Acoplamiento de modelos temáticos en la extracción de opiniones para el análisis de medios sociales

Xujuan Zhou
Escuela de Sistemas de Información,
Universidad del Sur de Queensland, Australia
xujuan.zhou@usq.edu.au

Md Mostafijur Rahman
Escuela de Ciencias Agrícolas, Computacionales y Medioambientales,
Universidad del Sur de Queensland, Australia
Md.Rahman@usq.edu.au

WI '17, Leipzig, Alemania

DOI: 10.1145/3106426.3106459

RESUMEN

Muchas de las plataformas de medios sociales, como Facebook y Twitter, facilitan que todo el mundo comparta sus opiniones sobre, literalmente, cualquier cosa. La detección de temas y opiniones en los medios sociales facilita la identificación de tendencias sociales emergentes y el análisis de las reacciones del público a políticas y productos empresariales. En este artículo, proponemos un nuevo método que combina la minería de opiniones y el modelado de temas basado en el contexto para analizar las opiniones públicas en los datos de las redes sociales. El modelado temático basado en el contexto se utiliza para categorizar los datos en grupos y descubrir comunidades ocultas en el grupo de datos. Los grupos de datos no deseados descubiertos por el modelo temático se descartan. A los grupos de datos restantes se les aplicará un método de minería de opinión basado en léxicos para detectar el sentimiento del público sobre las entidades. En nuestros experimentos se utilizó un conjunto de datos de tuits sobre las elecciones federales australianas de 2010. Nuestros resultados experimentales demuestran que, con la ayuda del modelado temático, nuestro modelo de análisis de redes sociales es preciso y eficaz.

CONCEPTOS DE CSS

• Sistemas de información → Sistemas y herramientas de computación colaborativa y social; - **Metodologías informáticas → Enfoques de aprendizaje en línea**;

PALABRAS CLAVE

Minería de opiniones, Modelización de temas, Análisis de redes sociales, Redes sociales en línea

Formato de referencia ACM:
Xujuan Zhou, Xiaohui Tao, Md Mostafijur Rahman y Ji Zhang. 2017. Coupling Topic Modelling in Opinion Mining for Social Media Analysis. En *las actas de WI '17, Leipzig, Alemania, 23-26 de agosto de 2017*, 8 páginas.
DOI: 10.1145/3106426.3106459

Xiaohui Tao
Escuela de Ciencias Agrícolas, Computacionales y Medioambientales,
Universidad del Sur de Queensland, Australia
xtao@usq.edu.au

Ji Zhang
Escuela de Ciencias Agrícolas, Computacionales y Medioambientales,
Universidad del Sur de Queensland, Australia
ji.zhang@usq.edu.au

1 INTRODUCCIÓN

Las redes sociales han evolucionado hasta convertirse en una fuente de información muy variada [1]. El uso del análisis de los medios sociales para obtener conocimiento de sus datos y tomar decisiones inteligentes es una de las actividades de investigación más interesantes e importantes. La detección de temas y opiniones en los medios sociales facilita la identificación de tendencias sociales emergentes y el análisis de las reacciones del público ante políticas y productos empresariales. Proporciona una forma rápida y fiable de transformar un conjunto de documentos sin etiquetar en una base de conocimientos bien estructurada. El análisis de medios sociales abarca un amplio conjunto de foros en línea, Twitter, Facebook, blogs u otros flujos de texto disponibles públicamente que se rastrean y analizan. Se trata de una plataforma cada vez más popular para transmitir opiniones y pensamientos, por lo que parece natural explorar Twitter en busca de tendencias potencialmente interesantes sobre temas destacados de la actualidad o la cultura popular [14].

Twitter es una de las redes sociales de mayor renombre. Es una red social y de información global, pública, distribuida y en tiempo real en la que los usuarios publican mensajes cortos llamados tweets [21]. Un tuit es un mensaje corto de 140 caracteres. Los usuarios registrados pueden leer y publicar tweets, pero los no registrados sólo pueden leerlos. En los últimos años, Twitter se ha utilizado como fuente de información ideal para conocer los intereses de la sociedad y las opiniones de la gente en general. Las empresas y organizaciones preocupadas por la gestión de la reputación consideran Twitter como una nueva forma potencial de marketing eWOM (boca a boca electrónico) [9]. También se ha utilizado como plataforma de vigilancia en línea para evaluar la opinión de la población sobre cuestiones de salud pública como las vacunas [4, 22] y seguir tendencias sanitarias como los brotes de gripe [16].

Aunque el análisis de sentimientos se centra a menudo en las críticas de películas o productos de consumo, éstas constituyen probablemente una fracción ínfima de los medios sociales. El resto incluye muchos intercambios amistosos en sitios de redes sociales, debates sobre política, deportes y noticias en blogs y foros en línea, así como comentarios sobre medios publicados en Tweeter, Facebook, YouTube, Flickr, Myspace e Instagram [8]. Twitter, un servicio de microblogging, es una de las redes sociales más extendidas [2]. Un modelo exitoso de clasificación de sentimientos basado en los amplios datos de Twitter podría proporcionar

una utilidad sin precedentes tanto para las empresas como para los grupos políticos y los internautas curiosos.

En este estudio, nos centramos en el análisis de sentimientos y la minería de temas basada en el contexto en el análisis de datos de Twitter para un evento electoral en el ámbito político. Se propone un modelo analítico de redes sociales que consta de dos componentes: el análisis de sentimientos y el modelado de temas basado en el contexto. El primer componente adopta un modelo de análisis de sentimientos para evaluar las opiniones del público; el segundo utiliza el modelo de Asignación de Dirichlet Latente (LDA) para estudiar los temas relacionados en los debates públicos en función del contexto. Como base del problema, se estudia el caso de las elecciones australianas a Primer Ministro de 2010 utilizando el modelo de análisis de redes sociales propuesto. Los resultados del estudio son interesantes y el modelo propuesto es prometedor. El trabajo que presentamos en este artículo ofrece una doble contribución, que se describe a continuación.

