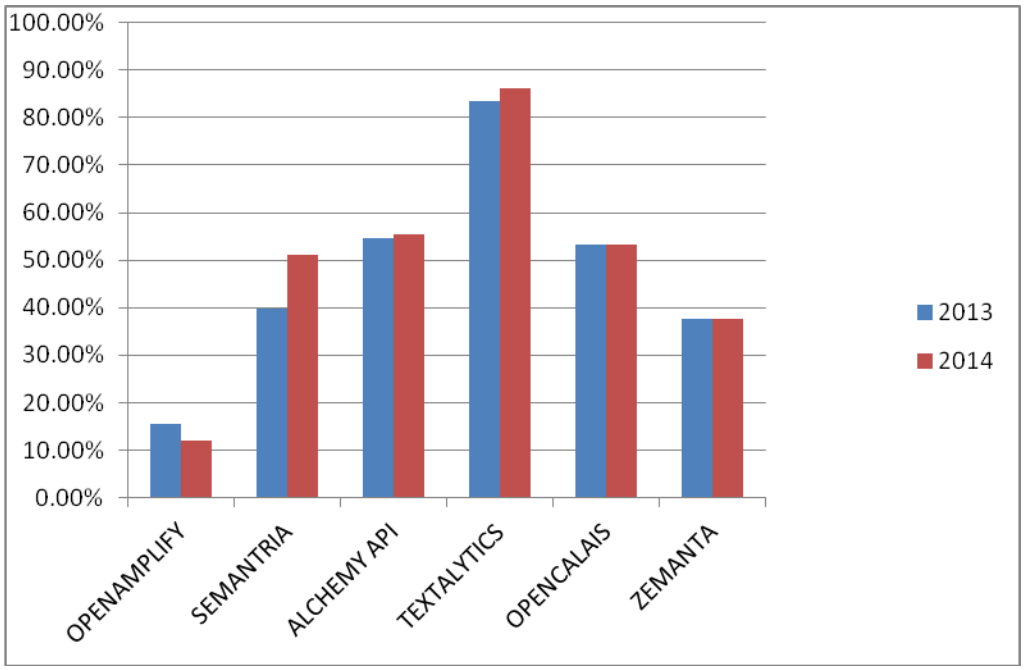
TBESTADI

PRODUCTO	TIPOANALI	SUBTIPOANALI	TOTAL_ANALISIS	TOTAL_CORRECTOS	TOTAL_INCORREC
SEMANTRIA	TOPIC	CATEGORIA	1	1	0
DAEDALUS	TOPIC	CATEGORIA	1	1	0
OPENAMPLIFY	TOPIC	CATEGORIA	1	0	1
ALCHEMY_API	TOPIC	CATEGORIA	1	0	1
SEMANTRIA	SENTIMIENTO	TIPOPOL	1	0	1
SEMANTRIA	SENTIMIENTO	VALORPOL	0	0	0
DAEDALUS	SENTIMIENTO	TIPOPOL	1	1	0
DAEDALUS	SENTIMIENTO	VALORPOL	1	1	0
OPENAMPLIFY	SENTIMIENTO	TIPOPOL	1	0	1
OPENAMPLIFY	SENTIMIENTO	VALORPOL	0	0	0

