Trabajo Fin de Master: Análisis de reputación en redes sociales en ámbitos locales

Trabajo Fin de Master

Javier Porras Gómez

Trabajo de investigación para el

Master en Lenguajes y Sistemas Informáticos

Universidad Nacional de Educación a Distancia

Dirigido por el

Prof. Dr. Enrique Amigó

Septiembre 2019

Resumen

El presente trabajo pretende estudiar diferentes criterios de búsqueda de términos de filtrado para la extracción y clasificación de mensajes en redes sociales pertenecientes a un mismo contexto político. En concreto se analizarán funciones de pesado, técnicas de análisis lingüístico y, finalmente, heurística

Además, pretende analizar la calidad de los datos obtenidos al filtrar mensajes en redes sociales aplicando dichos términos y que, además, hayan sido geolocalización tomando como caso de estudio la ciudad de Talavera de la Reina (Toledo) en periodo pre-electoral municipal.

Por tratarse de temática concreta y delimitada en un radio pequeño, la densidad de los datos termina siendo realmente baja, de manera que el uso de funciones de pesado y técnicas de análisis lingüístico terminan siendo ineficaces en la mayoría de los casos.

Agradecimientos

Agradecer a mi tutor Enrique Amigó su ayuda, su disposición, constante orientación y confianza depositada en mí y en este trabajo, así como a todo el equipo docente de la Universidad.

Agradecer además a mi familia por la paciencia y comprensión.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1. Introducción

El paradigma de la comunicación se revoluciona y varía desde el día en que se comienzan a popularizar las redes sociales ya que el alcance de los mensajes o campañas individuales son inmediatos y multitudinarios con una inversión extremadamente baja o incluso nula, de manera que las redes sociales se convierten en herramientas de marketing tremendamente eficaces, además proporcionan un feedback casi en tiempo real sobre los temas, campañas o cuentas personales o profesionales por parte del resto del público.

A partir de esta revolución social surgen, en primera instancia, innumerables formas de explotar las redes como vía de entrada para campañas empresariales, sociales o políticas de comunicación y marketing y, posteriormente, las consecuentes formas de analizar el alcance, impacto y eficacia de dichas campañas.

Para tales fines se desarrollan aplicaciones dentro de las propias redes sociales y paralelas a estas donde se puede hacer un seguimiento exhaustivo de cierta información que circula en la red, analizar comentarios, reacciones de los usuarios que son, en definitiva, clientes o clientes potenciales y que marcan la tendencia que las compañías pueden intentar seguir para mantener o incrementar la aceptación de sus productos. Con esta premisa surge el marketing y la monitorización de redes sociales, los cuales pueden ayudar a elevar la audiencia y convertir a personas interesadas, en clientes potenciales de una forma significativa con una inversión muy inferior a la que se hace en medios de comunicación convencionales, así como detectar y prevenir situaciones adversas contra usuarios o marcas.

Toda esta información fluye por la red con tantos usuarios y competidores vigilando la actividad y la información existente resulta interesante poder automatizar los procesos de recuperación y análisis de mensajes, así como la notificación de información relevante que surge a partir de estos.

En este proyecto se centra en la red social Twitter, que permitía la emisión de mensajes de un máximo de 140 caracteres en sus inicios, aunque ahora se extiende hasta 280 pudiendo incluir fotos y vídeos con una población de 326millones de usuarios activos y hasta 65 millones de mensajes diarios.

1.1. Justificación

Existen multitud de herramientas que se dedican a la recuperación de información vertidas en redes sociales, también existen multitud de aplicaciones que analizan las campañas y el alcance de las mismas, así como herramientas que notifican en base a ciertos parámetros la actividad en redes sociales. Cabe esperar que exista una fuerte demanda de aplicaciones que puedan rastrear cualquier tipo de campaña, tanto las propias como las de los competidores, así como en análisis del alcance y feedback de las mismas, pudiendo discernir el carácter de la interacción de los usuarios y, por último, notificar cuando pueda surgir un tema a partir de éstas o incluso independientes a las mismas. Estas aplicaciones deben tener interés para las compañías, de manera que puedan sacar partido de las mismas en beneficio propio.

Actualmente, las soluciones que tratan este tipo de problema necesitan ingentes cantidades de información, en contraposición del presente que se basa en información muy específica y geolocalizada vertida por usuarios y cuentas de microbolgs con influencia y temáticas de ámbito local.

Por estos motivos se plantea el presente trabajo que trata de monitorizar campañas, cuentas de usuario y su alcance en redes sociales, pudiendo comprobar el carácter de la información y notificar al usuario en base a la importancia, fortaleza o debilidad de los mismos.

1.2. Objetivos

El objetivo del proyecto consiste en investigar y analizar diferentes técnicas de recuperación de información y sus resultados filtrando diferentes cuentas de Twitter para monitorizar información que pueda afectar a la reputación de entidades de ámbito local.

Se pretende recuperar mensajes en las redes sociales en base a temáticas o cuentas de índole política, delimitando un radio de ámbito local para, posteriormente poder clasificarlos. Con estos datos se pretende crear un conjunto de documentos anotado y polarizado manualmente como documento de control para poder estudiar la calidad de las clasificaciones de las distintas aproximaciones comprobando, además, su calidad respecto a otras técnicas de recuperación y clasificación de información. Además, estos

datos podrán compararse con los resultados electorales de la campaña local sucedida durante el presente estudio.

Los objetivos, finalmente, serán:

- Determinar si la desambiguación de términos es favorable usando técnicas existentes o si, por el contrario, y dado las características de los datos, es preferible que la idiosincrasia de cada localización, en concreto, la ciudad o provincia donde se circunscribe el problema, sea determinante en la calidad de la información obtenida por medio de un sistema de filtrado.
- Evaluar una aproximación por geolocalización general por medio del uso de los metadatos de cada mensaje, es decir, determinar si la predicción de la geolocalización de los mensajes en redes sociales resulta una técnica que mejore el rendimiento del sistema.
- Estudiar diferentes aproximaciones al rankeo automatizado en base a criterior predefinidos tras el filtrado inicial y analizar los resultados comparándolos con documentos etiquetados manualmente.

Capítulo 2. Contextualización

En este capítulo se presenta una breve introducción a las redes sociales desde sus inicios hasta su monitorización pasando por la comunicación, el márquetin o el análisis de redes sociales, así como los distintos conjuntos de datos que serán planteados posteriormente, su procesado, su extracción y clasificación.

2.1. Redes Sociales

El término red, proviene del latín rete, que define una estructura con un denominado patrón. A grandes rasgos, encontramos diferentes tipologías de redes como familiares, laborales, informáticas, eléctricas, sociales, etc.

En concreto, las redes sociales son estructuras sociales compuestas de grupos de personas, las cuales están conectadas por uno o varios tipos de relaciones, tales como amistad, parentesco, intereses comunes o que comparten conocimientos.

Pueden definirse como un conjunto de entidades vinculados entre sí por medio de relaciones estructurados de la siguiente manera:

- **Entidades**: individuos, grupos, organizaciones y sociedades sujetos de los vínculos.
- **Vínculos**: enlaces relacionales entre pares de entidades pudiendo ser de tipo personal, de transferencia de recursos, asociaciones, interacciones, movilidad, conexiones, etc.
- **Conexión**: cada par de actores y el lazo que los une.
- **Malla**: espacio existente entre cada entidad y las entidades con las que se relaciona por medio de vínculos.
- **Subgrupo**: conjunto de atores y los lazos que los une
- **Grupo**: modelado de relaciones entre sistemas o subgrupos de actores.

Según el público en que se conectan encontramos redes sociales de tipo **generalista u horizontales**, donde cualquier entidad puede formar parte de una malla e interactuar con el resto de entidades, **abiertas**, donde se comparte información de cualquier índole y **temáticas o verticales**, que van dirigidas a un público concreto o especializado. Además, según la forma en que se crean los vínculos encontramos redes sociales **simétricas**,

donde las entidades objeto de un vínculo deben aceptar dicho vínculo para que este sea efectivo, o **asimétricas**, donde esa aceptación no es necesaria por ambas partes.

2.1.1. Inicios

El análisis de grupos sociales surge tras de la II Guerra Mundial a partir de la observación de pautas culturales o sociales fijas orientándose en conceptos dirigidos a la adaptación y adaptabilidad, interesándose también por interacciones iniciadas por individuos con pautas similares generadas por iniciativa propia.

Este interés se desarrolla simultáneamente con el interés por las sociedades modernas y el conjunto de estructuras sociales cada vez más variadas y complejas.

John Barnes describe en su estudio el sistema social de una pequeña comunidad noruega donde pudo distinguir tres agrupaciones sociales con estructuración territorial, industrial y parentelar o de amistad con vínculos cambiantes y sin organización ni coordinación estables, apareciendo así una primera definición de red:

> *"La imagen que tengo es de un conjunto de puntos algunos de los cuales están unidos por líneas. Los puntos de la imagen son personas o a veces grupos, y las líneas indican que individuos interactúan mutuamente. Podemos pensar claro está, que el conjunto de la vida social genera una red de este tipo"*[1]
>
> (Whitten, 1974)

Una de sus conclusiones es que entre la sociedad tradicional y la moderna hay diferencias en la malla de la red. En las tradicionales la red es más densa ya que hay más vínculos parentelares o de amistad, mientras que en las modernas los "agujeros" en la red son mayores ya que los vínculos no son tan estrechos.

[1] WHITTEN, NORMAN E., JR., and ALVIN W. WOLFE. 1974. "Network analysis," en Handbook of social and cultural anthropology. Edited by J. Honigmann. Chicago: Rand McNally

2.1.2. Sociología

En la sociología encontramos en un extremo el análisis de redes limitado a una metodología, técnica estadística o matemática, mientras, en el otro extremo, encontramos análisis que hacen un uso normativo o ético.

Los patrones de vinculación que afectan a la conducta social han ayudado a definir la estructura social. Aunque la sociología se ha propuesto siempre estudiar las organizaciones entre individuos, la mayor parte de los trabajos se centran en la perspectiva individual y atributiva. La mejor forma de estudiar una organización social es analizar los patrones de vínculos que unen a sus miembros profundizando en sus estructuras.

"La Red Social Informal (RSI) corresponde a una organización pluralista y descentralizada, y es un sistema de organización cuyo lazo solidario no se construye ni a través de la coacción ni a través de la culpa. El encuentro y el entretenimiento como sistema es el que reemplaza en la red a la coacción y a la culpa como formas de nexo."[2]

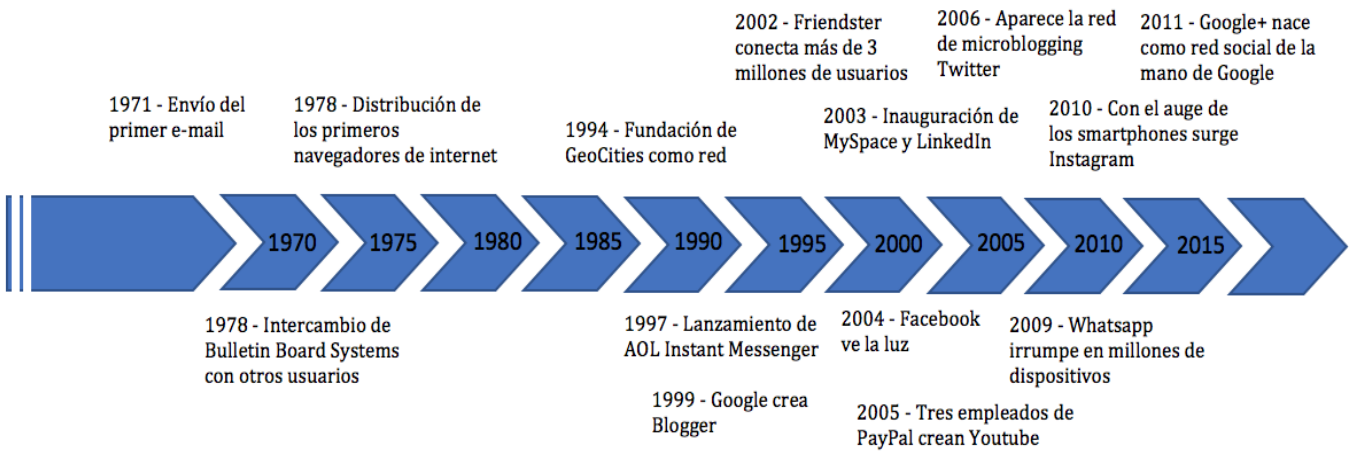
(Motta, 1995)

2.1.3. Evolución de las Redes Sociales

Las redes de interacción social se han convertido en uno de los elementos de Internet más difundidos, ofrecen a sus usuarios un lugar común para desarrollar comunicaciones constantes.

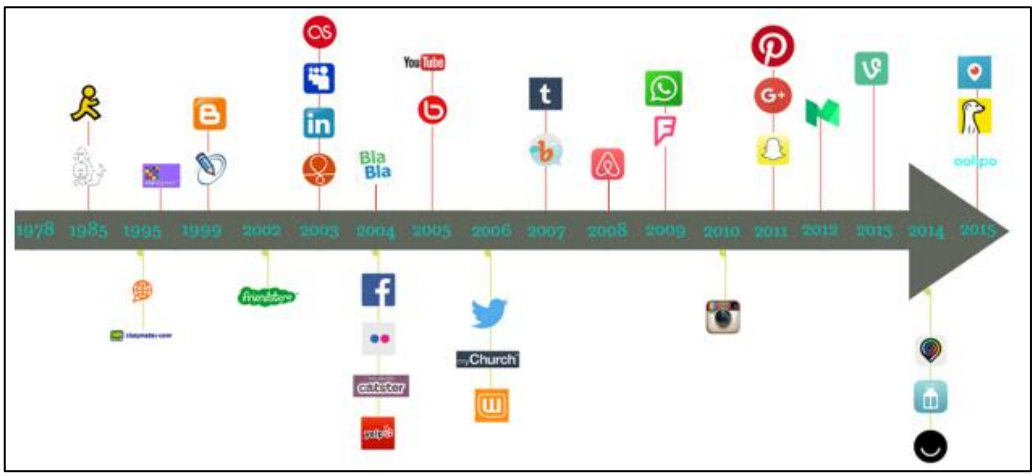
En los últimos años podemos destacar ciertos hitos reseñables y fundamentales en la evolución de las redes sociales tal y como son conceptualizadas en la actualidad

[2] MOTTA, R. 1995 "Redes.El lenguaje de los vínculos" Buenos Aires. Paidós



Img1.- Cronograma hitos de Internet

La auténtica explosión social se experimenta de forma exponencial en los últimos 25 años desde que, en 2002 se revolucionara internet con la aparición de sitios web promocionando conexiones entre usuarios de las comunidades virtuales tales como Friendster. Ya en el 2003 aparecen plataformas como MySpace, Ecademy, o LinkedIn hasta alcanzar los 200 sitios de redes sociales. En 2004 surge Facebook como modelo de apoyar a las redes universitarias, aunque pronto se extendió a todos los usuarios potenciales de Internet. Desde entonces diversas redes se han creado con muy diferentes y variadas temáticas, públicos, agrupaciones y finalidades, de funcionamiento similar a las originales y con popularidad variable.



Img2.- Cronograma redes sociales

2.1.4. Principales redes sociales

Existen más de 200 redes sociales hoy en día y, aunque muchas de ellas puedan ser temáticas y no permitan el acceso de forma abierta, la realidad es que hay ciertas redes sociales abiertas que han encontrado en estos días su mayor esplendor y cuentan con miles de millones de usuarios registrados. Entre ellas destacan especialmente, en orden de relevancia por el número de usuarios, las siguientes:

Nombre	Creador	Fecha de creación	Comprador	Fecha de compra	Usuarios	Tipo	Detalles
Facebook	Mark Zuckerberg	2004			2100 millones	Horizontal, abierta y simétrica	Pensada como comunidad de estudiantes de Harvard
Youtube	Hurley - Chen - Karim	2005	Google	2005	1800 millones	Horizontal y asimétrica	Contenido audiovisual
Whatsapp	Koum - Acton	2009	Facebook	2014	1300 millones	Horizontal, abierta y simétrica	Enviar y recibir mensajes mediante de texto y multimedia
Instagram	Systrom - Krieger	2010	Facebook	2012	1000 millones	Horizontal, y asimétrica	Compartir fotos y videos con efectos y filtros
Google+	Google	2011			343 millones	Horizontal, abierta y asimétrica	Clausurada en 2019
LinkedIn	Hoffman - Blue - Guericke - Ly - Vaillant	2003			260 millones	Vertical y simétrica	De temática profesional
Twitter	Jack Dorsey	2006			145 millones	Horizontal, abierta y asimétrica	Mensajes de hasta 280. Precursor del hashtag y de menciones

Tbl1.- Clasificación de las principales redes sociales por número de usuarios

2.2. Comunicación, marketing y análisis de redes sociales

Las redes sociales posibilitan que fluya todo tipo de información entre los usuarios y comunicar valores, necesidades o productos que una empresa, colectivo social o partido político quiera dar a conocer.

Con esta premisa surge el marketing en redes sociales, el cual puede ayudar a elevar la audiencia y convertir a personas interesadas, en clientes potenciales con una inversión muy inferior a la que se hace en medios de comunicación convencionales.

Poner en práctica una estrategia de social media repercutirá en los siguientes aspectos:

- **Construcción de imagen de marca**: atraer y satisfacer a la audiencia para generar empatía, ofreciendo contenido para construir una marca fuerte y positiva gestionando la **reputación online**.
- **Tráfico web**: el uso de redes sociales es una excelente forma de atraer visitantes a una web, lo que incrementa el posicionamiento.
- **Target**: conocer detalladamente los intereses de los consumidores y posibles consumidores para poder atraer su atención.
- **Crowdsourcing**: participación de la audiencia en la construcción de propuestas, imagen, necesidades o soluciones que beneficia tanto al producto y a la marca como al cliente o posible cliente.
- **Retorno de inversión (ROI)**: segmentar y llegar al público objetivo de manera más rápida.

Existe una regla nemotécnica para crear campañas en redes sociales y que deben regir todos y cada uno de sus objetivos, esta es **S.M.A.R.T.** o *Specific, Measurable, Assignable, Realistic* y *Time-related*, o en español Específicos, Medibles, Alcanzables, Relevantes y Basados en un periodo de tiempo delimitado.

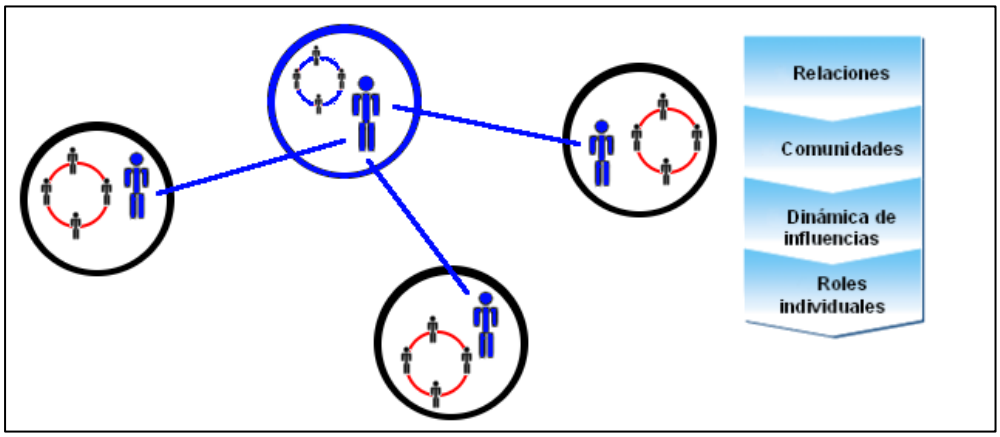
Por otro lado, el análisis de redes sociales (ARS) se originó a partir de varias perspectivas teóricas paralelas en sus inicios, pero divergentes en sus ámbitos y convergentes en sus trabajos a medida que se amplían los conocimientos en ellos.

Estos inicios estuvieron marcados por las teorías de la sociometría de *Jacob Moreno*, la teoría de grafos, las teorías del equilibrio estructural en psicología social o la antropología que, han encontrado un equilibrio en la confluencia de la teoría de la Gestalt, la teoría de campos y la dinámica de grupos del lado de la psicología social con el desarrollo de estructural y socio-antropológicos del lado de la antropología.

Se centra en la investigación de una perspectiva de la estructura social como el conjunto de vínculos que unen cualquier tipo de entidades para poner de manifiesto la realidad social a distintas escalas en momentos determinados.