- Como contribución teórica, proponemos un modelo de análisis de redes sociales de alto nivel que nos ayuda a mejorar nuestra accesibilidad a las opiniones y preferencias del público;
- Como contribución metodológica, el modelo propuesto ayudará a reducir la dimensionalidad en los medios sociales para centrarse
  analítica.

El resto del documento se organiza como sigue. En la sección 2 se hace un breve repaso de los trabajos relacionados. En la Sección 3, se presenta el modelo analítico de medios sociales propuesto con detalles técnicos. El estudio del problema de las elecciones se discute en la Sección 4 con análisis empíricos y estadísticos relacionados. Por último, en la sección 5 se esbozan las conclusiones.

2 TRABAJOS RELACIONADOS

2.1 Opinión Minería

La minería de opiniones (también conocida como análisis de sentimientos) es el estudio computacional de las opiniones, sentimientos y emociones expresados en un texto [13]. El campo del análisis de sentimientos y la minería de opiniones se adapta bien a varios tipos de aplicaciones de inteligencia. De hecho, la inteligencia empresarial parece ser uno de los principales factores que explican el interés de las empresas por este campo [15]. Los enfoques de la minería de opiniones pueden clasificarse a grandes rasgos en métodos basados en el aprendizaje automático y métodos basados en léxicos. Los métodos basados en el aprendizaje automático son tareas de aprendizaje supervisado. Utilizan la representación de características textuales junto con algoritmos de clasificación para inferir las opiniones expresadas en el texto [4, 22]. Las técnicas no supervisadas basadas en léxicos se basan en el supuesto de que la polaridad colectiva de una frase es la suma de las polaridades de las palabras o frases individuales de esa frase [11]. Tanto [19] como [23] propusieron sistemas basados en léxicos no supervisados. Los autores de [19] utilizaron Word Net para clasificar el texto basándose en la suposición de que las palabras con polaridad similar tienen orientaciones similares, pero los autores de [23] han utilizado la pista de subjetividad del léxico de polaridad previa de la lista de léxicos de opinión de Wilson para cuantificar la orientación semántica de las palabras otorgando a cada tipo de palabra una puntuación numérica. Los autores de [17] han utilizado un diccionario creado manualmente etiquetando a mano todos los adjetivos encontrados en su corpus de desarrollo. Todos ellos realizaron un preprocesamiento razonable en sus conjuntos de datos. El modelo basado en el léxico de Wilson superó en velocidad al sistema basado en WordNet y al basado en el léxico creado manualmente, aunque la precisión fue similar.

La minería de opiniones en Tweeter es diferente de la minería de opiniones de blogs y reseñas de productos debido al tamaño del texto. Es difícil diseñar un sistema que analice rápidamente el sentimiento a partir de los datos de Twitter, cuando a los usuarios de Twitter les encanta el uso informal del inglés, el uso de acrónimos, hashtag, palabras escritas de forma innovadora. Este uso informal del lenguaje evoluciona cada día, lo que lo hace aún más difícil. En los grandes conjuntos de datos se ocultan datos no deseados. Estos datos no deseados pueden afectar a la precisión y eficacia del sistema de análisis. Puede hacer que la visualización de los datos sea mucho más compleja de presentar.

2.2 Modelización de temas

El modelado de temas es un tipo de minería de textos, una forma de identificar patrones en un corpus o conjunto de datos. Es un intento de inyectar significado semántico al vocabulario [5]. Tras seleccionar un corpus, se pasa por una herramienta que agrupa las palabras del corpus en temas. Se trata de un método para encontrar y rastrear grupos de palabras (abreviado, "temas") en grandes corpus de textos y, a continuación, agrupar las palabras del corpus en temas. Una herramienta de modelización de temas busca en un corpus estos grupos de palabras y los agrupa mediante un proceso de similitud. En un buen modelo temático, las palabras tienen sentido, por ejemplo "ejército", "tanque", "capitán" y "trigo", "granja", "cultivos". El modelado de temas en sí es una técnica poderosa, pero cuando se combina con técnicas de minería de opiniones puede ser más útil, ya que ayuda a categorizar grandes conjuntos de datos y a detectar patrones ocultos subyacentes en grupos de datos. Una herramienta de modelado de temas como MALLET puede utilizarse para mejorar el proceso de minería de opiniones y análisis de sentimientos. Puede utilizarse para encontrar patrones ocultos en los datos que revelen nuevos conocimientos, así como para dividir los datos en grupos.
que puede ayudar a descartar grupos de datos no deseados.

Los modelos temáticos son una herramienta útil y omnipresente para comprender grandes corpus [7]. Un modelo temático es un mecanismo útil para identificar y caracterizar varios conceptos incluidos en una colección de documentos, lo que permite al usuario navegar por la colección de forma guiada por temas. Según [18], los temas, formados por palabras significativas, proporcionan al usuario una visión general del contenido de la colección de documentos. Cada documento se representa como una mezcla de temas construidos automáticamente y el usuario puede seleccionar documentos relacionados con un tema específico de interés y viceversa. Las similitudes entre documentos pueden encontrarse observando qué documentos están asignados a un tema específico, lo que permite al usuario encontrar otros documentos relacionados con un documento determinado. Esta metodología permite a los usuarios digerir un mayor número de documentos, ayudándoles a dedicar más tiempo a la lectura que a la búsqueda de información relevante.

La técnica de modelado de temas y la técnica de minería de opiniones se han utilizado en el método propuesto por [12], donde mostraron un nuevo marco de modelado probabilístico basado en LDA, llamado modelo conjunto de sentimiento/tema (JST), que detecta el sentimiento y el tema si- multáneamente a partir del texto. A diferencia de otros enfoques de aprendizaje automático para la clasificación de sentimientos, que a menudo requieren corpus etiquetados para el entrenamiento del clasificador, el modelo JST propuesto es totalmente no supervisado. Este trabajo es algo parecido al nuestro, ya que utilizaron el modelado temático para encontrar grupos ocultos de datos en el conjunto de datos, pero su minería de sentimientos es a nivel temático y no a nivel de entidad de los tweets.