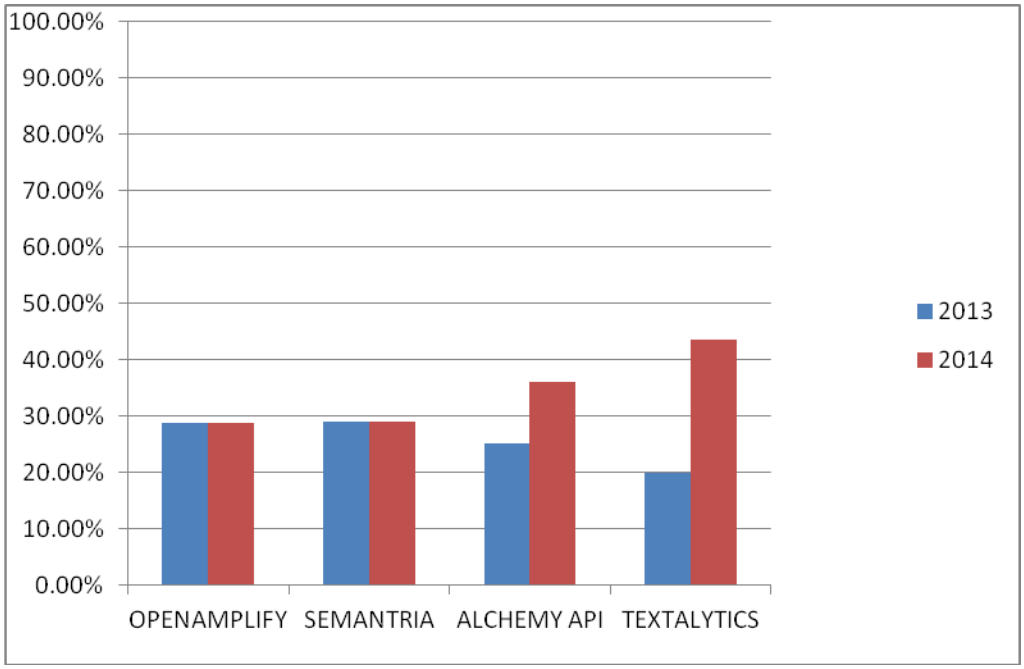
Cabe destacar la base de datos ha sido diseñada para ser reutilizable y poder incluir nuevos análisis, herramientas o fuentes de datos en el futuro sin impactar a los datos ya almacenados.

ANEXO III: COMPARACIÓN RESULTADOS 2013/2014

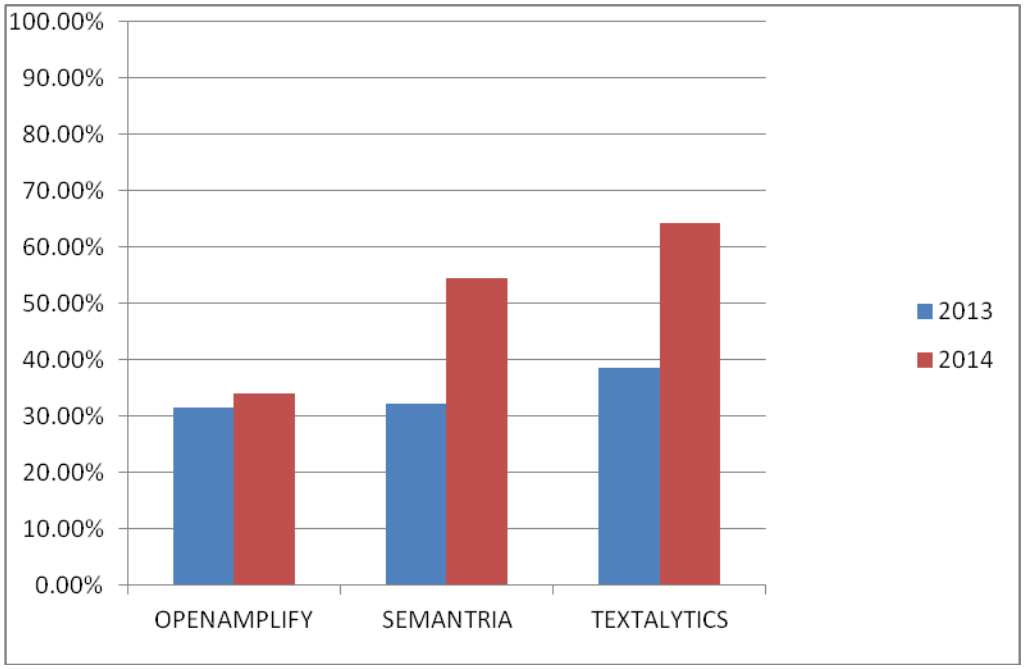
A continuación se comparan los resultados obtenidos en las pruebas de las herramientas en Agosto del 2014 con los conseguidos en Abril del 2013 para observar la evolución de las mismas y las facetas en las que han podido mejorar u empeorar.



Comparación del Valor-f de las herramientas obtenido en 2013 y 2014 para el reconocimiento de entidades nombradas en artículos



Comparación del Valor-f de las herramientas obtenido en 2013 y 2014 para la categorización de la temática del texto en tweets



Comparación del Valor-f de las herramientas obtenido en 2013 y 2014 para el análisis del sentimiento de tweets