Posteriormente se trató como una variedad propia de la teoría de la sociología estructural analizando las oportunidades y limitaciones de los actores como resultado de su conducta

obviando elementos culturales o subjetivos centrada en tres direcciones, la conducta de un actor en función de su posición en la red, la definición de subgrupos en la red y las motivaciones o naturaleza de los vínculos.



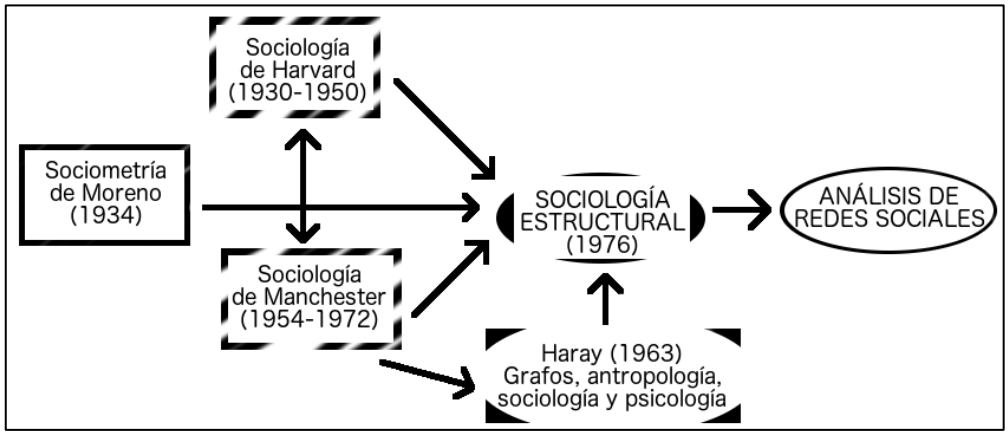
Img3.- Relaciones entre entidades y grupos de redes sociales

Existen diversos métodos de análisis que aportan una gran riqueza en el tratamiento de datos relacionales desde las perspectivas, por un lado, relacional basadas en las conexiones directas e indirectas entre unidades de una red sobre subgrafos de pares ordenados como modelo de cohesión y, por otro lado, posicional basadas en las similitudes en las pautas de las relaciones de unas unidades con otras, de manera que se identifican posiciones y los actores que ocupan cada una de ellas.

El estudio realizado por (Otte y Rousseau, 2002) permitió, desde la bibliometría y el análisis de redes sociales, comprobar las medidas de centralidad para observar los actores y sus relaciones con los demás.

Los enfoques relacional y posicional se pueden concebir como una unidad de cohesión en la vinculación entre actores con variantes **subjetiva**, donde se identifica la asociación de los miembros de un grupo y la motivación de esta asociación, y **objetiva** donde se identifican las posiciones de los individuos obviando las motivaciones. Además, también se pueden presentan desde las perspectivas **social**, que investiga la medida de conexión de los actores a otros lazos y **estructural**, que considera a los actores en base a una estructura de equivalencia, aunque no existan lazos directos entre algunos de ellos.

Hoy en día, el ARS también llega hasta la gestión del conocimiento en las empresas por medio de herramientas para el estudio de las relaciones humanas y de la manera en que interaccionan hasta obtener la unidad de investigación transversal.



Img4.- Evolución de la teoría de redes sociales

En el Análisis de Redes Sociales, los atributos observados a partir de los actores sociales son comprendidos en términos de patrones o estructuras de relaciones entre las unidades. Los vínculos relacionales entre actores son el foco primario y los atributos de los actores son secundarios[3]

(S. WASSERMAN, 1999)

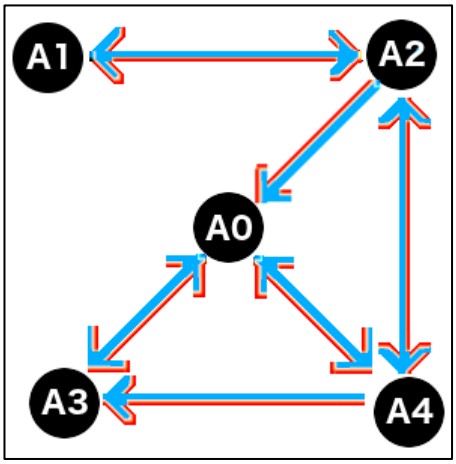
Por otro lado, el análisis de redes sociales presenta ciertas limitaciones entre las que destacan los **límites en la calidad de los datos**, donde la muestra no siempre será perfecta con procesos de extracción de topología automatizados ya que pueden darse errores de medición, mostrar datos incompletos o corrompidos, y otras inconsistencias que afectan al análisis posterior y los **límites de la muestra** ya que manejar gran cantidad de nodos puede ser contraproducente para el análisis y nos obliga a permanecer en un nivel meramente descriptivo y exploratorio.

De todo ello podemos encontrar ciertos conceptos asociados a las redes sociales y a sus formas de análisis como una conjunción de la teoría de grafos, matrices de datos unidos gracias al concepto de centralidad:

- **Red social como grafos**: Un grafo G consiste en un conjunto de nodos, N = {n1, n2,.., ng} y un conjunto de líneas, L = {l1, l2, ...lL} entre pares de nodos que se representanta como G (N, L). El uso de grafos está muy extendido porque

[3] S. WASSERMAN, K. FAUST. 1999 "Social network analysis: methods and applications". Cambridge University Press, 857p.

ofrece operaciones matemáticas con las que medir y comprobar propiedades o enunciados teóricos.



Img5.- Ejemplo de grafo dirigido de vínculos de entidades de una red social

- **Red social como matriz de datos**: se trata del conjunto de actores y sus enlaces, pudiendo incluirse nodos como atributos adicionales. Gracias a la figura Img4 podemos crear la siguiente matriz de datos:

	A0	A1	A2	A3	A4	Suma
A0				1	1	2
A1			1			1
A2	1	1			1	3
A3	1				1	2
A4	1		1			2
Suma	3	1	2	1	3	10

Tbl2.- Ejemplo de matriz de datos de vínculos de entidades de una red

- **Centralidad de cercanía en redes sociales**: medida que indica lo cerca que está un actor respecto de los demás. Por ejemplo, la centralidad aparente del nodo A0 de la imagen Img4 no es tal ya que los pasos que hay que dar para que A0 llegue a todos los nodos son 3, mientras que A2 Y A4 alcanzan a todos los nodos con un máximo de 2 saltos.

En internet, el ARS se utiliza para monitorizar e interpretar las interacciones online entre actores y el contenido publicado en los diferentes canales por medio de técnicas de investigación cuantitativa basada, principalmente, en los pilares de la teoría de grafos y en las representaciones de redes de actores, y se centra en la comprensión de las interacciones entre los actores de una red, así como de la propia estructura de la red misma. Gracias a los datos obtenidos es posible observar el nivel de interacción de la audiencia, sus preferencias, las horas y los días en los que hay más movimiento de información y poder así recoger una gran cantidad de datos sobre el público objetivo.

La manera de cuantificar y cualificar de una manera objetiva la presencia o importancia de un actor o subgrupo entre sus contactos en las redes sociales viene definido por una serie de **KPIs** (*Key Performance Indicators*) definidos para este fin y clasificados en función del objetivo u objetivos a observar, medir y evaluar tratando de buscar siempre los que mejor rendimiento ofrezcan, fijándose además en los que se vean respaldados con bajo las siglas **SMART**, descritas anteriormente, y se dividen en 3 categorías:

- **De gestión y comunidad**: esta categoría contiene dos indicadores. El primero es el **cumplimiento de periodicidad** de publicación, que hace seguimiento de periodicidad una publicación, y el segundo es el **tamaño de la comunidad** también denominado **crecimiento orgánico**, que cuantifica el número de seguidores o relaciones directas.
- **De interacción y alcance**: esta categoría tiene 5 indicadores. Tres de ellos **"*Likes*" promedio por publicación**, **"*Shares*" promedio por publicación** y **Comentarios promedio por publicación** se calculan como la división entre el número de dicha interacción en cada publicación entre el número total de publicaciones en los últimos 28 días. Las otras dos son **Usuarios alcanzados promedio por publicación**, que informa del número de usuarios que ven una publicación, y **Porcentaje de "engagement" relevancia** que calcula el porcentaje de interacción y depende de todos los anteriores de la manera que *(likes+shares+comentarios) / usuarios alcanzados x 100*.
- **De conversión y resultados**: del mismo modo que la anterior, contempla hasta 5 indicadores, y las 4 primeras son **CRT (Click Though Rate)**, que informa de la "tasa de clic" en los enlaces compartidos, **Porcentaje de conversión**, que informa del número de visitantes a un sitio que cumplen su objetivo, **CPL (Cost Per Lead)**, que informa del coste por potencial cliente logrado en un sitio web,

y **CPA (Cost Per Acquisition)**: informa del coste por cada comprador logrado y todos ellos se calculan como el cociente entre la medida individual y el total de clientes logrados en el mismo periodo y multiplicado por 100. Además, existe el **ROI (Return Of Investment)**, que estima si una campaña es conveniente, y se calcula como la diferencia entre el beneficio y el coste de la campaña en redes sociales, entre el mismo coste de la campaña en redes sociales y multiplicado por 100.

2.3. Conjuntos de datos

Existen dos tipos de datos en función del volumen que represente, por un lado, están los volúmenes pertenecientes a un mismo contexto de tal volumen que sería imposible revisar manualmente aplicando incluso estrategias de filtrado y son los denominados Big Data; por otro lado, los que cuentan con un volumen de información inferior a los pertenecientes al grupo anterior y se denominan Slow Data

El BigData describe un gran volumen de datos, tanto estructurados como no estructurados cuyo tamaño, complejidad y velocidad de crecimiento dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales.

Con el fin de utilizar eficazmente el Big Data, en la mayoría de los casos debe combinarse con datos estructurados (normalmente de una base de datos relacional) de una aplicación como un ERP (Enterprise Resource Planning) o un CRM (Customer Relationship Management).

El ritmo de crecimiento de los datos es exponencial, y pueden provenir de múltiples y diferentes canales, como datos almacenados desde hace años, redes sociales, click-stream, dispositivos móviles o sensores de todo tipo, y todos estos datos nos ayudan a descubrir cosas que les podrían haber llevado años en descubrir sin el BigData y sus herramientas.

Big Data nació con el objetivo de cubrir unas necesidades no satisfechas por las tecnologías existentes, como es el almacenamiento y tratamiento de grandes volúmenes de datos que poseen unas características muy concretas:

- Volumen: los datos, en muchas ocasiones desestructurados, que se obtienen y guardan hoy en día tienen un inmenso potencial, de manera que se es necesario definir las correctas estrategias de filtrado y ofrecer así un ahorro de tiempo.

- Velocidad: el tratamiento de datos requiere agilidad, y el tiempo de procesamiento de la información tiene que ser un factor fundamental debido al gran volumen de información manejada.
- Variedad: al existir tal volumen de información heterogénea de diferentes fuentes es preciso su tratamiento para lograr uniformidad tanto para el almacenamiento de información como para el análisis
- Veracidad: con tal volumen de información heterogénea y de fuentes diversas es preciso asegurar su veracidad ejerciendo una "depuración" en los datos.
- Valor: toda la información debe servir para aportar valor ya que de lo contrario no tendría sentido su almacenamiento, análisis o administración.

Los Slow Data son, proporcionalmente, tan comunes como los masivos debido a que cada individuo genera grandes cantidades de datos que nutren las bases de datos de las grandes compañías pero, al mismo tiempo, nutren las medianas y pequeñas empresas que recopilan datos. Cualquier minorista podría ver una decisión de compra individual hecha por un consumidor un par de veces a la semana. Otros modelos de empresa, como las de seguros, por ejemplo, podría ver una sola decisión directa en un mes, y todas ellas se enfrentan al desafío de alcanzar usuarios y al del Slow Data.

Las posibles acciones disponibles para mejorar la calidad, la cantidad y los resultados del Slow Data:

- **Inferencias**: entender sistemáticamente algo y predecir un resultado basado en un conjunto muy limitado de datos, no en el comportamiento que queremos entender, sino en comportamientos relacionados, como el uso de la aplicación o el sitio web, los resultados de marketing y el comportamiento de pago.
- **Uso de datos externos**: modelar las decisiones de los clientes ubicándolos en un contexto más amplio utilizando datos externos como educación, transporte, consumo de energía, viajes de vacaciones, salud,… que, aunque puedan parecer irrelevantes, facilitan la creación de una imagen más completa del comportamiento del cliente.
- **Diseño experimental**: consiste en variar los atributos de los productos tales como precio o características, enfocándose en una determinada población de clientes y capturando datos abundantes sobre los resultados. Gracias a esto es posible aprender rápidamente lecciones enfocadas y pragmáticas de los clientes.

Las técnicas y herramientas desarrolladas en torno al Big Data son sorprendentes, pero la mayoría de los problemas analíticos que enfrentan las organizaciones de hoy vienen determinados por el análisis del slow Data.

2.4. Twitter como fuente de datos

Twitter es la primera fuente de las últimas noticias y la fuente de redes sociales más importante para ORM. Por otro lado, representan un desafío para la minería de texto ya que los usuarios publican texto usando un lenguaje no estándar, la longitud de sus publicaciones está limitada a 280 caracteres y es posible difundir mensajes a un gran número de usuarios en muy poco tiempo.

Se ha establecido una sintaxis para la interacción, denominada *Twitter Syntax* en base a los siguientes estándares:

Acción	Menciones del usuario	Respuestas	Retweets	Hashtags
Descipción	Con el signo "**@**" antes de un nombre de usuario mencionado un usuario menciona a otro	Elegir la opción de responder en el hilo de un tweet anterior, mencionando al creador del tweet a mencionar al comienzo del tweet.	Replicación de un mensaje de otro usuario adjuntando "**RT**" y el nombre de usuario que publicó el tweet original al comienzo del retweet pudiendo encadenarse diferentes niveles de retweet para que la información relevante se distribuya por diferentes grupos.	Etiquetas temáticas que tienden a agrupar tweets en conversaciones o representan los términos principales del tweet por medio del símbolo "**#**" seguido del tema en cuestión.

Tbl3.- Elementos de sintaxis de interacción en Twitter

El análisis el fenómeno de microblogging y el modelado la propagación de información en la red social ha sido extensamente estudiado por (Kwak et al., 2010) utilizando una muestra de 467 millones de tweets de 20 millones de usuarios entre el 1 de junio y el 31 de diciembre de 2009. Yang y Leskovec utilizaron 500 millones de tweets para modelar la influencia de los nodos y la dinámica temporal de la difusión de información. (Cha et al., 2010) construyeron un conjunto de datos en el que se conectaban 54,981,152 usuarios para analizar su influencia, encontrando 1,963,263,821 enlaces por medio de 1,755,925,520 tweets.

A la hora de estudiar la información contenida en los tweets, es necesario saber que se comporta como un diccionario con más de 30 entradas, que algunas tienes valores simples como un valor numérico, booleano o Sting y otras tienen valores que son a su vez otros diccionarios o listas. A continuación se detallan estos campos:

2.5. Monitorización

La monitorización supone la recopilación sistemática de la información delimitada tras seleccionar una serie de palabras clave estableciendo un corte temporal, y haciendo un control diario y acumulativo de la información, y las compañías, organizaciones.

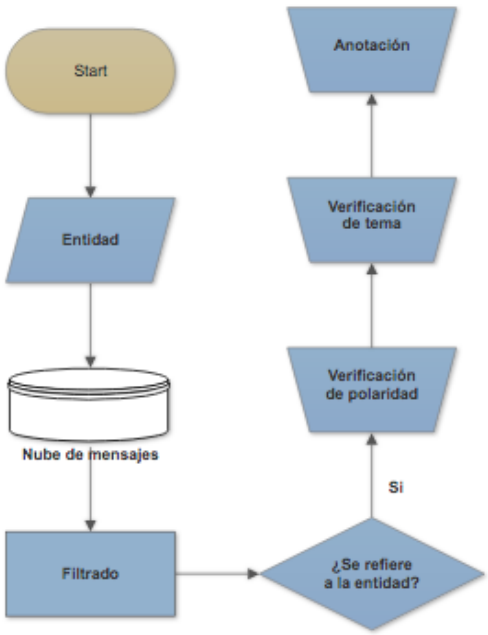
La Monitorización de la Reputación Online (Online Reputation Monitoring - ORM) consiste en la creación de una estrategia de búsqueda y rastreo, estableciendo un patrón de reconocimiento de información, basado en características independientes del texto, dónde se encuentre gracias al reconocimiento de keywords. Como las búsquedas están basadas en reglas sintáctica, es necesario definir y establecer filtros, reglas de inclusión y exclusión que consigan extraer información de la temática objetivo, de manera que sea posible la generación de dicha información se consiga de forma automática y continua, sin intervención manual y sin elevados conocimientos lingüísticos, para obtener un universo de datos relevante

El valor de la gestión de la información para la Reputación Online está delimitado por el estado del desarrollo de los sistemas de procesamiento de lenguaje natural y por la disponibilidad y acceso a tecnologías de base semántica

Por ejemplo, Google está lejos de ser un buscador semántico, lo que condiciona la forma en cómo los usuarios buscamos y, por tanto, encontramos y accedemos a la información.

Por este motivo, se ha creado la necesidad de supervisión de reputación online, de manera que es necesario realizar, a partir de un conjunto de textos que contengan una mención a la institución, tareas de:

- Filtrando de mensajes relacionados y no relacionados con la entidad
- Determinación de polaridad (positiva, neutral o negativa) de mensajes relacionados, Agrupación de mensajes relacionados por temas
- Asignación de prioridad en función de la reputación.



Img6.- Etapas de un proceso de anotación automática

Hoy en día, ya no es necesaria una revisión manual de los mismos gracias a herramientas de monitorización online con tecnología semántica que los clasifican automáticamente en positivos, negativos o neutros y que incluso pueden determinar la importancia relativa del comentario dependiendo de la fuente o el origen del mismo

Capítulo 3. Estado del arte

En este capítulo analizamos las aproximaciones existentes para la obtención de información, predicción geográfica, detección de temas y desambiguación, así como definición de polaridad y prioridad a partir de información publicada en redes sociales. Además, se mencionan trabajos relacionados para el filtrado, clasificación, agrupamiento o regresión de información y qué conclusiones se determinaron a partir de los resultados.

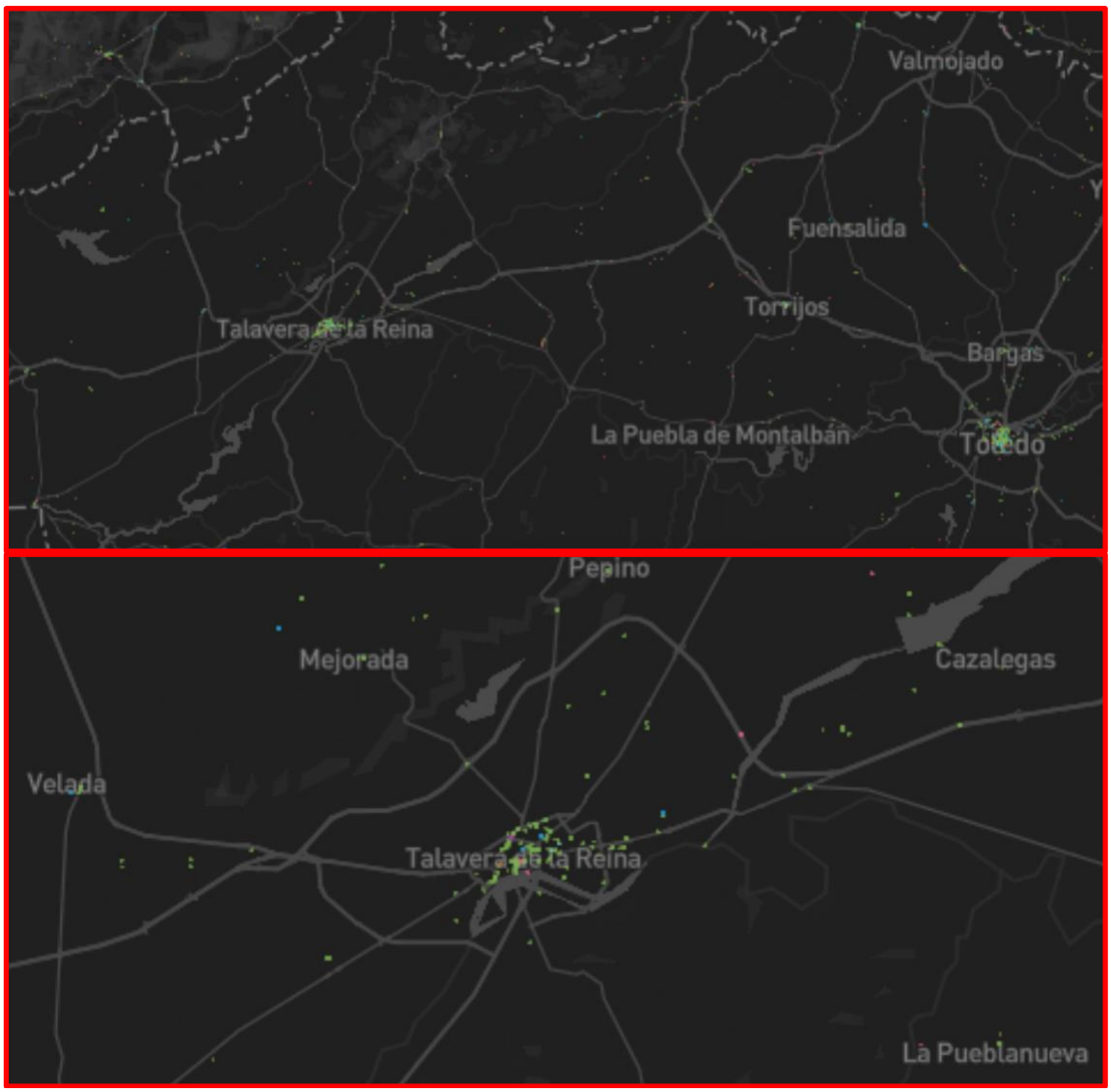
3.1. Geolocalización de tweets

El intercambio de ideas y contenido contextualizados por las ubicaciones son una preocupación los investigadores, que ha intentado averiguar esta ubicación por medio de metadatos y técnicas lingüísticas ya que, en muchos casos, las ubicaciones del perfil del usuario difieren de las ubicaciones físicas desde las cuales se está tuiteando, por lo que no se pueden usar, por si solas, como representantes útiles.