Como se desprende de estos debates, la clave del éxito es la capacidad de tomar decisiones rápidas e inteligentes, que se puede conseguir adoptando las técnicas de análisis de sentimientos y de análisis de opiniones.

modelización temática basada en el contexto. El trabajo que presentamos en este artículo está motivado por la demanda de dicha clave.

3 MODELO ANALÍTICO DE MEDIOS SOCIALES

3.1 Arquitectura de alto nivel

El modelo de análisis de redes sociales propuesto consta de dos partes: el modelo de análisis de sen- tidos y el modelado de temas basado en el contexto. Twitter ofrece una forma sencilla de acceder a través de la interfaz de programación de aplicaciones (API), que puede utilizarse para interactuar con el servicio muy fácilmente. El conjunto de tweets se descarga de Twitter a través de su API. En primer lugar, el conjunto de datos se preprocesa, lo que incluye tareas como la limpieza de datos para eliminar los datos ruidosos (por ejemplo, puntuaciones y símbolos), la eliminación de palabras vacías y la separación de palabras. A continuación, el conjunto de datos se analiza en dos módulos: el análisis de sentimientos y el modelado temático basado en el contexto. La figura 1 ilustra el marco del modelo propuesto.

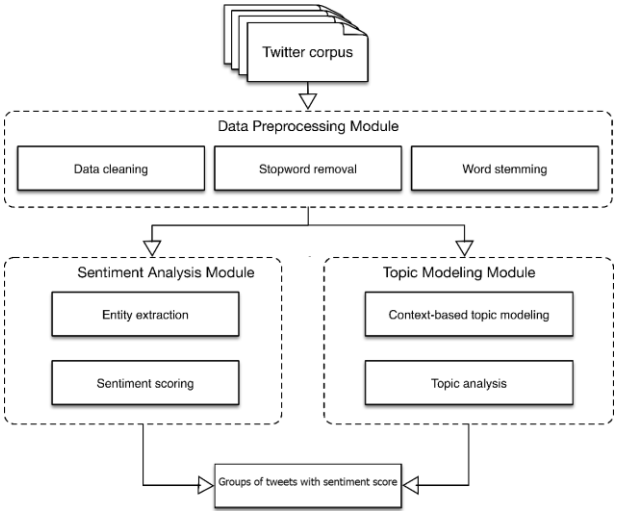

Figura 1: Arquitectura de alto nivel

3.2 Análisis del sentimiento

Como se muestra en la Fig. 1, en la primera etapa, se utilizará el Modelo de Análisis de Sentimiento de Tweets (TSAM) desarrollado por [23] para calcular el sentimiento y puntuar la entidad. Debido a la limitación de espacio, brevemente

Describimos TSAM aquí y se puede encontrar un proceso de análisis de sentimiento más detallado en [23].

recurso léxico de información de sentimiento para palabras, donde cada palabra se asocia con información de sentimiento positiva, negativa y neutra. En este proyecto, sólo se utiliza la pista de subjetividad del léxico de polaridad previa. La orientación semántica de las palabras se cuantificó dando a cada tipo de palabra una puntuación numérica. Así, a una palabra positiva y de fuerte subjetividad se le asigna la puntuación de orientación semántica de +1, a una palabra positiva y de subjetividad débil se le asigna la puntuación de

puntuación de la orientación semántica de +0,5, y una sub- orientación negativa y fuerte.

jectividad se le asigna la puntuación de orientación semántica de -1, a

A una palabra negativa y de subjetividad débil se le asigna una puntuación de orientación semántica de -0,5, y a una palabra neutra se le asigna una puntuación de orientación semántica de 0. Estas cadenas de texto pueden clasificarse en categorías.

gorías (positiva, negativa, neutra) y se puede diferenciar su fuerza o impacto asignándoles distintos pesos. Por ejemplo, la palabra "quiebra" puede tener un valor de ponderación mayor que "demanda", aunque ambas pertenezcan a la categoría "Negativa".

Dado un conjunto de tweets, T , que contiene un conjunto de frases, $T =$

$\{s1, s2, ..., si\}$; y cada frase si describe algo sobre un subconjunto de entidades $e = \{ei ...ej \mid ei , ej \in E\}$, donde E es el conjunto de todas las entidades. Una entidad puede ser una persona, una organización, una ubicación, un producto, un

acontecimiento, etc. Cada frase contiene también un conjunto de palabras de opinión, wk , $s =$

$\{w_1 , w_2 , ..., wl\}$. En primer lugar, se utiliza una función de puntuación del sentimiento de la frase (SSSF, Sentence Sentiment Scoring Function) para determinar la orientación del sentimiento expresado en cada

entidad ei en s (es decir, el par (ei , s)). A continuación, se utiliza una Función de Agregación de Sentimiento de Entidad (ESAF) para obtener las puntuaciones totales de sentimiento para una entidad ei dada [23].

3.2.2 Función de puntuación del sentimiento de la frase. En esta fase, el algoritmo de clasificación detecta todas las palabras que pertenecen a la lista del léxico de Wilson y extrae su polaridad. Los adjetivos son buenos indicadores del sentimiento y varios investigadores los han utilizado como características para la clasificación de sentimientos [10], [6]. Sin embargo, esto no implica necesariamente que otras partes del discurso no contribuyan a las expresiones de opinión o sentimiento. De hecho, los sustantivos (por ejemplo, "gema") y los verbos (por ejemplo, "amor") pueden ser buenos indicadores del sentimiento. Por eso, en este estudio utilizamos todas las partes de la oración. Sumamos la puntuación de orientación semántica de las palabras de opinión de la frase para determinar la orientación de la frase de opinión. La función de puntuación de una frase es la siguiente:

$$puntuación(es) = \sum_{(wj : wj \in s \wedge wj \in WL)} \frac{wj - sentOri}{dis(wj , ei)} \quad (1)$$

3.2.1 Extracción de características. En el modelo TSAM, en

lugar de utilizar todas las palabras que aparecen en los artículos de noticias o tuits, el TSAM sólo extrae las palabras de opinión como características para introducirlas en el algoritmo de minería de opinión. Se identifican y extraen las palabras de opinión que se utilizan principalmente para expresar opiniones subjetivas en la frase de opinión. Las palabras que codifican un estado deseable (por ejemplo, hermoso, impresionante) tienen una orientación positiva, mientras que las palabras que representan estados indeseables tienen orientaciones negativas (por ejemplo, decepcionante, horrible).

Para identificar las palabras con opinión, se utilizó la lista del léxico de opinión de Wilson [20] para decidir las orientaciones semánticas de las palabras. Esta lista es un

donde $_{wj}$ es una palabra de opinión, WL es el conjunto de todas las palabras de opinión de la lista del léxico de Wilson y s es la frase que contiene la entidad $_{ei}$, y $dis(_{wj}, _{ei})$ es la distancia entre la entidad $_{ei}$ y la palabra de opinión $_{wj}$ en la frase s, y $_{wj}.sentOri$ es la orientación semántica de la palabra $_{wj}$ (es decir, +1, o +0,5, o 0, o -1, o -0,5). Si una frase contiene más de una entidad, entonces la palabra de opinión cercana a la entidad tiene un valor menor de $dis(_{wj}, _{ei})$ e indica que esta palabra contribuye más a las puntuaciones de sentimiento de esa entidad.

Las $puntuaciones(s)$ se normalizan por el número de palabras de opinión, n, en la frase para reflejar las distribuciones de las puntuaciones de sentimiento de las palabras de opinión. Así, la puntuación de sentimiento normalizada será:

$$puntuación(s)_N = puntuación(s) \div n \quad (2)$$

3.2.3 Función de agregación del sentimiento de la entidad. En el conjunto dado de tuits, una entidad aparece en el conjunto de frases $s = \{s1, s2, ..., si\}$. Utilizamos la co-ocurrencia de una entidad y una palabra de sentimiento en la misma frase.
significa que el sentimiento está asociado a esa entidad. Esto no siempre es exacto, sobre todo en frases complejas. Aun así, el volumen de texto que procesamos nos permite generar puntuaciones de sentimiento precisas.

Para una entidad determinada ei, que puede aparecer en varias frases
$\{s1, s2, ..., si\}$, la puntuación de sentimiento normalizada para esta entidad en la frase sk es $score(ei, sk)_N$. La puntuación total del sentimiento de esta entidad se agregará mediante la función de agregación del sentimiento de la entidad
que se representa a continuación:

$$puntuación(_{ei}) = \sum_{(sk : sk \in s)} puntuación(_{sk})_N \quad (3)$$

Esta puntuación se normaliza por el número de frases, *m*, y luego la puntuación final del sentimiento de una entidad se situará en el rango
intervalo [+1, -1].

$$score(_{ei})_N = score(_{ei}) \div m \quad (4)$$

Con respecto a la intensidad (o fuerza) del sentimiento para una entidad determinada, ei, aparece en las frases, se aplica la siguiente regla heurística:

$$intensidad(_{ei}) = \begin{cases} \text{SPif } (+0{,}5 < puntuación(_{ei})_N < +1) \\ \text{Pif } (0 < puntuación(_{ei})_N < +0{,}5) \\ \text{Neuif } (puntuación(_{ei})_N = 0) \\ \text{Negif } (-0.5 < puntuación(_{ei})_N < 0) \\ \text{SN si } (-1 < puntuación(_{ei})_N < -0{,}5) \end{cases}$$

- *SN (Negativa Fuerte)* Las frases sobre la entidad ei contienen palabras o frases puramente negativas o sólo se permite una palabra ligeramente positiva.
- *N (Negativa)* Las frases contienen principalmente frases y palabras negativas. Puede haber algunas palabras positivas, pero la
las palabras o frases negativas superan a las positivas.
- *Neu (Neutral)* Las frases tienen un sen- timiento mediocre o equilibrado. Las palabras o frases positivas y negativas parecen equilibrarse entre sí, o no es ni positivo ni negativo en general. Aunque haya más frases negativas, las positivas utilizan un lenguaje más fuerte que las negativas.
- *P (Positivas)* Las frases tienen principalmente términos positivos. Puede haber algunos negativos; sin embargo, los positivos son
más fuertes y pesan más que las negativas.
- Las frases *SP (Strong Positive)* tienen palabras puramente positivas que expresan fuertes sentimientos afirmativos sin quejas. Puede tener pocas palabras negativas, pero la frase
tiene en su mayoría palabras o frases que suenan muy bien.

es un paquete de código abierto basado en Java para aplicaciones estadísticas de procesamiento de lenguaje natural en texto. Incluye una implementación extremadamente rápida y escalable del muestreo de Gibbs, métodos eficientes para la optimización de hiperparámetros documento-tema y herramientas para inferir temas para nuevos documentos a partir de modelos entrenados. En nuestro conjunto de datos de tuits, hay muchos tipos de entidades ("cargo", "empresa", "lugar", etc.) además de personas (Tony Abbott, Julia Gillard, Rick James, Kevin Rudd, etc.). Nuestras entidades objetivo son únicamente dos candidatos a Primer Ministro (Julia Gillard y Tony Abbott). Por lo tanto, utilizando el modelo LDA, se produjeron tres modelos temáticos: dos modelos temáticos de dos candidatos a primer ministro y un modelo temático que no está relacionado con los candidatos a primer ministro. Las palabras de estos modelos temáticos se utilizan para categorizar los tweets en
encontrar comunidades y patrones ocultos subyacentes.

Por ejemplo, la Tabla 1 muestra tres modelos temáticos. Cada modelo temático
contiene un conjunto de palabras que se muestran en las columnas uno, dos y tres, respectivamente. El número de palabras que un tuit tiene en común en un modelo temático es la puntuación de ese tuit frente a los tres modelos temáticos. Si un tuit tiene cinco palabras comunes en el modelo temático 0, dos palabras comunes en el modelo temático 1 y una palabra común en el modelo temático 2, ese tuit pertenece al grupo 0. El mayor número de palabras comunes en un modelo temático determina a qué grupo pertenecen los tuits. Si hay dos puntuaciones iguales en dos modelos temáticos
entonces ese tuit pertenece a ambos grupos de tuits.