En otoño de 2009 Twitter presentó las nuevas funciones de geoetiquetado, lo que ofrece una valiosa información para la identificación de conversaciones locales.

Los datos proporcionados suelen ser coordenadas que definen la longitud y latitud por medio de los datos del dispositivo móvil, el GPS, de la red inalámbrica a la que se está conectado, o también a través de los repetidores de telefonía móvil. Sin embargo, también se pueden geolocalizar mensajes de manera manual por parte del usuario sin que ésta tenga que ser, obligatoriamente, la ubicación real del mismo. Es por esto que la fiabilidad de esta información está limitada únicamente al uso automático de la ubicación.

Los diferentes apartados de los campos de los metadatos de un tweet de ejemplo donde figuran referencias de posibles geolocalizaciones para poder determinar, con la mayor precisión posible, la localización de los mismos, en función siempre de datos proporcionados por el propio usuario, ya sea en el momento de la creación de su cuenta como coordenadas, nombre de ciudad, país, etc., o en el momento de la publicación de un mensaje.



Img7.- Tweets geoetiquetados en la comarca de Talavera en tiempo real

Uno de los primeros métodos de inferencia de geolocalización propuestos es el de (Backstrom et al., 2010), que utilizan la red social de Facebook utilizando las direcciones que describen los propios usuarios en sus perfiles en el momento de las creaciones de sus cuentas, que debían incluir números de calles y, por lo tanto, indicar ubicaciones altamente precisas. Sin embargo, la realidad es que no siempre se informa sobre la realidad de la ubicación, como es de esperar, y muchos usuarios preferían incluso no indicar sus datos.

Con estas premisas, deberemos tener en cuenta todos los orígenes de los datos y las distintas técnicas para obtener un a información relativa a la geolocalización de mensajes en redes sociales:

3.1.1. Predicción usando metadatos

Las relaciones entre los usuarios ubicados se utilizan para entrenar un modelo de árbol de regresión que predice la distancia real entre los usuarios. Éstas distancias previstas se tratan como medidas de las capacidades predictivas como particiones entre usuarios, de manera que, al separar las relaciones en diferentes particiones, la información de ubicación en la partición con las relaciones más predictivas dominará el cálculo de probabilidad. En ausencia de particiones de relación, este método se reduce a (Backstrom et al., 2010).

(McGee et al., 2011) extienden el método de (Backstrom et al., 2010) explicando que los datos proporcionados por el usuario son significativamente menos precisos en Twitter, aunque en las publicaciones de GPS, los valores de latitud y longitud medios se seleccionan como su ubicación.

Entre el 10 de noviembre y el 16 de diciembre de 2011, (Graham et al., 2013) recolectaron más de 100 millones de tweets utilizando el método de *statuses/filter* de la API1 de transmisión de Twitter, que permite recopilar tweets desde un cuadro que debía ser especificado por el usuario, aunque en esta ocasión, los tweets muestreados solo incluían tweets con una ubicación de "dispositivo" GeoIP o GPS explícita. La limitación de la clasificación solo se notó durante los momentos que coincidieron con algunas horas pico de los días de semana.

3.1.2. Predicción usando reconocimiento de nombres de lugar

Dado que la localización en Twitter no siempre está disponible y no siempre es verídica, es necesario el desarrollo de algoritmos de predicción que valoren si existe, en la medida de lo posible, una relación real entre el usuario, el mensaje y el lugar geoetiquetado.

Otro enfoque tradicional para identificar ubicaciones mediante programación es extraer nombres de lugares utilizando un sistema de Reconocimiento de Entidades Nombradas (NER) pero como el presentado por (Ritter et al., 2011), que presentaron un sistema denominado T-NER que, haciendo uso de un conjunto de herramientas pudieron realizar tareas de detección de geolocalización en datos de tweets

obteniendo una puntuación F de 0,77, el doble de lo obtenido con el sistema NER Stanford. Sin embargo, en nuestro contexto, faltan muchos nombres de lugares debido a la corta extensión de los mensajes y, sobre todo, a la ausencia de información en los mismos, así como en el lenguaje utilizado por muchos usuarios, que proporciona contextos sintácticos y semánticos limitados, lo que hace que el proceso de anotación sea mucho más difícil y en ocasiones fallido.

Por otro lado, según lo desarrollado en los trabajos de (Davis, 2011) y posteriormente de (Rout et al., 2013), se trata de evaluar los métodos de geoinferencia que predicen una ciudad o producen una clasificación de ciudades de acuerdo con cada una de las cuales son la ciudad desde la cual se origina el tweet. Las inferencias se miden utilizando una medida de precisión que informa el porcentaje de inferencias de ubicación en las que la ubicación correcta se encuentra dentro de las n ubicaciones con la clasificación más alta de la predicción; cuando n = 1, la métrica es equivalente a la definición tradicional de precisión. Si bien la métrica tiene una interpretación clara, esta precisión no tiene idea de cuán distante está la ubicación pronosticada de la ubicación real, por lo que las respuestas cercanas se consideran tan incorrectas como las respuestas distantes. Esto limita la capacidad de la métrica para medir los primeros criterios y lo hace suficientemente impreciso como para poder descartarla como forma de predicción de la geolocalización por sí sola.

Así mismo, y por lo que se extrae de los trabajos de (McGee et al., 2011) primero, (Li et al., 2012) posteriormente, y finalmente (Rout et al., 2013), es posible evaluar, de manera complementaria a la anteriormente descrita, los sistemas de geolocalización de tweets midiendo el porcentaje de predicciones que se encuentran dentro de n unidades de distancia (por ejemplo, kilómetros o millas) de la ubicación real. Describiéndose una ventaja frente al anterior aplicable a los métodos de geoinferencia de informes de ciudad y GPS. Sin embargo, la métrica es sensible a la elección en el número de unidades, ya que los valores de unidades bajos pueden hacer que los puntajes de precisión de los métodos sean idénticos, sin importar qué tan por encima de dicha unidad estuvieran sus predicciones, y de manera análoga, los valores de unidad grandes pueden crear puntajes de precisión similares para dos métodos, incluso cuando uno es significativamente más preciso. Como resultado, muchos análisis reportan múltiples valores que dificulta resumir el desempeño general de un sistema y comparar los enfoques de los análisis que reportan diferentes unidades.

(Chi et al., 2016) describe métodos de Bayes Naive multinomiales sobre diferentes conjuntos de características usando únicamente datos de texto de tweet por medio de un conjunto existente de palabras indicativas de ubicación, nombres de ubicación, hashtags, menciones de usuarios y una combinación de todas estas características. Además, usan un método basado en la frecuencia para filtrar el conjunto de características combinadas. Los resultados experimentales mostraron que la selección de características sobre los métodos combinados logró los mejores resultados en todas las métricas respecto a otros algoritmos y metodologías descritas hasta la fecha para la medición de precisión de geolocalización de tweets de usuarios con alta frecuencia de posteo, como se demuestra en los resultados de aplicar los diferentes métodos sobre los mismos conjuntos de datos consistentes en 5.000 tweets de entre el 2013-2015 y otros 5.000 de 2016:

	Anterior a 2016			Durante 2016		
	Precisión	Mediana	Media	Precisión	Mediana	Media
Miura	0,345	105,9	1594,1	0,417	79,7	2196,7
Jayasinghe	0,315	689,8	2773,6	0,524	10,2	3329,0
Chi	0,121	3105,8	4867,5	0,170	3852,7	5810,3

Tbl4.- Resultados de comparativa de predicción sobre tweets

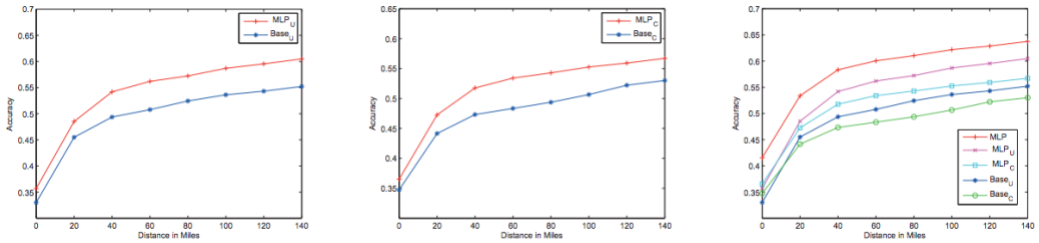
	Anteriores a 2016			Durante 2016		
	Precisión	Mediana	Media	Precisión	Mediana	Media
Miura	0,401	55,0	849,8	0,502	0,0	1218,8
Jayasinghe	0,370	267,8	2075,7	0,631	13,7	2409,0
Chi	0,196	611,7	2232,8	0,254	689,2	3487,6

Tbl5.- Resultados de comparativa de predicción sobre usuarios

Los casos de prueba de 2016 son más probables de predecir en comparación con los datos anteriores a 2016 en términos de precisión y error de distancia media. Como los datos de prueba de 2016 son de nuevos usuarios creados en 2016, esto sugiere que las personas están más dispuestas a compartir la información de su ubicación, ya sea en sus perfiles o en los datos de texto del tweet.

3.1.3. Predicción usando relación entre usuarios

Ya que un usuario puede figurar con más de una ubicación desde la que publica, y según el enfoque de (Li et al., 2012), para descubrir todas las ubicaciones de origen de un usuario se modelan las relaciones entre los usuarios y las ubicaciones construyendo un gráfico, intentando separar las relaciones de predicción de ubicación de aquellas que cumplen otras funciones en Twitter, obteniendo como resultado una estimación probabilística de la localización del usuario. La complejidad computacional del modelo depende de la complejidad del modelo de tema; sin embargo, en la práctica, se demostró que el modelo convergía rápidamente, aunque el tiempo de ejecución general del método estaba dominado por la tarea de buscar nombres de ubicación dentro de la publicación de un usuario.



Img8.- Precisión acumulativa a diferentes distancias

(Kong et al., 2014) proponen estrategias para ponderar cuál de los amigos de un usuario es probable que sea más predictivo de su ubicación, estableciendo que un usuario **a** es *amigo* de **b** si **a** ha mencionado **b** en al menos dos publicaciones. De este modo, los amigos se ponderan de acuerdo con un coeficiente de *estrechez social*, que se calcula como **la similitud de coseno de los *amigos de los dos usuarios***. Como no siempre se realizan etiquetas o menciones, la inferencia es poco precisa en usuarios con pocos vecinos etiquetados. Si se realizan varias *pasadas*, el método identifica significativamente más usuarios, aunque potencialmente a un costo elevado de precisión.

En el estudio de (Jurgens et al., 2015), este último método se probó con todos los usuarios etiquetados con ubicaciones, en cuyo caso la complejidad crece más allá de la de (Backstrom et al., 2010). Sin embargo, a medida que crece el tamaño de la red,

el cálculo del modelo de regresión en todos los pares de usuarios ubicados domina el tiempo de ejecución, lo que hace que el algoritmo no sea factible en la práctica.

3.1.4. Usuarios ruidosos

Por medio del uso de entradas vectorizadas en modelos lineales para tareas de etiquetado geográfico de nivel de usuario y, usando texto de tweets, ubicaciones autodeclaradas por el usuario, valores de zona horaria y autodescripciones del usuario como fuentes de entrada, (Miura et al., 2016) trataron de predecir la geolocalización de diferentes tweets, de manera que cada fuente de entrada se transforma de una representación de "*bolsa de palabras*" en un vector. Los vectores de la misma fuente se promediaron y los vectores de diferentes fuentes se concatenaron y se usaron como entrada a los modelos lineales.

Al mismo tiempo, en el modelo desarrollado por (Jayasinghe et al., 2016) Se adoptaron métodos de aprendizaje en conjunto con características cuidadosamente extraídas de varias fuentes disponibles en texto y metadatos. Es decir, implementaron métodos de propagación de etiquetas entre las publicaciones de tweet basados en metadatos, como "Text a diccionario geográfico", "Ubicación a diccionario geográfico", "Zona horaria", "UTC" y "de aplicación", las asignaciones de nombres de ubicación, los métodos de recuperación de texto que suponen que un texto similar proviene de la misma región, como una aproximación al método "Pigeo System" y los clasificadores basados en el idioma por medio de heurísticas. Una vez obtenidos los resultados de cada uno de los métodos individuales, combinaron las características resultantes de diferentes maneras. La mejor precisión se logra utilizando predicciones ordenadas con precisión, es decir, tomando la predicción del mejor método en la etapa de validación, y si no se encuentran resultados, retroceda al siguiente clasificador. Como los métodos se basan en servicios atendiendo a la ubicación o en información confiable, este enfoque trata de orientarse en aumentar la precisión, pero al tratarse de métodos de propagación finalmente consiguen aumentar la cobertura.

Como contamos con el problema de los usuarios ruidosos, con el ámbito de localización tan restringido como puede ser una ciudad de tamaño medio, y la previsible escasez de volumen de datos, podemos concluir que será de nuestro interés tener en cuenta las ubicaciones reales, así como la predicción de geolocalizaión por reconocimiento de nombres de lugar y por el uso de metadatos.

3.2. Detección y seguimiento de temas

Partiendo de la premisa de que la extracción automática de sintagma clave consiste en identificar un conjunto de términos relevantes que caracterizan uno o más documentos, podemos concluir que no podemos aplicarlo al enfoque de palabras clave en Twitter ya que el funcionamiento se presupone contrario. Es decir, a partir la extracción de información por medio del filtrado de una serie de palabras clave, se deberían obtener los mensajes relevantes. Sin embargo, dado que las técnicas de extracción automática de información en documentos se centran en documentos gramatical y léxicamente bien construidos, estos métodos comprometerían la integridad de los resultados del sistema ya que las palabras clave proporcionadas no pueden utilizarse como estándar de referencia precisamente porque los usuarios no siempre utilizan un lenguaje gramatical y léxicamente bien construidos.

Una de las características principales en la página de inicio de Twitter muestra una lista de los principales términos de los llamados temas de tendencias o *TrendingTopics* (TT) en todo momento. Estos términos reflejan los temas que más se están discutiendo en el momento en el flujo de flujo rápido del sitio. Twitter define los temas de tendencias como temas que son populares en cada instante. De esta manera, el planteamiento de este apartado surge de la necesidad de encontrar temáticas interesantes de estudio, no obstante, podemos prever que en datos discretos de densidad baja no podremos recurrir a metodologías o algoritmos de detección, extracción y seguimiento de temas de tendencia. Además de esto, los temas que puedan surgir con relativa importancia no son temas de índole general como pudieran ser campañas publicitarias o noticias nacionales o internacionales para que pudieran formar parte de los temas más importantes que pudieran ser detectados, sino que es la propia idiosincrasia de cada elemento de estudio la que debe determinar los temas más importantes.

Identificar los términos o palabras clave más destacados nos ayudará a abordar tanto el tema del filtrado como el de la detección de temas, y podremos validar la utilidad de usar palabras clave para monitorizar la reputación de entidades. Al aplicar diferentes temáticas sobre un mismo modelo de extracción de información, podremos verificar si aumenta o disminuye la precisión de los datos en función de la cantidad y calidad de los temas elegidos, de manera que esto refleje la idoneidad del conocimiento de los usuarios sobre los temas que en cada momento requieran de su investigación.

3.2.1. Similitud temática

Podemos entender también el problema de la detección de temas como el estudio de la similitud temática, y esto puede entenderse como una relación entre dos cadenas de texto relacionadas entre sí y separadas por una distancia léxica. La distancia de edición entre dos cadenas de texto A y B se basa en el conjunto mínimo de operaciones de edición necesarias para transformar A en B (o viceversa), como se demuestra, por ejemplo, en el estudio de (Amon et al., 2010), tras analizar funciones de similitud como Levenshtein, Smith Waterman, Jaro, 2-grams, 3-grams y soft TF-IDF entre otras, donde las operaciones de edición permitidas son eliminación, inserción y substitución de un carácter, de manera que, generalmente, **a mayor distancia menor es la similitud y viceversa**, pero algunas, como demostraron, tienden a fallar bajo la presencia de ciertas variaciones textuales como errores ortográficos, uso de abreviaturas, palabras faltantes, introducción de prefijos/sufijos sin valor semántico, reordenamiento de palabras y eliminación/adición de espacios en blanco y que provocan que se obtengan resultados erráticos.

3.2.2. Detección de eventos

En trabajos relacionados, como el de (Yamanaka et al., 2010), un método de extracción que detecta eventos en un área de observación dada, clasificando los mensajes con información GPS adjunta utilizando un modelo de máquina de vectores de soporte (SVM), se agrupan según la categoría de los mensajes y la posición y se detectan los temas principales para cada grupo. Sin embargo, este método necesita predefinir un conjunto de consultas para cada área y condición.

Un ejemplo del uso de la detección de eventos es el trabajo de (Hannon et al. 2011), como resumen automático de eventos que generan vídeos a partir de fragmentos de los tweets con mayor actividad al finalizar eventos deportivos, o el trabajo de (Zhao et al., 2011), que detectan eventos secundarios ocurridos durante los juegos de la NFL, utilizando un enfoque basado en el aumento de la actividad de tweets. (Chakrabarti y Punera, 2011) presentan un enfoque basado en los modelos de Markov para construir resúmenes de eventos en tiempo real a partir de tweets, aunque requiere conocimiento previo de eventos similares.

En el trabajo realizado por (Ishikawa et al., 2012) recopilan tweets asociados con geoetiquetas, eliminan los que no tengan relación geográfica con el lugar de estudio y realizan un análisis morfológico utilizando MeCab para descomponer el texto en partes del discurso. Utilizan un método de detección temas para determinar la frecuencia en un período determinado propuesto por (Kleinberg, 2002), que detecta si el intervalo de mensajes que llegan es más denso que en una condición normal comparándolos con otros flujos de documentos como artículos o noticias actuales. En este trabajo vamos a desarrollar esta técnica descargando los tweets con un componente geográfico basado en dos áreas diferentes y analizar las palabras contenidas en los mensajes durante un periodo de tiempo determinado con la intención de extraer temas o términos relacionados con esa ubicación.

Esta función organiza y puntúa favorablemente los documentos pertenecientes a un corpus muy grande y más aún si pertenece a un contexto de ámbito general, donde cada momento se tratan los temas más importantes a modo de Trending Topics y se tratarían de datasets de densidad alta o Big Data tal y como se extrae del trabajo de (Pérez, 2016). Sin embargo, en nuestro caso particular el resultado debería ser, presumiblemente, mucho más bajo o incluso incorrecto por tratarse de temas concretos no muy importantes a nivel global y de baja densidad o Slow Data.

Además, en redes sociales se puede también estudiar la distribución del primer dígito más significativo en el número de amigos y seguidores de los usuarios de dichas redes, lo que indicaría que se sigue la Ley de Benford. Sin embargo y, del mismo modo que sucede en el estudio de (Pérez, 2016) donde estudia solamente temas y usuarios como sucede en este trabajo, a lo que se suma que tratamos con conjuntos de datos de baja densidad, podemos llegar a la conclusión que este método no devolverá datos que se encuadren en el cumplimento de ésta ley.