La Tabla 2 muestra la puntuación frente a los modelos temáticos. El ID de línea (Line#) y la frase del tuit con la entidad se muestran en la columna uno. Las puntuaciones de los modelos temáticos 0, 1 y 2 se muestran en las columnas dos, tres y cuatro, respectivamente.

La comunidad de tuits denominada Grupo 0 contiene la mayoría de los

3.3 Modelización temática basada en el contexto

En la segunda fase, los resultados del TSAM serán utilizados por el componente de modelización de temas para el proceso de categorización de los tweets con vistas a la detección de comunidades ocultas. En este proyecto de investigación, se adopta el modelo de Asignación de Dirichlet Latente (LDA), una técnica de vanguardia para el modelado de temas basado en el contexto [3]. En la práctica, utilizamos MALLET, una de las herramientas de minería de texto más populares y de libre acceso, para gestionar la tarea de modelado temático basado en el contexto. MALLET

tweets sobre Tony Abbott, la comunidad denominada Grupo 2 contiene la mayoría de los tweets sobre Julia Gillard, y la comunidad de tweets denominada Grupo 1 contiene la mayoría de los tweets sobre los demás. Como estamos más interesados en los dos candidatos a primer ministro, Tony Abbott y Julia Gillard, podemos descartar la comunidad de tuits denominada Grupo 1. Este descubrimiento de la detección de una comunidad oculta de datos en nuestro conjunto de datos nos inspiró para realizar más modelizaciones temáticas en los grupos categorizados para profundizar en el descubrimiento de cualquier otra comunidad oculta subyacente. Curiosamente, encontramos un patrón útil en el Grupo 2, el patrón nos da más información sobre los datos para subcategorizar este grupo y darle un sentido más profundo que no se podía haber visto antes. Las puntuaciones de sentimiento obtenidas en la primera etapa se han inyectado en las comunidades de tuits para ver la polaridad de la opinión pública sobre las entidades de la comunidad.

Para detectar el patrón oculto subyacente o la comunidad de tweets en estos tres grupos de tweets recién categorizados, se realizaron modelos temáticos adicionales en cada grupo de tweets. Así, se crearon dos submodelos temáticos de cada grupo de tuits.

Tras categorizar dos modelos temáticos del Grupo 02, descubrimos comunidades ocultas de tuits; una comunidad de tuits habla predominantemente de una cosa, mientras que la otra comunidad de tuits habla de otra cosa. Se trata de un hallazgo importante; la técnica de minería de opiniones por sí sola no es capaz de conseguirlo; necesita acoplarse a la técnica de modelado de temas para lograrlo.

Tabla 1: Modelos temáticos con el conjunto de palabras incluidas en cada modelo producido por LDA

Tema Modelo 0	Tema Modelo 1	Tema Modelo 2
0.00432	0.00317	0.00405
ausvotes abbott persona	ausvotes mujer	ausvotes gillard julia
empresa abc elecciones tony	persona de tecnología	ministro house position
streaming cassidy crabb	censura internet	persona primer gobierno
cobertura en directo twitter	posición obligatoria bit	convocatoria de elecciones
tecnología verde uhlmann	james historia rick ir	llega industryterm
noticias bit medios de comunicación en línea	elección punter sucedió	abc news conducir gobierno

Tabla 2: Puntuación de las palabras coincidentes de los tweets según los modelos temáticos

Línea# (Tweet#)	Tema Modelo 0	Tema Modelo 1	Tema Modelo 2
Línea# 1 Tony Abbott: Lamento decir que rompimos la fe con el Howard battlers #ausvotes|Tony Abbott;Person	5	2	2
Line# 2 @theburgerman: Work Choices no sólo muerto sino incinerado dice Tony Abbott #ausvotes|Tony Abbott;Person	6	2	2
Línea # 3 @skynewsaust: Watch Tony Abbott online here en una conferencia del LNP de Qld #ausvotes|http;Technology	6	2	2
Line#29 así que mi primera elección federal y voy a ser 15000km ¡en todo el mundo no les gusta! #ausvotes|federal election;PoliticalEvent	8	2	2
Línea # 46 ¿VAMOS a apagar a Rick James por esto? Por esto la democracia APESTA #ausvotes|Rick James;Person	3	4	3
Línea# 57 Graham Richardson dice que Gillard es lo suficientemente australiana como para entrar en un bar en cualquier parte y "charm em" #ausvotes|Graham Richardson;Person	2	6	2

4 RESULTADOS EXPERIMENTALES Y DISCUSIONES

4.1 Fuente de datos y preprocesamiento

El conjunto de datos original utilizado en este trabajo se compone de 57.000 tuits australianos de un periodo de dos semanas durante las elecciones federales australianas de 2010. Cuando se anunció la fecha de las elecciones, el 21 de agosto de 2010, el 17 de julio de 2010, Twitter experimentó una oleada de tweets. Este conjunto de datos se compone concretamente de dos semanas de tuits, desde el sábado 17 hasta el sábado 31 de julio. Los tuits con el hashtag *#ausvotes* han sido con-
sidered only. Aunque podría haber tweets sobre la elección
sin ese hashtag. Nos falta la puntuación de sentimiento de esos tuits. Los datos de los tweets recopilados se dividieron en 57 archivos, cada uno de los cuales contenía unos 1.000 tweets, lo que hace un total de 57.000 tweets. Todos los tweets se ordenaron cronológicamente por id de tweet dentro del conjunto de datos para el experimento.