La característica común de todos los trabajos anteriores es que se contextualizan en un escenario de datos denso que proporciona suficiente información y redundancia para extraer los temas principales en busca de términos relevantes, pero en nuestro caso de estudio nos interesa solo un pequeño subconjunto de la transmisión de Twitter. Además, presumiblemente, los eventos no supondrán un elemento de importancia relevante para nuestro trabajo, de manera que, podemos a descartar finalmente la detección de temas para la presente investigación por estos motivos.

3.3. Desambiguación de entidades

La tarea de Monitorización de Reputación Online (ORM) se realiza siempre en torno a de una entidad. Además de esta entidad, los temas y eventos que se discuten en involucran otras entidades relacionadas. Como las entidades nombradas son frecuentemente ambiguas, su identificación de la entidad o concepto a la que se refiere otro término no es trivial.

La ambigüedad en los textos y, más concretamente, el problema de Desambiguación del Sentido de las Palabras (Word Sense Disambiguation - WSD), consiste en asignar el sentido apropiado ante la aparición de una palabra polisémica en un contexto dado. Por otro lado, el Reconocimiento de Entidad Nombrada (Name Entity Recognition - NER), que consiste en identificar y clasificar entidades, expresiones temporales o cantidades numéricas en un texto. Ambas se relacionan con la Desambiguación de Nombre de Entidad (Named Entity Disambiguation - NED), que implica la asociación de menciones o referencias de una entidad en uno o varios textos con el objeto al que realmente hacen referencia. Por ejemplo, en la oración:

*"El próximo **AS** del baloncesto en el siguiente número del **AS**"*

Las dos ocurrencias de la palabra "AS" se refieren a dos entidades diferentes. Mientras una hace alusión al número 1 o al mejor, la otra se trata de nombre de un periódico deportivo.

NED puede plantearse desde tres perspectivas distintas: (i) como un problema de agrupación automática de menciones referidas a una misma entidad (*clustering*), (ii) enlazar términos con entradas de una base de conocimiento (*entity linking*) y (iii) como un problema de filtrado

3.3.1. Clustering

La desambiguación por clustering consiste en agrupar páginas web resultado de una búsqueda de un nombre en tantos grupos como entidades compartan ese nombre, como afirma (Artiles, 2009), complementado con la extracción de atributos personales, expuesto posteriormente por (Artiles et al., 2009) y (Artiles et al., 2010).

La técnica más utilizada y la más competitiva es la agrupación jerárquica aglomerada (Hierarchical Agglomerative Clustering - HAC), como afirman (Gooi y Allan, 2004), (Artiles et al., 2007), (Artiles et al., 2009) y (Artiles et al., 2010), donde los documentos se representan como un conjunto de palabras, aunque también se pueden utilizar porciones más pequeñas del documento, sin embargo otros trabajos usan porciones más pequeñas del documento, como oraciones donde aparece el nombre ambiguo o ventanas predefinidas de palabras, como señalan (Gooi y Allan, 2004) y (Artiles et al., 2009). Es una técnica muy común a la hora de buscar, clasificar y desambiguar entidades de manera automática, es una buena candidata para abordar nuestro problema. La medida de similitud más común en este escenario es el coseno, según (Artiles et al., 2007) y (Artiles et al., 2009), mientras que algunos trabajos, como el de (Gooi y Allan, 2004), también utilizan la divergencia Kullback-Leibler.

Lamentablemente, la agrupación jerárquica aglomerada, además de no ser apropiada para textos cortos como tweets por falta de contexto, de no ser válido para textos informales o con errores ortográficos, léxicos y gramaticales, palabras acortadas o signos de puntuación inusuales, y de no ser válidas para conjuntos de baja densidad, podemos considerar descartarla también porque, al tratarse de temas con pocos documentados o poco comunes, es de esperar que no se encuentren muchos de estos temas en la red, de manera que podrían llevar a error con mucha facilidad, al contrario de lo expuesto en la tesis de (Spina, 2014) debido, en parte por la disparidad encontrada en los datos, sus densidades, tipos o idiosincrasias. Por otro lado, dado que este trabajo se centra en la búsqueda y clasificación de tweets de densidad baja y localizada geográficamente, cabe pensar que la mejor manera de desambiguar entidades es selección de las mismas y su clasificación por geolocalización. Hipótesis que estudiaremos en nuestros experimentos.

3.3.2. Entity linking

La vinculación o enlace de entidades consiste en asociar la mención de una entidad en un texto con la entidad correspondiente en una Base de conocimientos tales como Wikipedia para facilitar la comprensión del texto, mejorar la información, crear sistemas de recuperación de información o sistemas que den respuestas a ciertas preguntas. Los usos que podemos destacar con este tipo de desambiguación son:

- **Comprobación de menciones en ORM**: el filtrado en la monitorización es un tipo de vinculación de entidad donde se debe decidir si cada mención se refiere o no a la entidad de interés
- **Detección de temas**: es necesario verificar las entidades y los conceptos mencionados y si se pueden usar como una señal para mejorar el proceso gracias a bases de conocimiento

Los sistemas realizan tres pasos para vincular la entidad a la base de conocimiento, que son **expansión**, que extrae la estructura de la base de conocimiento o resuelve una referencia en el documento, **generación de candidatos,** que busca las posibles entradas en la base de conocimientos a la que la consulta podría enlazar, y **clasificación de candidatos**, que clasifica las candidatas calculando la similitud entre la consulta representada y las entidades, y definir el momento en que la entidad no existe en el base de conocimientos.

Debemos tener en cuenta esta solución debido a la globalización de la información y a que la selección de palabras elegidas para el análisis de reputación online podrá pertenecer a diferentes contextos, de manera que podríamos esperar que una entidad coincida con otras entidades de las que aparentemente no tiene relación más allá de la relación de semejanza entre sus nombres.

La ambigüedad puede resolver usando estrategias basadas en la consistencia de las entidades vinculadas o por medio de técnicas de aprendizaje automático aprovechando en ambos casos, y siempre que sea posible, las resoluciones no ambiguas, como búsqueda de la solución para los ambiguos. De ésta manera, y según los estudios realizados por (Ferragina et al., 2010), donde aborda el problema de la anotación de fragmentos de texto reducido usando Wikipedia como base de conocimiento y donde tras plantear el problema realiza y compara varios experimentos, concluye que casi el 95% de un total de 5.000 tweets analizados tienen al menos 3 frases con una entrada en Wikipedia, aunque esto no implica que sea una entidad, lo que demuestra que Wikipedia tiene una alta cobertura como catálogo para la desambiguación de tweets. Considerar el concepto de Wikipedia como el más probable para cada n-gramo como asunción de la mayor probabilidad puede ser una buena solución para la desambiguación, y parece ser un recurso efectivo para la desambiguación de entidades y la vinculación en Twitter porque tiene una alta

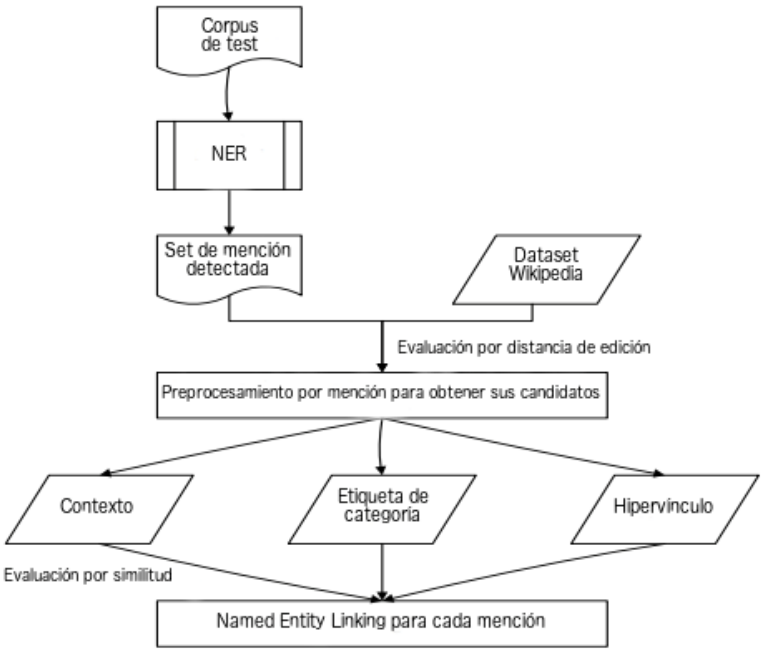
cobertura, se puede usar con precisión explotando su estructura de hipervínculo y pueden ser indexados para su uso en escenarios en tiempo real.

Wikipedia y otras fuentes de conocimiento contienen un contenido semántico estructurado o semiestructurado, que puede usarse como inventario de sentido para la desambiguación. Los trabajos más relativos en NED basados en Wikipedia son realizados por (Bunescu, 2006), donde entrenaron un núcleo SVM de desambiguación para explotar la alta cobertura y la rica estructura del conocimiento, o (Cucerzan, 2007), que extendió el trabajo anterior al agregar algunas características más ricas a la comparación de similitud. Además, según (Nguyen et al., 2008) y (Nguyen et al., 2010), resulta que la desambiguación funciona mejor usando las características de Wikipedia como títulos de entidad o de página, etiquetas de categoría o de enlace de salida en combinación con características de texto. (Alhelbawy et al., 2012) desarrollan una función de similitud de documentos basada en las menciones de entidades nombradas que se encuentran en dos documentos en lugar del modelo de espacio vectorial común que calcula la similitud del coseno. Por último, y tal y como se extrae de la tesis de (Spina, 2014), el uso de la probabilidad común en Wikipedia como método de desambiguación por Entity Linking parecen ser un recurso efectivo para la desambiguación de entidades y la vinculación en Twitter porque tiene una alta cobertura, se puede usar con precisión explotando su estructura de hipervínculo y pueden ser indexados para su uso en escenarios en tiempo real.

Según (Naderi, 2015), es posible crear un ranking de candidatos por medio del uso de técnicas supervisadas, no supervisadas o basadas en conocimiento y desambiguar con la generación de vector binario o con lo que denomina generación gráfica usando tanto información local como global, pero en ambos casos el resultado es que la desambiguación de términos no se realiza correctamente en el 100% de los casos.

Por último, como explica (Labori, 2016), las técnicas para la desambiguación que podemos utilizar son el cálculo de la similitud entre los cálculos del TF-IDF sobre los términos o con Naive Bayes, que pueden ser usadas en el problema de la desambiguación de entidades llegando a obtener sus mejores resultados con casi un 60% de precisión y un 74% de exactitud en un pequeño corpus de 196 oraciones en inglés anotadas manualmente y 312 entidades diferentes.

Img10 muestra un framework NED basado en Wikipedia donde se detectan las menciones en un copus de test, cada una de las cuales serán preprocesadas construyendo índices para la entidad candidata apropiada.



Img9.- Framework NED

Con la utilidad Wikifier, que se trata de una librería para Python que trata de anotar un documento con conceptos relevantes de la Wikipedia. Se utiliza un método basado en pagerank para identificar un conjunto coherente de conceptos relevantes considerando el documento de entrada como un todo. La ventaja de este enfoque es que la Wikipedia es una fuente de información de libre acceso, cubre una amplia gama de temas, tiene una rica estructura interna y cada concepto está asociado con un documento textual semiestructurado que se puede utilizar para ayudar en el proceso de anotación semántica.

Como desventaja podemos mencionar que habría que implementar un algoritmo que, haciendo uso de su librería, tomará como entrada de documento cada uno de los tweets recuperados porque de otra manera habría que tomar un conjunto de ellos como un único documento. La primera opción podría desambiguar de manera inconclusa porque no en todos los mensajes relevantes se encuentran términos existentes en la Wikipedia por la idiosincrasia de las búsquedas requeridas, y la segunda es poco escalable porque, primero tiene límite de 1000 caracteres como

parámetro de entrada de documentos y, segundo, no siempre detecta ciertas claves como entidad a desambiguar.

3.3.3. Desambiguación de entidades en Twitter

Como ya se ha explicado, la desambiguación de nombres de entidades para la monitorización de la reputación online es muy útil en textos relativamente largos como noticias o publicaciones en blogs, pero en el caso del microblogging, donde el contenido de una mención es mínimo, esta tarea es mucho más compleja.

Las investigaciones de NED en Twitter se han centrado en el uso de Wikipedia a modo de Entity Linking como principal base de conocimientos, de manera que los tweets están vinculados a páginas de Wikipedia como entidades o conceptos creando, en primer lugar, un índice invertido de Wikipedia de manera que es el texto el que apunta a un lugar de Wikipedia y que, cuando se detecta un texto vinculado, se recuperan las posibles acepciones del mismo para poder asignarle el que más sentido tenga.

Como se puede extraer del desarrollo experimental del artículo POPSTAR (Saleiro et al., RepLab 2013), y con un pequeño experimento con nuestro propio dataset, las técnicas de Entity Linking no supervisadas para la desambiguación no van dar, previsiblemente, unos resultados notablemente buenos respecto una la clasificación automática que no esté basada en algoritmos de aproximación de desambiguación principalmente por tratarse de textos categorizados como slow data por su baja densidad.

En el escenario de ejecución propuesto en este trabajo podremos comprobar en el siguiente apartado la baja precisión en esta forma de desambiguación debido tanto a que los términos usados para la extracción de información como la información contenida en los tweets se debe tratar como un problema de slow data, de manera que no siempre habrá una base de conocimientos extensa y consistente previa para desambiguar tweets. Además, y sobre todo, los términos tratados no siempre van a estar contenidos o descritos correctamente en bases de conocimiento como Wikipedia.

3.4. Prioridad reputacional y polaridad

El análisis de reputación tradicional es principalmente manual, aunque los medios en línea permiten procesar, comprender y agregar grandes flujos de datos y opiniones. Aunque, la minería de opinión ha logrado avances significativos en los últimos años, la minería e interpretación de opiniones es, en general, un problema mucho más difícil y menos comprendido, ya que las opiniones sobre personas y organizaciones no se pueden estructurar en torno a un conjunto fijo de características o aspectos, que requieren una mayor complejidad

La tarea de controlar la reputación de las entidades por medio del monitoreo para los analistas consiste en buscar en el flujo de tweets las posibles menciones a la entidad, filtrar aquellos que sí se refieren a la entidad y clasificarlos en función del grado en que son potenciales problemas que pueden tener un impacto sustancial en la reputación de la entidad y deben ser manejados por expertos en gestión de reputación. Además, no todos los temas discutidos tienen las mismas consecuencias en la reputación de la entidad, ya que las alertas deben ser tratados primero. La tarea Prioridad de tema se define como un problema de clasificación de acuerdo con su influencia potencial en la reputación de la entidad.

Según (Kreher y Stinson, 1998), si tenemos un conjunto de una salida que represente los temas detectados llamado C = $\{C_1 \ldots C_k\}$, la prioridad se representa como:

$$prioridad : C \rightarrow \{1, \ldots, N\}$$

Donde N <=k, por ejemplo, dos conjuntos pueden tener la misma prioridad, aunque los sistemas deben optimizar una función de prioridad que se aproxime al estándar aureo con tres niveles (alerta, medio, bajo).

Hay que tener en cuenta que los valores de prioridad se anotan en una escala graduada, donde las métricas de evaluación de recuperación de información (IR) estándar utilizadas para evaluar los sistemas de clasificación con relevancia gradual, no son aplicables en nuestro escenario, como explica (Järvelin y Kekäläinen, 2002).

La detección de la polaridad para la reputación es un paso esencial para asignar prioridad y se evalúa como una subtarea independiente como se explica en el trabajo de (Amigo et al., 2013) en RepLab 2013, donde se define la clasificación de prioridad de los temas,

como entrada para los expertos en monitoreo de reputación. Dichas subtareas son el filtrado, clasificación de polaridad por reputación y asignación de prioridad.

Según se describe en dicho trabajo, donde evalúan la polaridad de acuerdo con la precisión y R&S por medio de diferentes sistemas, donde sólo los tweets relacionados con la entidad en el conjunto de prueba han sido evaluados y donde los tweets relacionados sin respuesta del sistema están penalizados, la clase mayoritaria es el conjunto de datos "POSITIVO", donde el enfoque por machine learning mejoran al resto de sistemas de acuerdo con la precisión y el R&S. También trata de predecir la polaridad promedio de una entidad con respecto a otras entidades por medio de la correlación de Pearson entre el promedio estimado y los niveles reales de polaridad entre las entidades, donde algunos enfoques pueden estimar la reputación de polaridad promedio de una entidad con una correlación de 0.9 con la verdad fundamental.

	Acc	Ratio tw procesado	Nivel Corr Ent	R	S	F
Machine Learning + enlaces	0,69	1	0,88	0,48	0,34	0,38
Machine Learning	0,65	1	0,82	0,5	0,15	0,19
Clustering	0,62	1	0,82	0,38	0,27	0,29
Análisis Sintáctico + Sentimientos	0,44	1	0,52	0,31	0,4	0,34

Tbl6.- Precisión, proporción de tw procesados, correlación, confiabilidad sensibilidad y medida F

Esta clase de sistemas es previsiblemente adecuada para trabajos que involucren volúmenes de datos muy grandes, pero para la tarea que nos ocupa, como es el caso de datos pertenecientes al grupo slow data, estos resultados serán muy dispares. Además, el entrenamiento sobre un conjunto de documentos para poder clasificar de manera automática la polaridad de datos de densidad baja no garantiza en ningún caso que los documentos clasificados tengan suficiente entrenamiento o que este sea correcto para poder discernir automáticamente para el resto de textos.

Como describe también (Amigo et al., 2013) en RepLab 2013, donde se evalúa la clasificación de la prioridad en base a diferentes sistemas, la tarea prioritaria consiste en clasificar los tweets en tres niveles. La confiabilidad representa la proporción de relaciones de prioridad correctas por tweet, mientras que la sensibilidad representa la

proporción de relaciones capturadas por tweet. Solo los tweets relacionados se consideran en el proceso de evaluación del mismo modo que en la polaridad donde la mejor solución logra una alta puntuación tanto en las mediciones de I + S como de precisión. El enfoque de línea de base se ha mejorado sustancialmente para ambas medidas.

	R	S	F	Acc	Tw procesados
Machine Learning	0,39	0,32	0,34	0,63	0,97
Clustering	0,24	0,2	0,2	0,46	1
Todos: importancia media	0	0	0	0,52	1
Todos: sin importancia	0	0	0	0,44	1
Todos: alerta	0	0	0	0,04	1

Tbl7.- Confiabilidad, sensibilidad, medida F, precisión y nº de Tw procesados

Del mismo modo que con la polaridad, esta clase de sistemas es previsiblemente adecuada para trabajos que involucren volúmenes de datos muy grandes, pero para la tarea que nos ocupa, como es el caso de datos pertenecientes al grupo slow data, estos resultados serán muy dispares. Además, el entrenamiento sobre un conjunto de documentos para poder clasificar de manera automática la polaridad de datos de densidad baja no garantiza en ningún caso que los documentos clasificados tengan suficiente entrenamiento o que este sea correcto para poder discernir automáticamente para el resto de textos.

La polaridad para la reputación es sustancialmente diferente del análisis de sentimiento estándar, ya que, cuando se analiza la polaridad para determinar la reputación, se deben considerar tanto los hechos como las opiniones, de manera que a los sistemas no se les pedirá explícitamente que clasifiquen los tweets, sino que deben encontrar la polaridad de la reputación independientemente de que tengan contenido favorable o desfavorable. Además, los sentimientos negativos no siempre implican polaridad negativa para la reputación y viceversa.

Es posible evaluar la prioridad de cada tema relacionado con la entidad diferenciando en varios niveles de importancia por medio del análisis textual teniendo en cuenta factores característicos de evaluaciones de prioridad como son la polaridad, la centralidad y, por último, la reputación del autor. Este último factor puede ser ciertamente interesante en volúmenes de datos de gran envergadura con cientos de miles de documentos, pero para datos de densidad baja resulta poco probable que existan muchos autores con prioridad alta, de manera que no siempre será determinante en este trabajo.

La forma de evaluar los **sistemas,** es por medio de Exactitud (*Accuracy*) como una medida de alta interpretación, y la combinación de Confiabilidad y Sensibilidad (*R&S Reliability and Sensitivity*) como una medida estricta y fundamentada en la teoría, que asume que cualquier tarea de la organización consiste en una bolsa de relaciones entre documentos de manera que dos documentos están relacionados si tienen diferentes niveles de prioridad, polaridad o relación, o cuando aparecen en el mismo grupo.

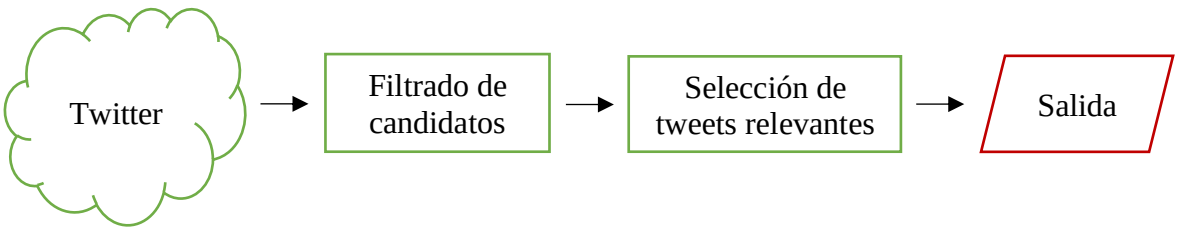
Existen propuestas para la clasificación de polaridad que utiliza un enfoque basado en el léxico para el **análisis de sentimientos**, mejorado con un **análisis sintáctico** completo y **detección de modificadores de negación y polaridad**, que también proporciona la polaridad a nivel de entidad. Otros enfoques utilizan **Machine Learning** agregando información de metadatos e incluso información externa mediante el uso de los enlaces proporcionados a Wikipedia y los sitios web oficiales de las entidades. También se han desarrollado algoritmos de **clustering** sobre conjuntos de tweets independientes, así como conjuntos de tweets agrupados.

Capítulo 4. Modelo propuesto

En este capítulo se describe el esquema general del modelo propuesto. Se describen también las diferentes variantes, que evaluaremos para discernir si resultan factibles o si existe alguna previsión de resultados favorables, en base al análisis de las diferentes técnicas realizado en el Capítulo 3. Por último, se detallan las diferentes herramientas que han sido necesarias para llevarlo a cabo los diferentes experimentos.

4.1. Esquema general

La base de conocimiento del presente trabajo son datos provenientes de la red social Twitter que, posteriormente serán procesados y analizados y evaluados, de manera que es necesaria su extracción de manera parametrizada. Para conseguirlo se procede en dos pasos (ver Img10), (i) un primer paso de extracción de mensajes en base a la geolocalización real proporcionada por el usuario, es decir, por medio de las coordenadas que definen la longitud y latitud desde la que se ha enviado el mensaje. A este conjunto se añaden aquellos mensajes filtrados en base al uso de palabras clave, términos relevantes o cuentas de usuario destacadas, con los que acabamos teniendo conjunto de mensajes candidatos. (ii) El segundo paso consiste en la selección de tweets relevantes, de entre los anteriores, por medio del ranking automático basado en la coincidencia o reiteración del uso de las palabras clave, términos relevantes o cuentas de usuario definidas en la fase anterior, la predicción de la geolocalización de aquellos mensajes que no hayan sido geolocalizados en el paso anterior y la interacción medida por el conteo de *likes*, *retweet*, comentarios o mención por parte de los usuarios en la red social con cada uno de los mensajes.



Img10.- Esquema del sistema del modelo propuesto

4.2. Geolocalización

Como se ha explicado anteriormente, es preciso el uso y aplicación de técnicas de geolocalización, tanto de manera directa en la primera fase del modelo propuesto, como por medio de la predicción en la segunda fase.

En cuanto a la definición de la geolocalización por parte del usuario en el momento de publicar el mensaje, los algoritmos de inferencia de geolocalización que usan Twitter desde 2009 requieren una fuente de datos de ubicación. A pesar de que las publicaciones con anotaciones de longitud y latitud tenían una frecuencia muy baja, de alrededor del 1%, este valor se está viendo incrementado en los últimos años, pero sigue dependiendo de los propios usuarios que, en ocasiones, indican unos valores que no corresponden con la realidad. Además, en ocasiones podemos tomar esta información de otros campos dentro de los metadatos de un tweet es la mejor solución.

En lo que a la predicción de la geolocalización respecta, las evaluaciones de los métodos de inferencia geográfica han tenido poca estandarización en los métodos utilizadas para medir el rendimiento, lo que reduce significativamente la comparabilidad entre los resultados. Además, los métodos existentes a menudo prueban diferentes capacidades, por lo que cuando solo se informa una métrica, se proporciona una vista incompleta del rendimiento de un método.

Se trata, por tanto, de identificar la ubicación de un tweet revisando **la información proporcionada en los metadatos** del mismo como criterio para lograr una correcta predicción. Ésta forma parece factible para una gran cantidad de mensajes que tengan dicha información. No obstante, las predicciones no siempre serán acertada porque, en ocasiones, el usuario puede añadir información errónea. Por este motivo, se suelen combinar varias de éstas para determinar la veracidad de los datos, tales como la zona horaria, la latitud y longitud, la ciudad, datos del perfil, cuentas de usuario a las que sigue, etc. Además, es posible revisar **la cantidad de publicaciones identificadas** como un criterio de verificación adicional, necesario para distinguir los enfoques en función de la cantidad de datos que se pueden ubicar.

Como resultado, las métricas basadas en datos de post y usuario permiten aumentar precisión con la que se puede localizar un usuario determinado y, lo que más nos interesa, sus publicaciones.

4.3. Filtrado de candidatos

El objetivo perseguido en esta fase es el de abordar la necesidad de filtrar mensajes de Twitter para lo que se aplicará un filtrado basado en geolocalización para que la recuperación de los mensajes se base en aquellos cuyo usuario haya decido utilizar geolocalizar el mensaje y éste se sitúe dentro de un radio determinado. Esto nos devolverá una serie de mensajes de los que conocemos con certeza su cercanía a un punto, y de los que podremos analizar términos de mayor relevancia para poder determinar los criterios de filtrado y descartar los que no sean relevantes para nuestro estudio.

A este filtro se añade otro que contengan palabras clave, términos relevantes o hagan menciones a cuentas de usuario destacadas. Para ello planteamos la posibilidad de seleccionar dichas palabras clave, términos y cuentas de usuario por medio de diferentes técnicas como son el uso de algoritmos de pesado, desambiguación de entidades, pero como la densidad de los datos con los que contamos no es muy grande, necesitamos plantear una técnica heurística como alternativa.

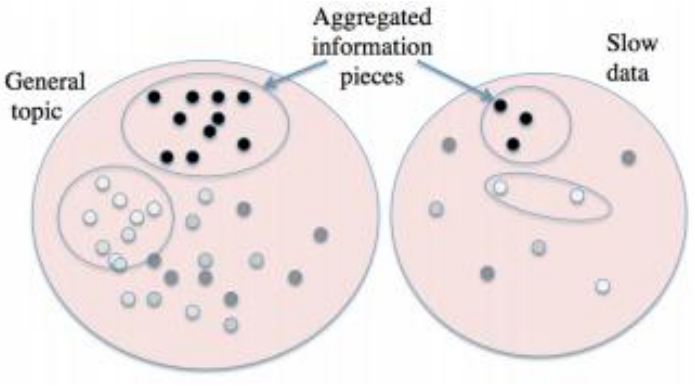
4.3.1. Funciones de pesado de términos

La detección y el seguimiento de temas ha alcanzado mucha popularidad en los últimos años debido a la tendencia de Big Data y al interés de las empresas e instituciones por controlar los contenidos generados en las redes sociales. Se ha propuesto una variedad de técnicas, basadas principalmente en la concurrencia y redundancia de palabras.

Sin embargo, estos enfoques tienden a fallar cuando el volumen de datos sobre un evento o entidad es escaso en el tiempo. Llamamos a este nuevo contexto como **Escenario de *Slow-data***, donde el desafío es encontrar cadenas de información escasa en flujos de texto de larga duración por medio de la gestión de reputación en línea y su vinculación con las métricas tradicionales de evaluación y detección de temas con los criterios de satisfacción del usuario

Dado que proponemos investigar la mejor opción posible para desarrollar un sistema de para el análisis de ORM en ámbitos locales, es necesario tener en cuenta, primero, que el volumen de los datos obtenidos no será especialmente significativo como

podría pasar con un país o con una gran compañía, de manera que es imprescindible estudiarlo desde el prisma del slow data.



Img11.- Esquema de coincidencia de subconjuntos

En la búsqueda de términos relevantes podría resultar útil identificar términos característicos de un documento en relación al resto de la colección, por lo que sería razonable el uso de técnicas de filtrado basadas en el contenido semántico como TF-IDF bajo la intuición de que un término que aparece en sólo unos pocos documentos podría ser descriptivo.

Hay dos factores que pueden influir en la densidad de las tendencias:

- Los temas son parte de la cola de temas destacados de Twitter, incluso, presumiblemente, no estarán en dicha cola debido a que la detección de temas está orientada a entidades donde los temas se refieren a una entidad determinada en un corto período de tiempo. Además, muchos de los temas tienen un volumen muy pequeño en comparación con el tamaño del flujo de Twitter específico de la entidad. Esto causa la dispersión de los datos, lo que hace que la detección de temas sea un problema difícil.
- Segundo, la cobertura es crucial para minimizar el efecto de no detectar una amenaza potencial, de manera que todos los tweets deben ser analizados por sistemas expertos o manualmente para asígnale un tema a cada uno.

A pesar de que podrían obtenerse buenos datos al aplicar algoritmos de pesado sobre el conjunto de mensajes para poder obtener un listado de palabras clave, términos relevantes y cuentas de usuario destacadas que deben ser definidas para realizar el filtrado de candidatos, al centrarse el interés del presente trabajo en los datos de ámbito local, podemos pensar que los términos que se obtendrán no serán relevantes,

por lo que se plantean, a continuación, otras técnicas. No obstante, y para verificar si realmente esto es cierto, funciones de pesado de términos serán utilizadas para determinar si es posible su uso en la extracción de términos relevantes.

4.3.2. Desambiguación de entidades

Pretendemos filtrar y clasificar una serie de mensajes relacionados con un tema con una baja densidad de ámbito local. Es muy frecuente, sin embargo, que se pueda mencionar en más de un contexto y en muchos lugares del mundo en el mismo momento, de manera que no siempre es sencillo averiguar si se está hablando de algo interesante para nosotros. Por poner un ejemplo real del ejercicio y relacionado con las pruebas que se están llevando a cabo para este trabajo, existen distintas acepciones y usos de la palabra "**Talavera**":

Lugares	**España**	Talavera de la Reina, ciudad de la provincia de Toledo. Talavera la nueva, EATIM de Talavera de la Reina Talavera, municipio de la provincia de Lérida. Talavera la Real, municipio de la provincia de Badajoz. Talavera la Vieja, municipio de la provincia de Cáceres desaparecida en 1963. Tierras de Talavera, comarca de España.
	Argentina	Nuestra Señora de Talavera, antigua ciudad de la provincia de Salta Isla Talavera, provincia de Buenos Aires
	Bolivia	Villa talavera (puna). Departamento de Potosí
	Estados Unidos	Talavera (Oklahoma), Estados Unidos de América
	Filipinas	Talavera, municipio en la provincia de Nueva Écija. Talavera, barrio de Taganaán, municipio de la provincia de Surigao del Norte. Talavera, provincia de Surigao del Norte, región de Caraga
	Paraguay	Isla Talavera, departamento de Misiones, Paraguay
	Perú	Talavera de la Reyna, distrito de la provincia de Andahuaylas, Perú
Personalidades destacadas	Fray Hernando de Talavera (1428-1507), arzobispo y confesor de la reina Isabel la Católica Talavera Vernon Anson (1809-1895), Almirante de la Marina Jonathan Talavera Pérez, Empresario Mexicano. Arcipreste de Talavera, escritor español. Hernando de talavera, prelado español. Juan Talavera y de la Vega, arquitecto español del siglo XIX. Juan Talavera y Heredia, arquitecto español del siglo XX, hijo de Juan Talavera y de la Vega. Alfredo Talavera Díaz, futbolista mexicano. María Talavera Broussé, compañera sentimental del político mexicano Ricardo Flores Magón. Manuel Antonio Talavera, cronista paraguayo. María Eloisa Talavera Hernández, política mexicana. Juan de Talavera, arquitecto y escultor del siglo XV. María Talavera Solís. 2008 Salvador Talavera 1945- Francisco Ibáñez Talavera, dibujante y escritor español. Francisco Javier Errázuriz Talavera, empresario y político chileno.	

	Hugo Ricardo Talavera, futbolista paraguayo. Juan Andrés Fontaine Talavera, ministro del gabinete chileno. Tracee Talavera, gimnasta mexicano-estadounidense.
Batallas	Batalla de Talavera (1809), combate entre ejércitos angloespañol y francés Batalla de Talavera (1936), durante la Guerra Civil Española. Regimiento Los Talaveras de la Reina, regimiento español en la Reconquista española de Chile
Deportes	Talavera CF, asociación de clubes de fútbol de Talavera de la Reina, activa desde 1948. Talavera FS, un club de fútbol sala con sede en Talavera de la Reina, fundado en 1990. UD Talavera, asociación de clubes de fútbol de Talavera de la Reina, fundada en 1993.
Cosas	Cerámica típica de Talavera de la Reina Talavera poblana, cerámica de mayólica típica del estado de Puebla, México. HMS Talavera, buque de la armada británica. Murallas de Talavera, recinto amurallado de la Ciudad de Talavera de la Reina, España.
Otros usos	HMS Talavera (1818), un barco de la Royal Navy en servicio 1818–1840 Talavera (gimnasia), una barra de equilibrio llamada así por Tracee Talavera La cerámica de talavera, un tipo de cerámica mexicana. Regimiento de Talavera o talaveras, una unidad española involucrada en la batalla de Rancagua. Talavera (araña), un género de arañas saltarinas.

Tbl8.- Distintas acepciones de la palabra Talavera o asociadas a ella

Además, existen un total de 1031 empresas con la palabra "Talavera" contenida en su nombre sólo en España. Hemos encontrado además 15 empresas que contienen la palabra "Talavera" en México, 15 más en Argentina, 3 en Bolivia, 5 en Oklahoma (Estados Unidos), 15 en Filipinas, 5 en Paraguay y 15 en Perú, así como centenares de calles con nombre de Talavera repartidas por el mundo. Del mismo modo, habrá que tener en cuenta que, por un lado **Talavera** puede ser un apellido y por lo tanto puede haber muchas personas mencionadas como tal y, por otro lado, "Talavera" puede encontrarse con otras acepciones debido, para el ejemplo de Talavera de la Reina, al uso de sus nombres en la antigüedad, como "Aebura", "Caesarobriga", "Ebora", "Talabayra" "Talabaira", a las que podríamos encontrar y, por lo tanto añadir a todo lo anterior, múltiples acepciones, apellidos o nombres de empresas.

Una vez descartada la desambiguación por clustering, proponemos intentar desambiguar por medio de técnicas de desambiguación por entity linking de manera aislada para comprobar si sería útil desarrollar alguna de ellas en el trabajo de ORM localizado teniendo en cuenta las características planteadas por geolocalización para entidades poco representativas con baja densidad y teniendo en cuenta también que el contexto es extremadamente reducido dado la longitud máxima de un texto proveniente de microbloggin.

4.3.3. Heurísticas

Por heurística entendemos aquellas estrategias, métodos o criterios no instaurados de manera estandarizada y que son utilizados para hacer más sencilla la solución de problemas. Dichas técnicas son producto del conocimiento experto sobre un tema a solucionar con un buen rendimiento, procurando cierto grado de confianza al encontrar soluciones de alta calidad con un costo razonable.

Teniendo en cuenta de los estudios de las técnicas de filtrado anteriormente analizados y a las sugerencias de diferentes expertos en la materia, se plantea un sistema de clasificación alternativa en la que se obtienen palabras clave, términos relevantes y cuentas de usuario destacadas por medio de heurística.

La selección de dichos términos, entidades y cuentas de usuario se trata de términos definidos por los mandatarios de diferentes partidos políticos de la ciudad, así como los titulares de diferentes medios de comunicación de alcance local, siendo el número de términos, cuentas o entidades elegidas como referentes el mismo por cada uno de los expertos, medios de comunicación, políticos o expertos consultados.

4.4. Selección de tweets relevantes

En esta fase del sistema resulta necesario, en gran parte de los mensajes recuperados, el uso de sistemas de predicción de geolocalización. Para ello se utiliza un sistema similar al utilizado por (Backstrom et al., 2010) revisando los datos descritos en el perfil de cada usuario como último criterio de predicción. Finalmente nos decantamos para la mayoría de los casos en los que se requiere predicción de geolocalización por usar un sistema probabilístico basado en el estudio de (Davis, 2011) que requiere poco tiempo de búsqueda y procesamiento, de manera que, a raíz toda de la información de geolocalización encontrada en los metadatos del tweet, se detectan distancias respecto del origen para cada uno de ellos y se asigna un valor medio de entre todas ellas.

Sobre todos los mensajes resultado del primer filtrado será posible clasificarlos para la selección de los tweets relevantes distinguiendo entre 3 condicionantes que describan hasta 3 niveles de clasificación automática para medir y comprobar de la siguiente manera:

1.- Mensajes que citen dos o más de las entidades, usuarios, hashtag o palabras clave.
2.- Mensajes que estén localizados en un radio no superior a 150km del centro de la población objeto de estudio
3.- En caso de cumplirse las dos anteriores, mensajes que tengan una interacción relevante entre menciones, respuestas y likes.

Se pretende, una vez recuperados, filtrados y clasificados los datos, crear un conjunto de documentos anotado y polarizado manualmente como documento de control para poder estudiar la calidad de las clasificaciones de las distintas aproximaciones comprobando, además, su calidad respecto a otras técnicas de recuperación y clasificación de información. Además, al tratarse de un trabajo enfocado al mundo de la política local, se prevé posible una aproximación a una relación entre los datos obtenidos en el conjunto o las aproximaciones con los resultados electorales de la campaña local sucedida durante el presente estudio.

4.5. Herramientas empleadas

En esta sección se exponen las herramientas empleadas para la extracción de información de la red social twitter, para extraer la geolocalización de los mismos y, para analizar la relevancia de términos para discernir su idoneidad en el sistema respectivamente. Dichas herramientas están basadas en librerías para Python, con lo que se puede conseguir una única aplicación integrada con la que sería posible automatizar el proceso de extracción, filtrado y clasificación.

4.5.1. Tweepy

Para poder acceder y recuperar la información requerida de Twitter es necesario el uso de la librería para Python llamada Tweepy.

Tweepy es la librería más conocida para acceder a la API de Twitter desde Python. Se encarga de comunicarse con la API Streaming de Twitter mediante la clase StreamListener o a la REST API. La primera captura tweets basados en uno o varios términos de búsqueda bajo condicionantes “OR” entre ellos proporcionando información en tiempo real. La segunda funciona esencialmente mediante consultas a sus servidores para una petición concreta pudiendo indicar también términos o grupos de términos bajo condicionantes “OR” entre ellos para su obtención, pero

también valores de fechas para determinar una acotación temporal, así como una referencia geográfica y un radio alrededor de ésta y se recibe una respuesta.

Para este proyecto se ha utilizado la Streaming API ya que no tiene las limitaciones que tiene la REST API en número de peticiones por unidad de tiempo, por lo que es más fácil para conseguir grandes cantidades de tweets. Además, es la forma que permite cubrir un determinado evento (por ejemplo, unas elecciones o un partido de fútbol). Por otro lado, tiene una desventaja evidente ya que requiere un tiempo de ejecución mayor puesto que el programa tiene que estar en ejecución durante el tiempo que dure el evento que se quiere cubrir, sin embargo, es posible filtrar y analizar posteriormente los mensajes con información geográfica relevante sin obviar los que no tienen esta información por diversos motivos y que, sin embargo, pueden ser igualmente relevantes.

Twitter representa los tweets en formato JSON, es un formato muy sencillo orientado al intercambio de objetos. Un objeto JSON está constituido por una colección de pares de nombre/valor. En varios lenguajes esto es conocido como un objeto, registro, estructura, diccionario, tabla hash, lista de claves o un registro asociativo.

Una vez obtenida la información se pretende filtrar y procesar los tweets para que sean más manejables, de manera que sea posible discriminar la información supuestamente irrelevante, además de simplificar la información.

Los principales métodos que se utilizan en el presente estudio son:

- API.search, que devuelve una colección de Tweets relevantes que coinciden con una consulta específica. Este servicio de búsqueda no está destinado a ser una fuente exhaustiva de Tweets. No todos los Tweets serán indexados o estarán disponibles a través de la interfaz de búsqueda.
- API.streaming, que se utiliza para descargar mensajes de Twitter en tiempo real.