Una tarea no trivial de la recopilación de datos de tuits para el análisis de sentimientos es la extracción de las entidades relevantes de los tuits. Para identificar y extraer las entidades que aparecen en las frases se utilizó Open Calais. Open Calais es una de las herramientas de extracción de entidades más conocidas. Open Calais extrae entidades de la entrada textual (lenguaje natural) y devuelve un documento XML que contiene metainformación sobre las entidades en formato RDF, incluido el nombre y el tipo. Encontrará información detallada en http://www.opencalais.com. En la primera etapa se utilizaron los 57000 tweets originales. Después de la primera etapa, quedan 457 tweets relevantes con sus puntuaciones de sen- timiento y entidades etiquetadas para la stage ha aplicado la técnica de modelado temático al conjunto de datos restante para descubrir el patrón oculto subyacente y comprender mejor lo que el público intenta decir.

Al conjunto de datos de tweets se le aplicaron técnicas de preprocesamiento de datos como la eliminación de palabras vacías, la eliminación de signos de puntuación y símbolos, y la eliminación de la raíz de los tweets. La Tabla 3 muestra qué partes se eliminaron de los tweets.

Medimos el rendimiento del sistema con su precisión de la siguiente manera:

$$precisión = \frac{NTSCL}{TNTTS} \quad (5)$$

segunda etapa. La segunda

donde NTSCL es el número de tuits que el sistema ha etiquetado correctamente y TNTTS es el número total de tuits de un conjunto de prueba.

4.2 Resultados del conjunto de datos categorizados

El modelo temático puede ayudar a categorizar grandes conjuntos de datos para descubrir grupos de datos no deseados. En este estudio, el conjunto de datos de tweets se clasifica en tres modelos temáticos: Grupo 0, Grupo 1 y Grupo 2. Las figuras 2, 3 y 4 representan todas las entidades del Grupo 0, Grupo 1 y Grupo 2, respectivamente. Se observa que la mayoría de los tweets del grupo 0 son sobre Tony Abbott, la mayoría de los tweets del grupo 1 son sobre cualquier persona que no sea Julia o Tony, y la mayoría de los tweets del grupo 2 son sobre Julia Gillard. En el grupo 0, la entidad objetivo es el candidato a primer ministro Tony Abbott, que obtuvo un 56% de opiniones positivas, un 38% de opiniones negativas y un 6% de opiniones neutras (véase la Fig. 2). En el grupo 2, la entidad objetivo es la candidata a primera ministra Julia Gillard, que obtuvo un 60% de opiniones positivas, un 24% de opiniones negativas y un 16% de opiniones neutras (véase la Fig. 4). Por orden

Tabla 3: Preprocesamiento de datos para eliminar partes no deseadas de los tweets

Muestra	Tipo	Tarea
A, an, on, at, in etc.	Palabras clave	Estas palabras no llevan ningún sentimiento
RT	Retweet	Volver a publicar el tweet de otro usuario
@	Mencione	Etiqueta utilizada para mencionar a otro usuario
#	Hashtag	Los hashtags se utilizan para etiquetar un tuit con un tema determinado.
URL	URL	Normalmente, un enlace a un recurso externo
, . ? ;	Puntuación	Estos no llevan ningún sentimiento
&, *, \,), (, $	Símbolo	Estos no llevan ningún sentimiento

para realizar análisis posteriores en el Grupo 0 y el Grupo 2 de forma más eficiente, se descartó el Grupo 1, ya que era un grupo no deseado.

Al comparar las figuras 2 y 4 se observa que Julia Gillard tiene un 64% de votos positivos, un 24% de negativos y un 16% de neutros, mientras que Tony Abbott tiene un 54% de votos positivos, un 38% de negativos y un 6% de neutros. Esta comparación sugiere que Julia Gillard tiene más posibilidades de ganar las elecciones que Tony Abbott y, en realidad, así fue en las Elecciones Federales Australianas de 2010.

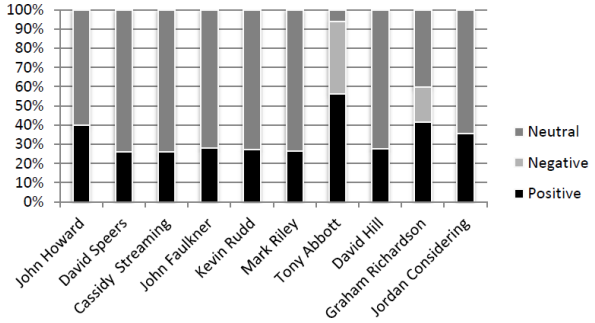

Figura 2: Comparación porcentual del sentimiento de las entidades del Grupo 0

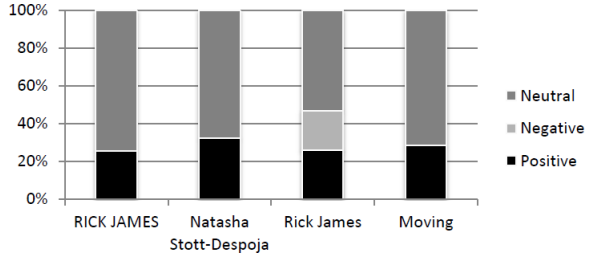

Figura 3: Comparación porcentual del sentimiento de las entidades del Grupo 1

La Fig. 5 es la comparación general de todas las entidades que teníamos antes de hacer el modelado temático; las entidades

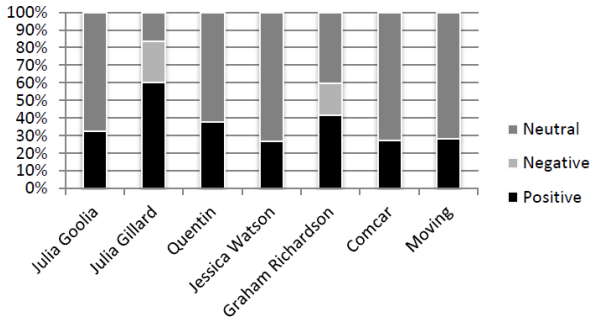

Figura 4: Comparación porcentual del sentimiento de las entidades del Grupo 2

menos interesantes agrandan el gráfico. Fig 6 es la representación de la comparación de todas las entidades después de hacer el modelado temático y descartar el grupo menos interesante de entidades; haciendo el gráfico más pequeño para presentar.

sin afectar al resultado. Incluso en este pequeño conjunto de datos, podemos eliminar el 21% de las entidades y, tras realizar el modelo temático, presentar un gráfico menos abarrotado sin que ello afecte al resultado. La figura 5 muestra todas las entidades, mientras que la figura 6 es una versión más clara con sólo los temas y comunidades interesantes.

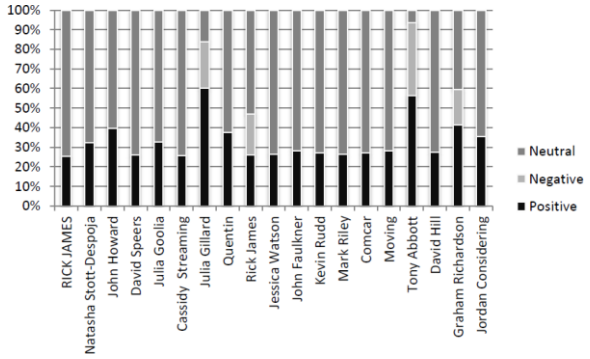

Figura 5: Gráfico con el grupo de entidades no deseadas antes de la modelización temática

4.3 Modelización temática para la detección de comunidades

La modelización temática adicional de los grupos categorizados se llevó a cabo para profundizar en el descubrimiento de cualquier otra comunidad oculta subyacente. Curiosamente, encontramos un patrón útil en el Grupo 2, que nos permite conocer mejor los datos para subcategorizarlo.

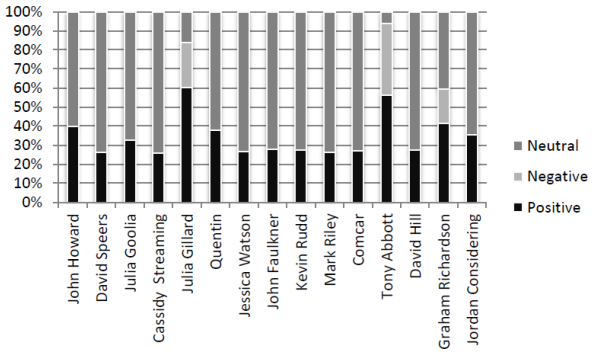

Figura 6: Gráfico tras descartar el grupo de entidades sin interés mediante el modelado por temas

grupo y adquieren un sentido más profundo que no se podía haber visto antes. La Fig. 7 es el resultado de la modelización temática del grupo 2 (grupo de Julia Gillard). El esfuerzo de hacer más modelización de temas en ese grupo reveló una comunidad oculta de tuits en ese grupo. El sesenta y tres por ciento de los tweets de ese grupo se referían a la candidatura de Julia al cargo de Gobernadora General (GG) y el treinta y siete por ciento restante a otros intereses electorales.

Figura 7: Subgrupo oculto en el grupo de Julia Gillard

4.4 Debates

Con la ayuda del modelado temático, pueden descartarse los temas, entidades o tuits no deseados. De este modo, se puede mejorar la precisión de los cálculos y el sistema puede funcionar más rápido, ya que manejará menos datos cuando se trate de conjuntos de datos muy grandes. El modelo temático puede descubrir grupos o patrones ocultos agrupando temas. Por ejemplo, el conjunto de datos utilizado en este estudio contiene todos los tweets que tienen el hashtag "#ausvote", de modo que los tweets están relacionados con las elecciones. Tras un primer intento de modelización temática, pudimos dividir los tweets en tres grupos. Hicimos más modelizaciones temáticas sobre los datos agrupados y encontramos patrones o comunidades ocultas subyacentes a los tuits. Esta técnica puede utilizarse para ayudar en la campaña de las elecciones federales. Por ejemplo, en la campaña de las elecciones federales de 2016, el LNP tiene mandatos que harán que aumenten los costes de la educación y la sanidad. En todos los tuits con "#ausvote", si se utiliza el modelado de temas, puede ayudar a categorizar cuántos tuits del LNP tratan sobre esos temas.

dos mandatos impopulares. Este análisis puede ayudar al LNP a cambiar su mandato o a anunciar la razón por la que lo está haciendo para educar a la gente y hacerles entender que es necesario para la economía del país y para ganar más votos para ganar las elecciones. En el conjunto de datos de este proyecto (Elecciones Federales Australianas de 2010), tras realizar un modelado de temas, es visible que un gran grupo de tuits sobre Julia Gillard tenía que ver con su elección a Gobernadora General. Esto demuestra que es posible descubrir grupos ocultos e interpretar su significado.

Esta técnica de descubrimiento de grupos ocultos también puede utilizarse para descubrir actividades delictivas o terroristas inminentes que estén planeando los delincuentes mediante el modelado temático de los datos de comunicación en Internet. Este uso de la modelización temática puede ser enormemente beneficioso para la sociedad, ya que evitará muertes, lesiones y destrucción de bienes. Además, este tipo de trabajo no requiere grandes inversiones, por lo que es una forma muy económica de luchar contra el terrorismo o los actos delictivos a gran escala.

Al descubrir grupos ocultos y categorizar datos de conjuntos de datos muy grandes, el modelo temático puede mejorar la velocidad y la precisión del resultado. Nuestro conjunto de datos no es muy grande. Las ventajas del modelo temático aumentan con el tamaño del conjunto de datos. Si analizáramos nuestro resultado, seguiríamos viendo los beneficios mencionados a pequeña escala. Nuestro objetivo es comparar a los dos candidatos a Primer Ministro Julia Gillard y Tony Abbott. Tras calcular el sentimiento y etiquetar la entidad de los tuits, realizamos un modelado temático y encontramos un grupo oculto que puede descartarse sin que afecte a nuestro resultado. Si se tratara de un enorme conjunto de datos con miles de entidades, sería realmente engorroso representarlas en un gráfico. Al detectar el grupo de entidades menos interesante, podemos eliminarlas del gráfico y producir un gráfico más pequeño y fácil de presentar sin que ello afecte al resultado. Si realizáramos el modelado temático antes del cálculo del sentimiento, podríamos realizar menos cálculos eliminando los grupos menos interesantes. Menos cálculo significa menos uso de procesamiento informático que podría conducir a un resultado más rápido y menos coste energético para ahorrar costes financieros directos, haciendo que el sistema sea más económico y eficiente.