Métodos	Parámetros	Descripción
API.search	q	Filtro de consulta basada en topics, palabras, frases o cuentas de usuario
	[geocode]	Devuelve tweets de usuarios ubicados dentro de un radio dado
	[lang]	Devuelve tweets escritos en el idioma especificado
	[locale]	Se usa para especificar el idioma de la consulta que está enviando
	[result_type]	Tipo de resultados de búsqueda (popular, reciente o mixto)
	[count]	Número máximo de resultados por página
	[until]	Fecha tope de la que se desean recuperar mensajes
	[since]	Fecha desde la que se desean recuperar mensajes
	[max_id]	Devuelve mensajes con id inferior al especificado
	[include_entities]	(V/F) incluir o no el nodo de entidades.
API.streaming	follow	Lista de IDs de usuario separados por comas de los que se quiere recuperar mensajes.
	track	Lista de palabras clave, frases, topics o cuentas de usuario separados por comas a rastrear
	locations	Especifica un conjunto de delimitadores a rastrear.
	delimited	Especifica si los mensajes deben estar limitados por longitud.
	stall_warnings	Especifica si se deben entregar advertencias de bloqueo.

Tbl9.- Métodos y parámetros de Tweepy

4.5.2. Geopy

Geopy es una librería para Python que permite localizar las coordenadas de direcciones, ciudades, países y puntos de referencia en todo el mundo utilizando geocodificadores de terceros y otras fuentes de datos.

Geopy incluye clases de codificación geográfica para OpenStreetMap Nominatim, ESRI ArcGIS, API de codificación geográfica de Google (V3), Baidu Maps, Bing Maps API, Yahoo! PlaceFinder, Yandex, IGN France, Geonombres, NaviData, OpenMapQuest, What3Words, OpenCage, SmartyStreets, geocoder.us, y GeocodeFarm geocoder services.

Entre otras cosas, gracias a esta librería es posible geolocalizar un lugar utilizando dirección y coordenadas, encontrar una dirección correspondiente a un conjunto de coordenadas o, lo que es más importante para nosotros, es posible calcular la distancia geodésica entre dos puntos utilizando las fórmulas distancia Vincenty o distancia ortodrómica. Su valor predeterminado es Vincenty disponible en la clase geopy.distance.distance, y la distancia calculada disponible como atributos (por ejemplo, millas, metros, etc.).

4.5.3. Wikifier

La utilidad Wikifier se trata de una librería para Python que trata de anotar un documento con conceptos relevantes de la Wikipedia. Se utiliza un método basado en pagerank para identificar un conjunto coherente de conceptos relevantes considerando el documento de entrada como un todo. La ventaja de este enfoque es que la Wikipedia es una fuente de información de libre acceso, cubre una amplia gama de temas, tiene una rica estructura interna y cada concepto está asociado con un documento textual semiestructurado que se puede utilizar para ayudar en el proceso de anotación semántica.

Como desventajas podemos mencionar que habría que implementar un algoritmo que, haciendo uso de su librería, tomará como entrada de documento cada uno de los tweets recuperados porque de otra manera habría que tomar un conjunto de ellos como un único documento. La primera opción podría desambiguar de manera inconclusa porque no en todos los mensajes relevantes se encuentran términos existentes en la Wikipedia por la idiosincrasia de las búsquedas requeridas como hemos podido ver con, por ejemplo, el término "Talavear" y sus diferentes acepciones, y la segunda es poco escalable porque, primero tiene límite de 1000 caracteres como parámetro de entrada de documentos y, segundo, no siempre detecta ciertas claves como entidad a desambiguar.

Capítulo 5

En este capítulo detallaremos, por un lado, las técnicas utilizadas en el modelo propuesto y su usabilidad y, por otro lado, los datos obtenidos bajo circunstancias similares y en periodos de tiempo de 24 horas localizado alrededor de la ciudad española de Talavera de la Reina, que son objeto de estudio, resultado de aplicar el modelo propuesto para analizar.

Como se explicaba en el capítulo 4, este proyecto se basa en dos partes, siendo la primera referente al filtrado por geolocalización por *topic terms,* y la segunda referente a la clasificación y rankeo de datos obtenidos, de manera que se detallarán en este orden las técnicas utilizadas.

5.1. Marco experimental

La base de funcionamiento inicial del sistema es similar en ambos casos. Se trata del uso de la librería Tweepy para Python, que hace uso del API de Twitter, con la que es posible recuperar información procedente de Twitter en dos modalidades (i) por descarga directa (API.search) en la versión estándar está limitada a un máximo de 7 días. Además, también se limita en el número de mensajes hasta un máximo de 200 en cada consulta y 900 cada 15 minutos. La modalidad elegida ha sido (ii) por streaming, con las mismas posibilidades de filtrado, pero con lectura directa en cada publicación y sin limitaciones en el número de tweets. Para poder llevar a cabo la recuperación de información es necesario la creación de una aplicación vinculada a Twitter con la que se obtienen las claves y los tokens.

Inicialmente haremos uso de dos conjuntos de tweets para asistir en el análisis de filtrado basado en funciones de pesado y en desambiguación de entidades, todos ellos provenientes de la red social Twitter recogidos en un periodo de 24 horas el 23 de mayo de 2019, durante la campaña electoral para las elecciones municipales de 2019 atendiendo únicamente a la geolocalización manual por parte del usuario, sin tener en cuenta ningún término relevante, obteniéndose, por un lado 333 mensajes en un radio de 20Km de la ciudad de Talavera de la Reina que denominaremos aproximación **Local** y 4304 mensajes para un radio de 150Km que denominaremos aproximación **Comarcal** entre los que, evidentemente, están incluidos los mensajes de la aproximación **Local**. Además, usamos una aproximación a partir de la aproximación **Local** limitada a los 1000 primeros términos para analizarla haciendo uso de la herramienta Wikifier para el filtrado basado en desambiguación de entidades que llamaremos aproximación **1000Local**. Por último,

obtendremos diversas aproximaciones filtrando tweets recogidos en el mismo periodo en base a diferentes topic terms definidos en por medio de heurística.

5.2. Filtrado basado en funciones de pesado de términos

En nuestros experimentos de detección de temas superan significativamente los sistemas de agrupación en clústeres basados en datos textuales para documentos agrupados pertenecientes a un contexto amplio denominados Big Data, pero en documentos de densidad media o baja pertenecientes al conjunto de datos denominados Slow Data, los resultados pueden ser menos significativos.

Para esta tarea se presentan dos vertientes distintas. La primera analizando unigramas de todos los mensajes vertidos en twitter durante un periodo de 24 horas con geoetiquetado por parte del usuario en un radio de 20km y otro de 150km de distancia del centro de la localidad y que se utilizará únicamente para compararlo con el resto y que denominaremos **Conjunto0**.

Como ventaja podemos destacar que de esta manera podemos estar seguros que, gracias a que los mensajes son geolocalizados, la información contenida en éstos tendrá vinculación con la zona de interés. Por otro lado, como desventaja, al no poder categorizar en base a contenidos concretos, cabe esperar que existan muchos términos irrelevantes. En la siguiente representación se muestran los valores obtenidos al calcular el TF-IDF de los 10 términos con mayor frecuencia.

	Local				Comarcal				Agrupado				
	Térm	Frec	TF	TF-IDF	Térm	Frec	TF	TF-IDF	Térm	Frec	TF	IDF	TF-IDF
1	dong	67	0,023	0,023	madrid	222	0,002	0,002	dong	277	0,003	1	0,003
2	dia	10	0,003	0,003	dong	210	0,002	0,002	madrid	224	0,002	1	0,002
3	gracias	8	0,003	0,003	día	152	0,002	0,002	dia	162	0,002	1	0,002
4	3	7	0,002	0,002	ahora	131	0,001	0,001	hoy	134	0,001	1	0,001
5	vida	6	0,002	0,002	hoy	128	0,001	0,001	gracias	132	0,001	1	0,001
6	hoy	6	0,002	0,002	gracias	124	0,001	0,001	ahora	132	0,001	1	0,001
7	mejor	5	0,002	0,002	ver	119	0,001	0,001	mejor	121	0,001	1	0,001
8	nuevo	5	0,002	0,002	mejor	116	0,001	0,001	ver	121	0,001	1	0,001
9	verano	4	0,001	0,001	verano	107	0,001	0,001	verano	111	0,001	1	0,001
10	cosas	4	0,001	0,001	2019	93	0,001	0,001	2019	94	0,001	1	0,001
	TOTAL	**2908**			**TOTAL**	**90571**			**TOTAL**	**93479**			

Tbl10.- TF-IDF de documentos geoetiquetados

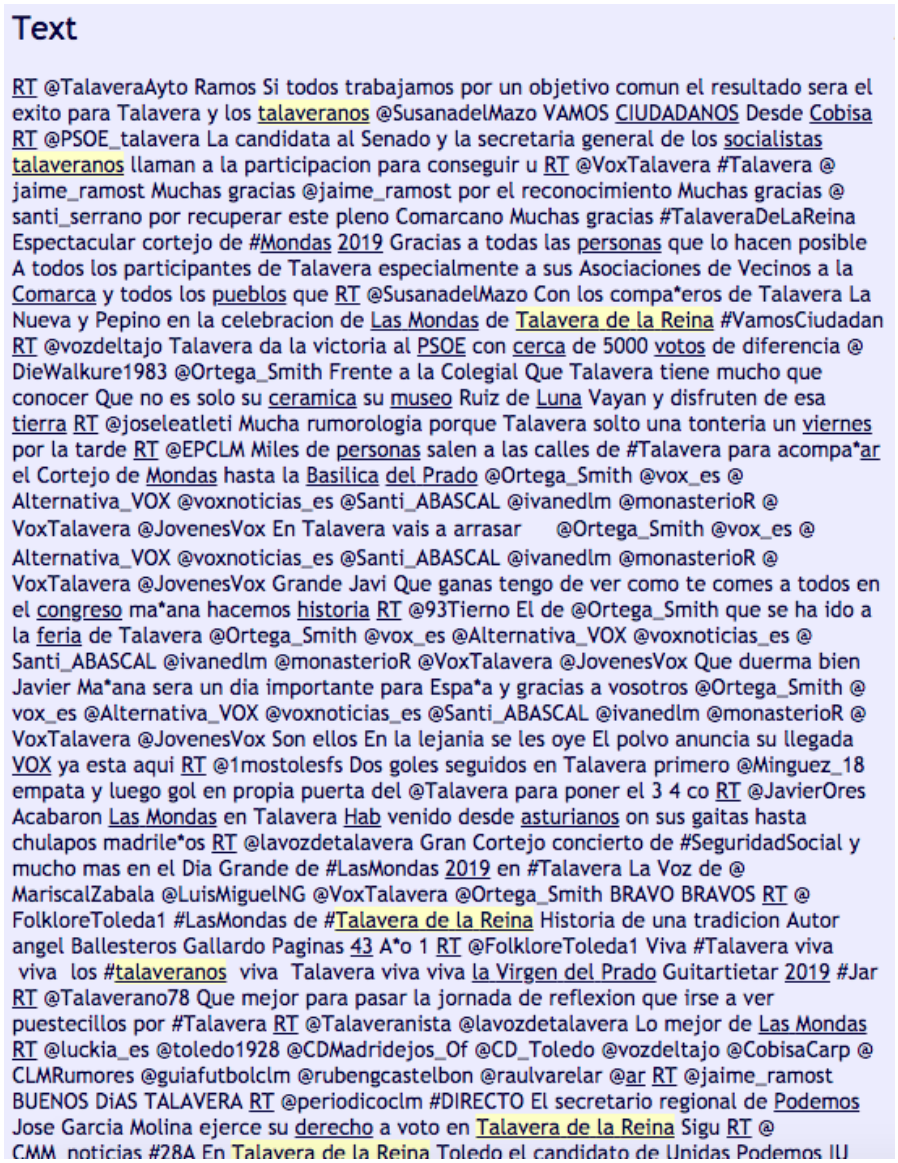
De los 333 documentos recuperados sobre el primer grupo de tweets de la aproximación **local** con un total de 1321 términos diferentes de entre 2908 palabras podemos comprobar únicamente se infiera una vez la palabra Talavera y una vez se menciona al principal equipo de futbol de la ciudad @CFTalavera_ pero en ningún caso se menciona otro tema relacionado directa o indirectamente con la ciudad de Talavera de la Reina, objeto de estudio, ni con ninguno de sus temas de actualidad o cuentas de usuario de interés, de manera que la relevancia de estos datos de baja densidad con geoetiquetado tienen poca.

De los 4304 documentos recuperados sobre el segundo grupo de tweets de la aproximación **comarcal** con un total de 19291 términos diferentes de entre 90571 palabras, eliminando stop-words, podemos extraer las mismas conclusiones que las obtenidas para los tweets locales ya que los datos relativos a Talavera son exactamente los mismos. De hecho, los valores TF-IDF para dichos términos es de 0,00001. Es por este motivo por el que podemos descartar el filtrado por TF-IDF para la primera fase de este trabajo.

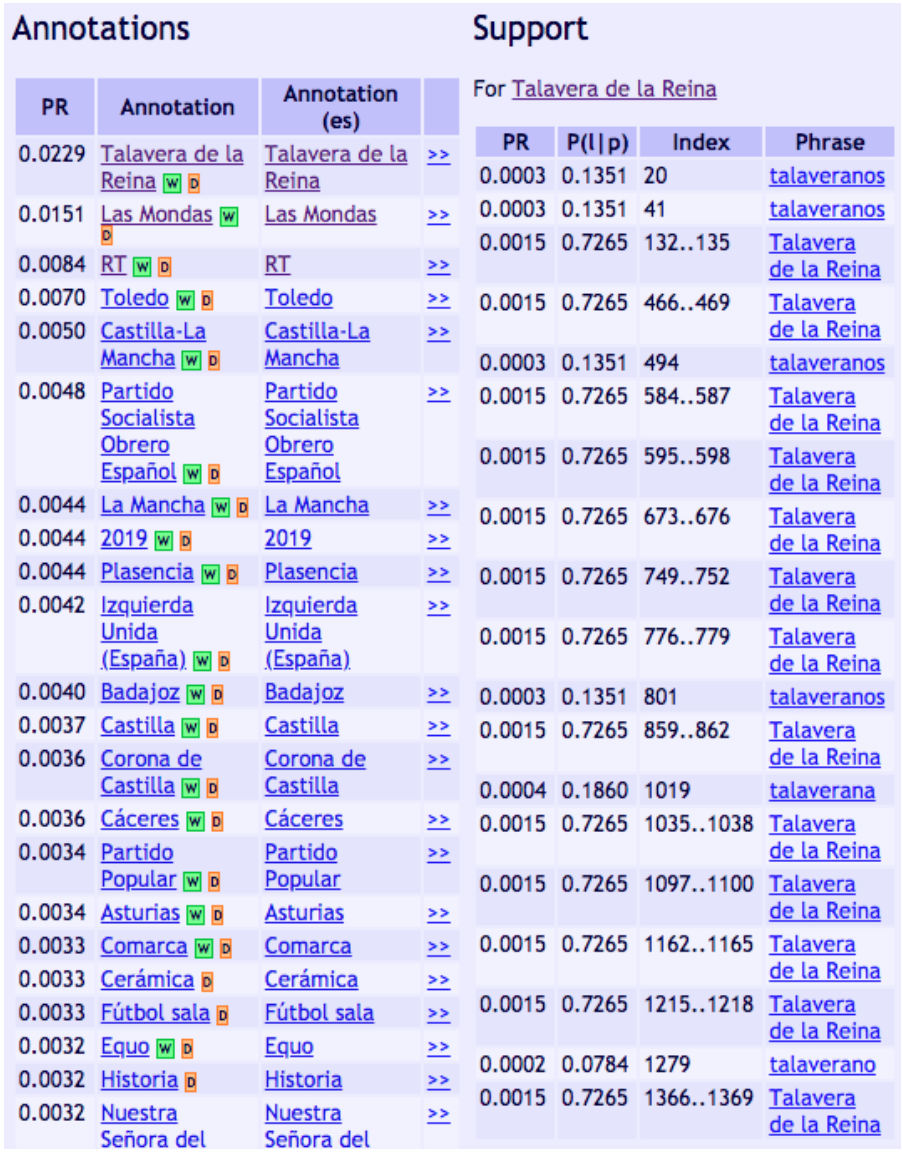
5.3. Filtrado basado en desambiguación de entidades

A partir de los 1000 primeros términos, debido a la limitación de la herramienta Wikifier mencionada anteriormente, de la aproximación **Local**, que hemos llamado **1000Local**

analizamos los resultados de la desambiguación de términos usando dicha herramienta, observamos y analizamos los siguientes resultados:



Img12.- Conjunto de datos analizados con wikifier



Annotations

PR	Annotation	Annotation (es)	
0.0229	Talavera de la Reina W D	Talavera de la Reina	>>
0.0151	Las Mondas W D	Las Mondas	>>
0.0084	RT W D	RT	>>
0.0070	Toledo W D	Toledo	>>
0.0050	Castilla-La Mancha W D	Castilla-La Mancha	>>
0.0048	Partido Socialista Obrero Español W D	Partido Socialista Obrero Español	>>
0.0044	La Mancha W D	La Mancha	>>
0.0044	2019 W D	2019	>>
0.0044	Plasencia W D	Plasencia	>>
0.0042	Izquierda Unida (España) W D	Izquierda Unida (España)	>>
0.0040	Badajoz W D	Badajoz	>>
0.0037	Castilla W D	Castilla	>>
0.0036	Corona de Castilla W D	Corona de Castilla	>>
0.0036	Cáceres W D	Cáceres	>>
0.0034	Partido Popular W D	Partido Popular	>>
0.0034	Asturias W D	Asturias	>>
0.0033	Comarca W D	Comarca	>>
0.0033	Cerámica D	Cerámica	>>
0.0033	Fútbol sala D	Fútbol sala	>>
0.0032	Equo W D	Equo	>>
0.0032	Historia D	Historia	>>
0.0032	Nuestra Señora del	Nuestra Señora del	>>

Support

For Talavera de la Reina

PR	P(l|p)	Index	Phrase
0.0003	0.1351	20	talaveranos
0.0003	0.1351	41	talaveranos
0.0015	0.7265	132..135	Talavera de la Reina
0.0015	0.7265	466..469	Talavera de la Reina
0.0003	0.1351	494	talaveranos
0.0015	0.7265	584..587	Talavera de la Reina
0.0015	0.7265	595..598	Talavera de la Reina
0.0015	0.7265	673..676	Talavera de la Reina
0.0015	0.7265	749..752	Talavera de la Reina
0.0015	0.7265	776..779	Talavera de la Reina
0.0003	0.1351	801	talaveranos
0.0015	0.7265	859..862	Talavera de la Reina
0.0004	0.1860	1019	talaverana
0.0015	0.7265	1035..1038	Talavera de la Reina
0.0015	0.7265	1097..1100	Talavera de la Reina
0.0015	0.7265	1162..1165	Talavera de la Reina
0.0015	0.7265	1215..1218	Talavera de la Reina
0.0002	0.0784	1279	talaverano
0.0015	0.7265	1366..1369	Talavera de la Reina

Img13.- Valores de desambiguación obtenidos con Wikifier

Link Targets

In the Wikipedia, where do links with a particular anchor text point to?

Anchor text: **Talavera** (Where does it occur?)
P(link | phrase) = 0.0732

Note: 6 candidates were ignored because they were linked to less than 2 times (the `minLinkFrequency` parameter) with this phrase as the anchor text.

Note: this phrase was ignored due to being highly ambiguous. Details...

PR	PR (lin.)	Cosine	Target
			Talavera de la Reina
			Arcipreste de Talavera
			Talavera
			Talavera Club de Fútbol
			Cerámica de Talavera de la Reina
			Talavera (Lérida)
			Batalla de Talavera (1809)
			Talavera la Real
			Talavera de la Reyna
			Talavera (Filipinas)
			Tierras de Talavera
			Nuestra Señora de Talavera
			Castilla-La Mancha Fútbol Sala
			Isla Talavera
			Alfredo Talavera
			Distrito de Talavera de la Reyna
			Toma de Talavera
			Talavera Fútbol Sala
			Club de Fútbol Talavera de la Reina
			Talavera de Puebla

Img14.- Posibles alternativas de desambiguación de Wikifier

Como se observa en la primera de las tres imágenes, los términos del documento "Talavera de la Reina", "#Talavera de la Reina", "talaveranos" y "#talaveranos" los reconoce vinculados a la entidad "Talavera de la Reina", sin embargo, "Talavera",

"#Talavera" o "#TalaveraDeLaReina" no los vincula a estos. Además, de entre los términos más relevantes obtenidos, tiene gran importancia los dos primeros, careciendo de importancia para el presente estudio el resto.

Otro ejemplo de desambiguación incorrecta sería con el término "RT" que se trata de un término específico de Twitter a tener en cuenta y que hace alusión a un mensaje reenviado por una cuenta desde otra diferente y que, sin embargo, Wikifier lo desambigua como las siglas de un canal de televisión llamado "Russia Today", con lo que no tiene relación alguna.

Otro problema es que, en los documentos recuperados de Twitter, la clave "Talavera" no siempre se refiere a la primera de las opciones que se vinculan en Wikifier, sin embargo, aunque sea generalmente la opción acertada, en muchas ocasiones se refiere a otros de los descritos anteriormente, con lo que podemos concluir que no es un método certero en este sentido.

Esto coincide, en parte, con el trabajo de (Ferragina et al., 2010) que concluye que casi el 95% de un total de 5.000 tweets analizados tienen al menos 3 frases con una entrada en Wikipedia, aunque esto no implica que sea una entidad, lo que demuestra que Wikipedia tiene una alta cobertura como catálogo para la desambiguación de tweets y se puede usar con precisión explotando su estructura de hipervínculo que pueden ser indexados para su uso en escenarios en tiempo real, sin embargo esto requiere mayor tiempo de análisis y una librería sin restricciones. Además, para conjuntos relativamente altos la cobertura es alta y se puede alcanzar una buena precisión, pero para conjuntos de densidad baja, como es el caso, se observa una precisión y cobertura bajas.

Por último, y muy importante, no es capaz de reconocer entidades como nombres de usuario, cosa que es fundamental para poder desambiguar en Twitter.

5.4. Filtrado basado en heurística

En base a las conclusiones obtenidas anteriormente y, una vez descartados el uso de funciones de pesado de términos y el filtrado basado en desambiguación de entidades, procedemos a la selección de palabras clave, términos relevantes y cuentas de usuario destacadas, basado en heurística tras la consulta de expertos, como ha sido expuesto anteriormente.

5.4.1. Descripción de aproximaciones

Los diferentes conjuntos de tweets utilizados para el filtrado basado en heurística basado en los mismos criterios de adquisición expuestos anteriormente son los siguientes:

- **TalaMundial**: aproximación basada en el filtrado de mensajes por el uso del término "Talavera" en sus propios mensajes o en sus metadatos sin discriminación por distancia, con un total de 53983 términos analizados sobre 3769 documentos o tweets.
- **TalaLocal**: aproximación también basada en radio con filtrados de manera análoga a la aproximación anterior discriminando, de manera automática, los que están escritos fuera de un radio de 20km de la ciudad de Talavera de la Reina o los que están escritos por usuarios que han descrito su ubicación habitual fuera del mismo radio, con un total de 5034 términos analizados sobre 1346 documentos o tweets.
- **PolAnotTalaLocal**: aproximación también basada en radio con filtrados de manera análoga la aproximación anterior, **con polaridad anotada manualmente** atendiendo, en dicha anotación, principalmente a mensajes de índole política y de amplitud delimitada al ámbito local de la ciudad de Talavera de la Reina consistente en 1346 tweets inicialmente capturados. Efectivamente, al tratarse de una ciudad de menos de 100.000 habitantes y de obtener 81 tweets tras la selección de tweets relevantes, estamos en condiciones de datos pertenecientes al grupo de densidad baja o SlowData.
- **PolAnotHeurTalaLocal**: aproximación también basada en radio con filtrados de manera análoga a la anterior aproximación, pero cuyos *topic terms* son "Talavera", "#Talavera", "#TalaveraDeLaReina", "talaverano", "#talaverano", "talavena", "#talaverana", "talaveranos", "#talaveranos", "talaveranas", "#talaveranas", "#mondas", "#lasmondas", "#ceramica", "@talaveraAyto", "@pp_talavera", "@jaime_ramost", "@psoe_talavera", "@titaelez", "@cs_talavera", "@susanadelmazo", "@ahora_talavera", "@sonsolesarnao", "@voxtalavera", "@xtalavera3", "@circulotalavera", "@vozdeltajo", "@lavozdetalavera", "@tribunatalavera" consistente en 7506 tweets inicialmente recuperados de los que se extraen 87 tras la selección de candidatos.

- **PolAnotHeurToleLocal:** aproximación también basada en radio con filtrados de manera similar a la anterior aproximación, pero cuyos *topic terms* son "Toledo", "#Toledo", "@toledoAyto", "@grupopptoledo", "@claudiaalonso", "@psoetoledolocal", "@milagrostolon" "@cs_toledociudad", "@estebanpanos_cs", "@unidastoledo" consistente en 2436 tweets inicialmente recuperados relacionados con la ciudad de Toledo para una evaluación cualitativa para poder exponer un caso de uso, con un total de 103 tras la selección de candidatos.

5.4.2. Resultados de la evaluación

Para analizar la bondad del sistema por medio del recuento de positivos (TP), falsos positivos (FP) negativos (TN) y falsos negativos (FN) para calcular la precisión, cobertura, Exactitud y medida F, tenido en cuenta los datos de las distintas aproximaciones, definiendo como TP los que recibieron, tras la selección de tweets relevantes, puntuaciones 2 o 3 de las aproximaciones **TalaMundial** y **TalaLocal**, no anotadas manualmente. Como FP para las aproximaciones **PolAnotTalaLocal** y **PolAnotHeurTalaLocal** no hay, y para las aproximaciones **TalaMundial** y **TalaLocal** será la diferencia entre el total que tenga puntuación 2 o 3 y los TP. Como TN en las aproximaciones **PolAnotTalaLocal** y **PolAnotHeurTalaLocal** no existen, y en las aproximaciones **TalaMundial** y **TalaLocal** son todos aquellos con puntuación 0 o 1 que no tengan relación con los criterios de búsqueda y filtrado. Por último, serán FN de las aproximaciones **PolAnotTalaLocal** y **PolAnotHeurTalaLocal** los que tengan una puntuación 0 o 1 y de las aproximaciones **TalaMundial** y todos los que tengan puntaución 0 o 1 pero que tengan relación con los criterios de búsqueda y filtrado, obteniéndose los siguientes resultados:

	TP	FP	TN	FN	Prec	Cob	Exac	F
TalaMundial	692	87	2899	91	0,8883	0,8837	0,9528	0,9971
TalaLocal	247	31	1035	33	0,8885	0,8821	0,9525	0,8853
PolAnotTalaLocal	77	0	0	4	1,0000	0,9506	0,9506	0,9746
PolAnotHeurTalaLocal	81	0	0	6	1,0000	0,9310	0,9310	0,9643

Tbl11.- Precisión, Cobertura, Exactitud y F sobre conjuntos 3 y 5

Tal y como se detalla en el capítulo 7, observamos que los resultados de las métricas entre los distintos documentos son realmente satisfactorios en ambos casos, ya que el número de mensajes "perdidos" en la aproximación **TalaMundial**, aunque importante, no es preocupante y podría llegar a mejorarse significativamente, así como los resultados de la aproximación **PolAnotHeurTalaLocal** no son mejores que los de la aproximación **PolAnotTalaLocal**, como se podría esperar.

Capítulo 6. Caso de uso

En esta sección se pretende poner en práctica lo descrito y analizado en el Capítulo 4 y Capítulo 5 por medio de un sistema que devuelva datos clasificados según lo expuesto anteriormente en un periodo de campaña electoral municipal y que puedan ser cuantificar y comparados con datos reales una vez finalizada una campaña electoral municipal.

El sistema o método D'Hondt es un método de promedio mayor para asignar escaños en sistemas de representación proporcional por listas electorales. Los métodos de promedio mayor se caracterizan por dividir a través de distintos divisores los totales de los votos obtenidos por los distintos partidos, produciéndose secuencias de cocientes decrecientes para cada partido y asignándose los escaños a los promedios más altos. Fue creado por el jurista belga Victor d'Hondt en 1878.

Los sistemas de representación proporcional intentan asignar los escaños a las listas de manera proporcional al número de votos recibidos. En general, no es posible alcanzar la proporcionalidad exacta, ya que no es posible asignar un número decimal de escaños.

De los métodos comúnmente utilizados para la conversión proporcional de votos en escaños, el método d'Hondt, siendo bastante proporcional, tiende a favorecer un poco más que otros a los grandes partidos. Sin embargo, hay dos circunstancias que favorecen muchísimo más a dichos partidos: las circunscripciones electorales pequeñas y la barrera electoral.

Tras escrutar todos los votos, se calculan cocientes sucesivos para cada lista electoral. La fórmula de los cocientes es

$$\frac{V}{s+1}$$

Donde **V** representa el número total de votos recibidos por la lista, y **s** representa el número de escaños que cada lista se ha llevado de momento, inicialmente 0 para cada lista.

El número de votos recibidos por cada lista se divide sucesivamente por cada uno de los divisores, desde 1 hasta el número total de escaños a repartir. La asignación de escaños se hace ordenando los cocientes de mayor a menor y asignando a cada uno un escaño hasta que estos se agoten. A diferencia de otros sistemas, el número total de votos no interviene en el cómputo.

Sabiendo, además, que la ley establece los partidos que obtengan menos de un 3% de votos serán eliminados del reparto, y que en poblaciones de entre 50.000 y 100.000 habitantes le corresponden 25 concejales, y las ciudades de Talavera de la Reina y Toledo cuentan con algo más de 80.000, ambos cuentan con características de muestreo similares.

Extrapolando entonces el método D'Hont a las ciudades de Talavera de la Reina y Toledo, obtenemos el siguiente reparto de votos:

	PSOE	PP	Cs	VOX	UP	ATalavera	xTalavera
Votos en Talavera	18175	7749	4294	4737	1778 (<3%)	1171 (<3%)	1029 (<3%)
Votos en Toledo	19258	10976	5463	3604	3505		

Tbl12.- Votos obtenidos en las ciudades de Talavera y Toledo en las elecciones municipales de 2019

Y el siguiente reparto de concejales:

	/1	/2	/3	/4	/5	/6	/7	/8	/9	/10	/11	/12	/13	/14	Total
PSOE	18175 [1]	9087 [2]	6058 [4]	4543 [6]	3635 [9]	3029 [10]	2596 [11]	2271 [14]	2019 [16]	1817 [18]	1652 [19]	1514 [22]	1398 [24]	1298 [25]	14
PP	7749 [3]	3874 [8]	2583 [12]	1937 [17]	1549 [21]	1292	1107	969	861	775	704	646	596	554	5
Cs	4294 [7]	2147 [15]	1431 [23]	1074	859	716	613	537	477	429	390	358	330	307	3
VOX	4737 [5]	2368 [13]	1579 [20]	1184	947	790	677	592	526	474	431	395	364	338	3

Tbl13.- Cálculo de concejales con el método D'Hont en la ciudad de Talavera de la Reina en las elecciones municipales de 2019

	/1	/2	/3	/4	/5	/6	/7	/8	/9	/10	/11	/12	Total
PSOE	19258 [1]	9629 [3]	6419 [4]	4814 [7]	3851 [8]	3209 [12]	2751 [13]	2407 [16]	2139 [18]	1925 [19]	1750 [24]	1604 [25]	12
PP	10976 [2]	5488 [5]	3658 [10]	2744 [14]	2195 [17]	1829 [20]	1568	1372	1220	1098	998	915	6
Cs	5463 [6]	2731 [15]	1821 [21]	1366	1093	911	780	683	607	546	497	455	3
VOX	3604 [9]	1802 [22]	1201	901	721	601	515	451	400	360	328	300	2
UP	3505 [11]	1752 [23]	1168	876	701	584	501	438	389	351	319	292	2

Tbl14.- Cálculo de concejales con el método D'hont en la ciudad de Toled en las elecciones municipales de 2019

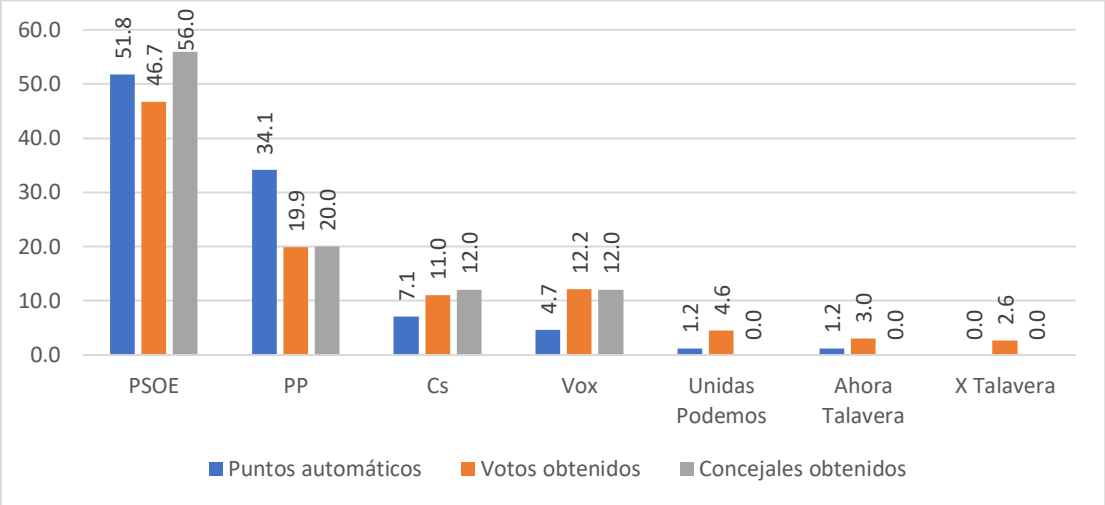
Teniendo como referencia esta información y los datos de la aproximación **PolAnotTalaLocal** con una clasificación y polaridad anotadas manualmente, es posible compararlos con los documentos de la aproximación **PolAnotHeurToleLocal** pertenecientes a un conjunto de documentos extraídos bajo condiciones análogas, clasificados y anotados manualmente en el mismo periodo de tiempo, podemos realizar una evaluación cualitativa tras la que podemos observar y comparar los datos en base a una clasificación automática determinada por los parámetros de coincidencia, localización y popularidad de manera que se puedan asignar valores de entre 0 y 3 a cada uno de los documentos, siendo 0 el valor más bajo y 3 el más alto y distribuyendo dicha puntuación de la siguiente manera:

- Existencia de dos o más de las entidades, hashtags o palabras clave definidas inicialmente
- Localización inferior o igual a 150km
- Número de interacciones superiores a 10

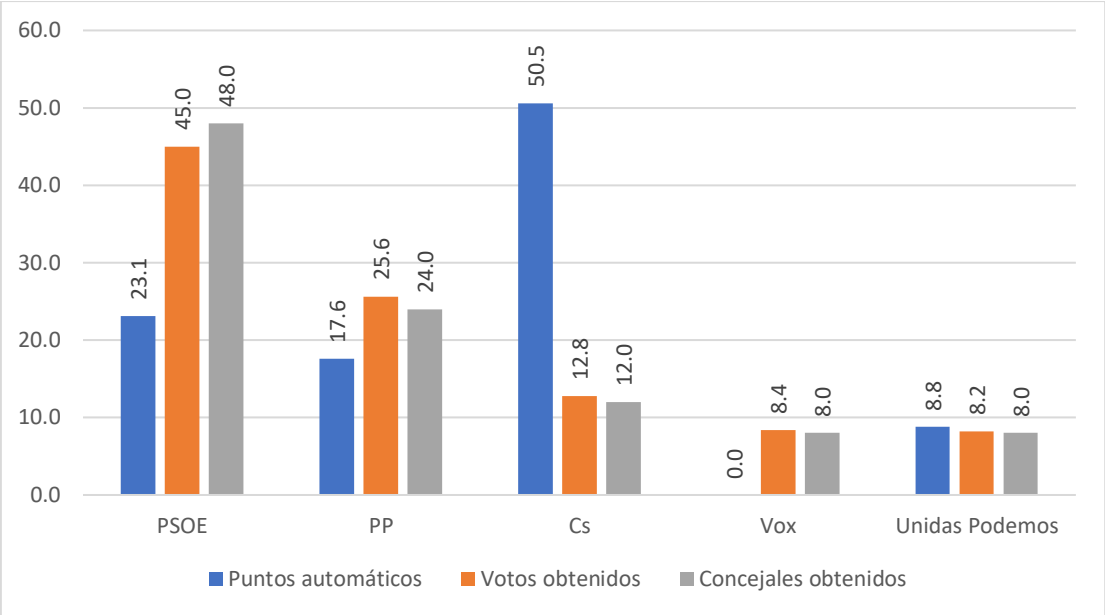
Con estos criterios se ha realizado una comparativa entre las puntuaciones de los tweets de cada ciudad separadas por partido político y observando los resultados electorales en cada ciudad:

	Talavera de la Reina				Toledo			
Partido	Nº Mensajes	Puntos	Votos	Concejales	Nº Mensajes	Puntos	Votos	Concejales
PSOE	39	44	18175	14	28	21	19.258	12
PP	31	29	7749	5	16	16	10.976	6
Cs	4	6	4294	3	30	46	5.463	3
Vox	3	4	4737	3	0	0	3.604	2
Unidas Podemos	2	1	1778	0	11	8	3.505	2
Ahora Talavera	2	1	1171	0				
X Talavera	0	0	1029	0				
TOTAL	**81**	**85**	**38933**	**25**	**85**	**91**	**42806**	**25**

Tbl15.- Comparativa de puntuación y resultados de dos ciudades



Img15.- Comparativa porcentual de los resultados del sistema y los datos obtenidos tras las elecciones en Talavera de la Reina



Img16.- Comparativa porcentual de los resultados del sistema y los datos obtenidos tras las elecciones en Toledo

Es curioso comprobar cómo, a excepción de lo sucedido con Ciudadanos en la ciudad de Toledo por una alta actividad (suponiblemente inusual por una visita con tintes políticos por parte del presidente regional de su partido) y el de Vox en la misma ciudad, que no fue mencionado durante toda la jornada, el resto de valores porcentuales se corresponde significativamente con los datos finalmente escrutados el día de las eleccione

Capítulo 7. Conclusiones y trabajo futuro

En este capítulo se detallan tanto las conclusiones obtenidas al analizar el sistema, así como el trabajo futuro propuesto.

Como se puede comprobar por medio de los datos comparativos entre los documentos de la aproximación **PolAnotTalaLocal** y los documentos de la aproximación **PolAnotHeurTalaLocal**, el hecho de aumentar el número de entidades, temas o usuarios en el momento del filtrado no significa un incremento en la precisión de los datos, ni siquiera un aumento significativo en la diferencia del número de documentos obtenidos

Tras observar todas las posibilidades y analizar los resultados es posible concluir que, teniendo en cuenta la densidad de los datos a valorar en un conjunto de documentos tan limitado para unos términos que no siempre serán fácilmente desambiguables por no ser muy comunes o muy extendidos y por tratarse de cuentas de usuario y temas de ámbito local, es posible concluir que resulta factible que la desambiguación de términos se realice de manera localizada por parte de las personas implicadas y conocedoras del lugar, sus costumbres y tradiciones

Dado que no se ha encontrado un algoritmo que resuelva significativamente favorable la problemática de la detección, extracción y seguimiento de temas debido, en parte, a la densidad de los datos manejados y a la idoneidad de los anteriores para datos pertenecientes a conjuntos de Big Data y, de manera análoga al problema anterior, esta tarea será presumiblemente más precisa si se realiza por personas conocedoras de las problemáticas intrínsecas de los diferentes partidos políticos y de la idiosincrasia de cada localidad objeto de estudio.

Sin embargo, observando los resultados anteriores se extrae que Miura logró la mejor distancia de error mediana para la predicción de la geolocalización, que la mejor precisión fue lograda por Jayasinghe, que también logró una distancia de error mediana muy competitiva. La distancia entre los resultados fue bastante modesta haciendo uso de metadatos debido a que se asigna a una representación a nivel de ciudad y evidencian que los metadatos desempeñan un papel importante en el etiquetado geográfico. A pesar de ello, el método propuesto por Chi proporciona información sobre el rendimiento de los métodos basados en texto puro, por lo que podemos concluir determinando que, a falta de una geolocalización fiable por parte del usuario, ésta resulta una solución muy convincente y útil.