5 CONCLUSIONES

Este estudio ha demostrado que la técnica de modelado temático puede utilizarse en el proceso de extracción de opiniones para mejorar las características del sistema. En contraste con la mayoría de los trabajos de investigación existentes en este campo de la minería de opinión, que tienden a pasar por alto la detección del grupo oculto subyacente de datos en el conjunto de datos, este trabajo se centró en la detección del patrón oculto en el conjunto de datos. El resultado experimental muestra que el modelo temático puede descubrir una comunidad oculta de tweets mediante la categorización de los tweets en grupos, y que el modelado temático posterior de un grupo de tweets reveló un patrón subyacente profundamente oculto en ese grupo de tweets. Esta técnica de modelado temático mejoró aún más el sistema TSAM y, sin duda, puede utilizarse para mejorar el rendimiento de cualquier sistema de minería de opiniones o análisis de sentimientos, ya sea un sistema no supervisado basado en léxicos, un sistema supervisado, un sistema supervisado basado en corpus o un sistema híbrido. Los resultados de este trabajo son la capacidad de categorizar los datos de forma significativa y de encontrar el grupo oculto subyacente de datos. Estos resultados nos permitirán tomar decisiones informadas y hacer predicciones sobre acontecimientos sociales emergentes.

REFERENCIAS

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow y Rebecca Passonneau. 2011. Sentiment analysis of twitter data. En *Proceedings of the workshop on languages in social media*. Asociación de Lingüística Computacional, 30-38.

[2] Amir Asiaee T, Mariano Tepper, Arindam Banerjee y Guillermo Sapiro. 2012. Si eres feliz y lo sabes... tuitea. En *Proceedings of the 21st ACM interna- tional conference on Information and knowledge management*. ACM, 1602-1606.

[3] Michael I. Jordan; David M. Blei, Andrew Y. Ng. 2003. Asignación latente de Dirichlet.
Journal of Machine Learning Research 3(1) (2003), 993-1022.

[4] Adam G Dunn, Julie Leask, Xujuan Zhou, Kenneth D Mandl y Enrico Coiera. 2015. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. *Revista de investigación médica en Internet* 17, 6 (2015).

[5] Shawn Graham, Scott Weingart e Ian Milligan. 2012. Getting started with Topic Modeling and MALLET. *The Programming Historian* 2 (2012), 12.

[6] Vasileios Hatzivassiloglou y Janyce M Wiebe. 2000. Effects of adjective orienta- tion and gradability on sentence subjectivity. En *Proceedings of the 18th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics. Association for Computational Linguistics, 299-305.

[7] Liangjie Hong y Brian D Davison. 2010. Estudio empírico del modelado de temas en Twitter. En *Proceedings of the flrst workshop on social media analytics*. ACM, 80-88.

[8] Xia Hu, Jiliang Tang, Huiji Gao y Huan Liu. 2013. Unsupervised sentiment analysis with emotional signals. En *Proceedings of the 22nd international conference on World Wide Web*. ACM, 607-618.

[9] Bernard J. Jansen, Mimi Zhang, Kate Sobel y Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci.* 60, 11 (2009), 2169-2188.

[10] J. Kamps, M. Marx, R. Mokken y M. de Rijke. 2004. Using WordNet to mea sure semantic orientation of adjectives. En *Proceedings of the 4th International Conference on Language Resources and Evaluation*. 1115-1118.

[11] Chetan Kaushik y Atul Mishra. 2014. A scalable, lexicon based technique for sentiment analysis. *arXiv preprint arXiv:1410.2265* (2014).

[12] Chenghua Lin y Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. En *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 375-384.

[13] Bing Liu. 2010. *Manual de procesamiento del lenguaje natural* (segunda edición). Capítulo Sentiment Analysis and Subjectivity.

[14] Ravi Parikh y Matin Movassate. 2009. Sentiment analysis of user-generated twitter updates using various classification techniques. *CS224N Final Report* (2009), 1-18.

[15] S. Shahheidari, H. Dong y M. N. R. B. Daud. 2013. Twitter Sentiment Mining: A Multi Domain Analysis. In *Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference on*. 144-149. DOI:http://dx.doi.org/10.1109/CISIS.2013.31

[16] Alessio Signorini, Alberto Maria Segre y Philip M Polgreen. 2011. El uso de Twitter para rastrear los niveles de actividad de la enfermedad y la preocupación pública en los EE.UU. durante la pandemia de gripe A H1N1. *PloS one* 6, 5 (2011), e19467.

[17] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll y Manfred Stede. 2011. Métodos basados en léxicos para el análisis de sentimientos. *Lingüística computacional* 37, 2 (2011), 267-307.

[18] JW Uys, ND Du Preez y EW Uys. 2008. Leveraging unstructured information using topic modelling. En *Management of Engineering & Technology, 2008. PICMET 2008. Conferencia Internacional de Portland*. IEEE, 955-961.

[19] Andrea Vanzo, Danilo Croce y Roberto Basili. 2014. A context-based model for Sentiment Analysis in Twitter.. En *COLING*. 2345-2354.

[20] Theresa Wilson, Janyce Wiebe y Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. En *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Asociación de Lingüística Computacional, 347-354.

[21] Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer y Pankaj Gupta. 2014. Modelado de temas de alta precisión a gran escala en Twitter. En *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1907-1916.

[22] Xujuan Zhou, Enrico W Coiera, Guy Tsafnat, Diana Arachi, Mei-Sing Ong, Adam G Dunn y otros. 2015. Using social connection information to improve opinion mining: La identificación de sentimiento negativo acerca de las vacunas contra el VPH en Twitter.. En *MedInfo*. 761-765.

[23] Xujuan Zhou, Xiaohui Tao, Jianming Yong y Zhenyu Yang. 2013. Análisis de sentimiento en tweets para eventos sociales. En *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*. IEEE, 557-562.