Por último, es evidente que, por un lado, los TN encontrados en la aproximación **TalaMundial** es muy elevando, pero esto es debido a que el sistema recupera una gran cantidad de documentos cuando tienen relación con los términos de filtrado, pero se ha observado que en muchas ocasiones estos términos pueden coincidir, principalmente, con otras ciudades que comparten nombre con Talavera de la reina o apellidos de deportistas, principalmente en Centroamérica, sin embargo, no hay una cantidad muy significativa de FN, los cuales se han detectado, por ejemplo, con mensajes escritos por un usuario vinculado a un periódico local de Talavera de la Reina sin ninguna información relacionada con geolocalización para comprobarla o predecirla y con la que los usuarios no siempre tienen mucha interacción. Los FP han surgido principalmente por mensajes con mucha interacción y con reiteración de términos relacionados con otros lugares o con otras personas como los descritos en los TN.

No obstante, a la luz de los resultados, podemos afirmar que el sistema es capaz de filtrar y rankear de manera satisfactoria los mensajes según los criterios descritos notándose una especial diferencia respecto del conjunto anotado manualmente en cuanto a la cobertura y la MedidaF, lo que indica que el sistema es bastante óptimo, aunque, evidentemente, cabrían muchas mejoras tales como la implementación de algoritmos de predicción de geolocalización más especializados.

Por otro lado, el aumento del número de *Topic Terms* no solo no hace que el sistema mejore, sino que, si la selección de términos se hace correctamente, como en la aproximación **PolAnotTalaLocal**, el incremento de su número, únicamente empeoran la puntuación de las métricas. Esto es lógico hasta cierto punto ya que si, por ejemplo, utilizamos términos vinculados a Talavera de la Reina, como es la cerámica, pero que también lo están con muchos lugares y culturas del mundo, es muy probable que se recuperen algunos mensajes relevantes, pero a costa de muchos más mensajes irrelevante

El trabajo futuro planteado pasa por desarrollar los algoritmos de geolocalicación y realizar un seguimiento exhaustivo dependiente e independiente de características especiales como festividades locales o nacionales o la resolución de una decisión polémica a nivel local, provincial, regional o nacional o la visita de alguna personalidad para comprobar la manera en que esto afecta con el fin de desarrollar un sistema que sea capaz de identificar tanto las virtudes como las debilidades como alertas a tener en cuenta con el fin de mejorar la comunicación con los ciudadanos y solucionar los problemas que puedan surgir de la manera más rápida posible. Además, el preprocesamiento actual tuits es bastante simple y se podría

plantear una normalización gramatical de los tweets que podría ayudar a clasificar la polaridad de manera automática entrenando clasificadores binarios y combinar los resultados obtenidos.

ANEXO

Attributo	Tipo de dato	Descripción
created_at	String	UTC time when this Tweet was created
id	Int64	Integer que representa el identificador único del Tweet
id_str	String	String que representa el id del Tweet
text	String	Texto de la actualización del estado en formato UTF-8
source	String	Herramienta usada para postear el Tweet
truncated	Boolean	Indica si el valor del atributo "texto" está recortado
in_reply_to_status_id	Int64 *(Nullable)*	Si el tweet es una respuesta, indica el integer del ID del tweet original
in_reply_to_status_id_str	String *(Nullable)*	Si el tweet es una respuesta, indica el string del ID del tweet original
in_reply_to_user_id	Int64 *(Nullable)*	Si el tweet es una respuesta, indica el integer del ID del usuario del tweet original
in_reply_to_user_id_str	String *(Nullable)*	Si el tweet es una respuesta, indica el string del ID del usuario del tweet original
in_reply_to_screen_name	String *(Nullable)*	Si el tweet es una respuesta, indica el nombre del usuario del tweet original
user	User object	Objeto del usuario que ha escrito el tweet
coordinates	Coordinates *(Nullable)*	Localización geográfica indicada por el usuario en el momento de la publicación
place	Places *(Nullable)*	Indica que está asociado a un lugar, aunque no necesariamente donde se publicó
quoted_status_id	Int64	Presente si se referencia a otro tweet. Contiene el int del ID del referenciado
quoted_status_id_str	String	Presente si se referencia a otro tweet. Contiene el ID del referenciado como texto
is_quote_status	Boolean	Indica si es un tweet referenciado
quoted_status	Tweet	Presente si se referencia a otro tweet. Contiene el objeto "Tweet" del referenciado
retweeted_status	Tweet	Si es un retweet, contiene la representación del tweet original
quote_count	Integer *(Nullable)*	Indica cuantas veces ha sido referenciado por otros usuarios
reply_count	Int	Indica el número de respuestas asociado al tweet
retweet_count	Int	Indica el número de veces que el tweet ha sido retweeteado
favorite_count	Integer *(Nullable)*	Indica el número de “likes” del mensaje

Attributo	Tipo de dato	Descripción
entities	Entities	Entidades que se han analizado del texto del Tweet
extended_entities	Extended Entities	Matriz de datos multimedia para Tweets referenciados con fotos, vídeos o GIFs
favorited	Boolean *(Nullable)*	Indica si el usuario ha autenticado este tweet
retweeted	Boolean	Indica si el Tweet ha sido retweeteado por el usuario autenticado
possibly_sensitive	Boolean *(Nullable)*	Si el tweet contiene una URL, indica si puede contener información confidencial
filter_level	String	Indica el valor máximo de "Filter_level" que se puede usar para transferir el tweet
lang	String *(Nullable)*	Cuando existe, indica un identificador de idioma BCP 47 detectado.
matching_rules	Array of Rule Objects	Presente en productos filtrados, proporciona la regla que contiene y su identificador
current_user_retweet	Object	Indica el ID del tweet del retweet del usuario (si existe) de este tweet
scopes	Object	Pares clave-valor que indica la entrega contextual prevista del Tweet que contiene
withheld_copyright	Boolean	Si está presente y es "verdadero", indica que se retiene por a una queja de DMCA
withheld_in_countries	Array of String	Si está presente, indica una lista de paises en los que se retiene el contenido.
withheld_scope	String	Cuando está presente, indica si el contenido retenido es el "estado" o un "usuario".

Tbl16.- Campos presentes en un tweet

Attributo	Tipo de dato	Description
id	Int64	Integer que representa el identificador único del usuario
id_str	String	String que representa el id del usuario
name	String	Nombre de usuario definido por el mismo
screen_name	String	Alias que el usuario muestra en su descripción
location	String *(Nullable)*	Localización definida en el perfil del usuario. No es verificada
derived	AEO	Proporciona los metadatos de enriquecimiento geográfico del perfil
url	String *(Nullable)*	URL proporcionada por el usuario para su perfil
description	String *(Nullable)*	Descripción de la cuenta para el perfil del usuario
protected	Boolean	Si es "Verdadero" indica que el usuario proteje sus mensajes
verified	Boolean	Si es "Verdadero" indica que el usuario es propietario de una cuenta verificada
followers_count	Int	Indica el número de seguidores de la cuenta
friends_count	Int	Indica el número de cuentas de usuario a las que sigue
listed_count	Int	Indica el número de listas públicas a las que pertenece
favourites_count	Int	Indica el número de tweets a los que otros usuarios han dado "like"
statuses_count	Int	Indica el número de tweets (incluyendo retweets) del usuario
created_at	String	Indica la fecha de creación de la cuent en formato UTC
profile_banner_url	String	URL donde está la imagen de la página principal del usuario
profile_image_url_https	String	URL donde está la imagen de perfil del usuario
default_profile	Boolean	Si es "Verdadero", indica que el usuario no ha modificado el tema del perfil
default_profile_image	Boolean	Si es "Verdadero", indica que el usuario no ha modificado la imagen del perfil
withheld_in_countries	Array of String	Si está presente, indica una lista de paises en los que se retiene el contenido.
withheld_scope	String	Si está presente, indica que el contenido retenido es el usuario

Attributo	Tipo de dato	Description
utc_offset	null	Deprecated
time_zone	null	Deprecated
lang	null	Deprecated
geo_enabled	null	Deprecated
following	null	Deprecated
follow_request_sent	null	Deprecated

Tbl17.- Campos del objeto “User” de un tweet

Attributo	Tipo de dato	Description
coordinates	Collection of Float	Longitud y latitud del lugar donde se escribe el tweet
type	String	Tipo de dato de las coordenadas *("Point")*

Tbl18.- Campos del objeto “Coordinates” de un tweet

Attributo	Tipo de dato	Description
id	String	Identificador del lugar en formato texto
url	String	URL con información adicional de la localización
place_type	String	Tipo de localización representada
name	String	Nombre aproximado de la representación de la localización
full_name	String	Nombre completo aproximado de la representación de la localización
country_code	String	Representación del código de país
country	String	Nombre del país que contiene la localización
bounding_box	Object	Pares de longitud-latitud que delimita la localización y el tipo de delimitación
attributes	Object	Hash con atributos de localización

Tbl19.- Campos del objeto “Place” de un tweet

8. Bibliografía

(Lozares, 1996) Carlos Lozares
La teoría de Redes Sociales
Universidad Autónoma de Barcelona – 1996

(Whitten y Wolfe, 1988) Whitten y Wolfe
Network Analysis
Icaria – 1988

(Barnes, 1954) John Barnes
Class and Committees in a Norwegian Island Parish
Human Relations, nº7 – 1954

(Requena, 2003) Felix Requena Santos
Orígenes sociales del análisis de redes
CIS – Monografías nº 198 – 2003

(Otte y Rousseau, 2002) Evelien Otte and Ronald Rousseau
Social network analysis: a powerful strategy, also for the information sciences
Journal of Information Science; 28; 441 - 2002

(Scott, 2000) John Scott
Social Network Analysis: A Handbook
Sage Publications Inc, - 2000

(Wasserman y Faust, 1994) Stanley Wasserman, Katherine Faust
Social Network Analysis: Methods and Applications
Cambridge University Press – 1994

(Montoya y Campanha, 2016) J.B. Montoya, T. Campanha
Flujo documental entre áreas administrativas de una entidad bancaria: una aproximación desde el análisis de redes sociales (ARS)
Universidade Estadual Paulista Júlio de Mesquita Filho - UNESP, Brasil – 2016

(Sánchez, 2003) Luis Sánchez Menendez
Análisis de redes sociales: o cómo representar las estructuras sociales subyacentes

AACTE España, Apuntes de Ciencia y Tecnología, Nº 7 – 2003

(Naderi, 2015) Ali M. Naderi

Unsupervised Entity Linking using Graph-based Semantic Similarity
TALP Research Center, Department of Computer Science - Technical University of Catalunya - 2015

(Miceli, 2008) Jorge E. Miceli

Los problemas de validez en el análisis de redes sociales: Algunas reflexiones integradoras

U.B.A. REDES- Revista hispana para el análisis de redes sociales Vol.14, #1, Argentina - 2008

(Plank, 2017) Barbara Plank

Short Text Classification with One Model for All Languages

Center for Language and Cognition - University of Groningen - 2017

(Del Fresno, 2011) Miguel Del Fresno García

Cómo investigar la reputación online en los medios sociales de la web 2.0.

Cuadernos de Comunicación Evoca, vol. 5, n. 1, pp. 29-33 - 2011

(Yu et al., 2009) Ting Yu, Kedar A. Patwardhan, Ser-Nam Lim, Nils Krahnstoever.

Monitoring, recognizing and discovering social networks.

IEEE Computer Society Conference on Computer Vision and Pattern Recognition - 2009

(Carrillo de Albornoz et al., 2014) Jorge Carrillo-de-Albornoz, Enrique Amigó, Damiano Spina, Julio Gonzalo

ORMA: *A Semi-Automatic Tool for Online Reputation Monitoring in Twitter*

UNED NLP & IR Group Madrid, Spain - 2014

(Spina, 2014) Damiano Spina Valenti

Entity-based filtering and topic detection for online reputation monitoring in twitter

PDH Thesis UNED – 2014

(Kleinberg, 2002) J. Kleinberg

Bursty and Hierarchical Structure in Streams

Proc. the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 91–101 - 2002.

(Cerezo Gilarranz et al., 2001) Julio Cerezo Gilarranz, Fernando Polo, David Martínez, Paloma Llaneza, Miguel del Fresno, Delia Rodríguez, Antonio Fumero, Mari Luz Congosto.
Identidad digital y reputación online
Cuadernos de comunicación. Evoca – 2001

(Ferragina et al., 2010) P. Ferragina and U. Scaiella. Tagme
On-the-fly annotation of short text fragments (by wikipedia entities).
Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM'10), pages 1625–1628, 2010.

(Labori, 2016) Javier Labori Marquez
Desambiguación de Entidades Nombradas
Universidad de La Habana - La Habana - 2016

(Saleiro et al., RepLab 2013) Pedro Saleiro, Luís Rei, Arian Pasquali, Carlos Soares, Jorge Teixeira, Fábio Pinto Mohammad Nozari, Catarina Félix, Pedro Strecht
POPSTAR at RepLab 2013: Name ambiguity resolution on Twitter
Universidad Nacional de Educación a Distancia - 2013

(Brank et al., 2017) Janez Brank, Gregor Leban, Marko Grobelnik
Anotando documentos con conceptos relevantes de Wikipedia . Actas de la Conferencia Eslovena sobre Minería de Datos y Almacenes de Datos (SiKDD 2017), Ljubljana, Eslovenia, 9 de octubre de 2017.

(Pérez, 2016) José Alberto Pérez Melián.
Análisis de frecuencia de hashtags en Twitter.
Escola Tècnica Superior d'Enginyeria Informàtica Universitat Politècnica de València - 2016

(Amon et al., 2010) Iván Amon y Claudia Jiménez.

Funciones de Similitud sobre Cadenas de Texto: Una Comparación Basada en la Naturaleza de los Datos
Universidades Pontificias de Bolivia y Colombia. - 2010

(Carrillo de Albornoz et al., 2016) Jorge Carrillo de Albornoz, Enrique Amigó, Laura Plaza, y Julio Gonzalo
Tweet Stream Summarization for Online Reputation Management
Universidad Nacional de Educaci´on a Distancia. - 2016

(Jurgens et al., 2015) David Jurgens, Tyler Finnethy, James McCorriston, Yi Tian Xu, Derek Ruths
Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice
School of Computer Science McGill University - 2015

(Backstrom et al., 2010) Backstrom, Sun, and Marlow
Find me if you can: improving geographical prediction with social and spatial proximity
WWW '10 Proceedings of the 19th international conference on World wide web - 2010

(Mcgee et al., 2011) Mcgee, J., Caverlee, J., and Cheng, Z.
A geographic study of tie strength in social media.
Proceedings of the 20th acm international conference on information and knowledge management. - 2011

(Kong et al., 2014) L. Kong, Z. Liu, Y. Huang. Spot,
Locating social media users based on social network context
Proceedings of the VLDB Endowment, vol. 7(13), - 2014

(Li et al., 2012) Li, R.; Wang, S.; and Chang, K. C.-C.
Multiple location profiling for users and relationships from social network and content.
Proceedings of the VLDB Endowment 5(11):1603–1614. – 2012

(Ishikawa et al., 2012) Shota Ishikawa, Yutaka Arakawa, Shigeaki Tagashira, Akira Fukud
Hot Topic Detection in Local Areas Using Twitter and Wikipedia
Faculty of Information Science and Electrical Engineering - Kyushu University, Japón - 2012

(Bunescu et al., 2006) R. Bunescu and M. Pasca
Using Encyclopedic Knowledge for Named Entity Disambiguation
Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, April 3-7; Trento, Italy - 2006

(Cucerzan, 2007) S. Cucerzan
Large-Scale Named Entity Disambiguation Based on Wikipedia Data
Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30; Prague, Czech Republic - 2007

(Nguyen et al., 2008) T. Nguyen and H. cao
Named entity disambiguation on an ontology enriched by Wikipedia
Proceedings of the IEEE International Conference On Research, Innovation, & Vision for the Future Information & Communications Technol, July 7-11; Hochiminh City, Vietnam - 2008

(Nguyen et al., 2010) T. Nguyen and H. cao
Exploring Wikipedia and text features for named entity disambiguation
Proceedings of the Second international conference on Intelligent information and database systems, March 24-26; Hue City, Vietnam - 2010

(Alhelbawy et al., 2012) A. Alhelbawy and R. Gaizauskas
Named Entity Based Document Similarity with SVM-Based Re-ranking for Entity Linking
Proceedings of the first International Conference on Advanced Machine Learning Technologies and Applications, December 8-10; Cairo, Egypt. - 2012

(Yamanaka et al., 2010) T. Yamanaka, Y. Tanaka, Y. Hijikata, and S. Nishida
A Supporting System for Situation Assessment using Text Data with Spatio-temporal Information
Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 22, No. 6. pp. 691–706, 2010

(Rout et al., 2013) Rout, D.; Bontcheva, K.; Preo¸ tiuc-Pietro, D.; and Cohn, T.
Where's @Wally?: a Classification Approach to Geolocating Users Based on Their Social Ties.
Proceedings of the 24th ACM Conference on Hypertext and Social Media - 2013

(Davis, 2011) Davis Jr, C.; Pappa, G.; de Oliveira, D.; and de L Arcanjo, F.
Inferring the location of twitter messages based on user relationships. Transactions
GIS 15(6):735–751. - 2011

(Amigo et al., 2013) Enrique Amigó , Jorge Carrillo de Albornoz , Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke y Damiano Spina
Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems
Universida Nacional de Educación a Distancia – 2013

(Miura et al., 2016) Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma.
A simple scalable neural networks based model for geolocation prediction in Twitter.
Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), Osaka, Japan. - 2016

(Jayasinghe et al., 2016) Gaya Jayasinghe, Brian Jin, James Mchugh, Bella Robinson, and Stephen Wan.
CSIRO Data61 at the WNUT geo shared task.
Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), Osaka, Japan. – 2016

(Cha et al., 2010) Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K. P.
Measuring User Influence in Twitter: The Million Follower Fallacy.
Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM). Washington DC, USA. - 2010

(Kwak et al., 2010) Kwak, H., Lee, C., Park, H., Moon, S.
What is twitter, a social network or a news media?
WWW '10: Proceedings of the 19th international conference on World wide web. ACM, New York, NY, USA, pp. 591–600. – 2010

(Graham et al., 2013) Mark Graham, Scott A. Hale, Devin Gaffney
Where in the world are you? Geolocation and language identification in Twitter
Oxford Internet Institute, University of Oxford – 2013

(Ritter et al., 2011) A. Ritter, S. Clark, O. Etzioni
Named entity recognition in tweets: an experimental study
Proceedings of the Conference on Empirical Methods in NLP, 2011, pp. 1524–1534.

(Artiles, 2009) Artiles, J.,
Web people search.
Ph.D. thesis, UNED University. - 2009.

(Artiles et al., 2010) Artiles, J., Borthwick, A., Gonzalo, J., Sekine, S., Amigó, E.,
Weps-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks.
NLP Group of UNED University, Madrid, Spain – 2010

(Artiles et al., 2009) Artiles, J., Gonzalo, J., Sekine, S., 2009.
Weps 2 evaluation campaign: overview of the web people search clustering task.
2nd Web People Search Evaluation Workshop, 18th WWW Conference – 2009

(Gooi y Allan, 2004) Gooi, C., Allan, J.,
Cross-document coreference on a large scale corpus.
Proceedings of HLT/NAACL. Vol. 4. – 2004

(Artiles et al., 2007) Artiles, J., Gonzalo, J., Sekine, S.,
The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task.
Proceedings of the 4th international wokshop on Semantic Evaluations pages 64-69. 2007

(Sokolova y Lapalme, 2009) Sokolova, M., y Lapalme, G.
A systematic analysis of performance measures for classification tasks
Information Processing and Management 45 (2009) 427–437 - 2009

(Chi et al., 2016) Lianhua Chi, Kwan Hui Lim, Nebula Alam, and Christopher J. Butler
Geolocation prediction in Twitter using location indicative words and textual features.
Proceedings of the 2nd Workshop on Noisy Usergenerated Text (WNUT), Osaka, Japan. – 2016

(Hannon et al., 2011) J. Hannon, K. McCarthy, J. Lynch, and B. Smyth
Personalized and automatic social summarization of events in video.
Proceedings of the 16th international conference on Intelligent user interfaces (IUI '11), pages 335–338 - 2011

(Chakrabarti and Punera, 2011) D. Chakrabarti and K. Punera
Event summarization using tweets
Proceedings of the fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11), pages 66– 73 - 2011.

(Zhao et al., 2011) S. Zhao, L. Zhong, J. Wickramasuriya, and V. Vasudevan
Human as real-time sensors of social and physical events: A case study of twitter and sports games
Arxiv preprint arXiv:1106.4300 - 2011.
(Kreher y Stinson, 1998) D. L. Kreher and D. R. Stinson
Combinatorial algorithms: generation, enumeration, and search, volume 7
CRC press - 1998.

(Järvelin y Kekäläinen, 2002) K. Järvelin and J. Kekäläinen

Cumulated gain-based evaluation of IR techniques

ACM Transactions on Information Systems (TOIS), 20(4):422–446 - 2002.