Iván Arias Rodríguez

El análisis de opinión como herramienta para la predicción de resultados electorales. Una contribución para el español.

Bajo la Dirección de:

Ana María Fernández-Pampillón Cesteros

Víctor Peinado Herencia

Grado en Lingüística y Lenguas Aplicadas, Universidad Complutense de Madrid,

17 de junio de 2015, convocatoria de junio.

Universidad Complutense de Madrid	**Resumen del Trabajo de Fin de Grado**

Autor: Iván Arias Rodríguez

Título: El análisis de opinión como herramienta para la predicción de resultados electorales. Una contribución para el español.

Fecha: 17 de junio de 2015

Número de páginas: 36

Departamento: Filología Románica, Filología Eslava y Lingüística General

Directores: Ana María Fernández-Pampillón Cesteros

Víctor Peinado Herencia

En los últimos años, las redes sociales, y en especial Twitter, se han convertido en un foro de debate global donde se discute sobre los temas de actualidad. Los partidos políticos y sus candidatos, así como los ciudadanos en general, utilizan Twitter como una herramienta de expresión política. Por otra parte, el denominado *análisis de opinión* es una tarea que está ganando cada vez más importancia dentro del ámbito del procesamiento del lenguaje natural, y que consiste en detectar información subjetiva del texto, como su positividad o su negatividad.

En el presente trabajo se estudia si se puede obtener algún tipo de correlación entre los temas tratados y las opiniones (favorables o desfavorables) extraídas de mensajes publicados en Twitter sobre unas elecciones (en nuestro caso las elecciones a rector de la Universidad Complutense de Madrid de mayo de 2015), y los resultados reales de las elecciones, con intención de implementar un sistema que pudiera hacer predicciones sobre porcentajes de voto obtenidos por cada candidato.

Palabras clave: análisis de opinión, polaridad, análisis de sentimiento, procesamiento del lenguaje natural, PLN, Python, Twitter, elecciones, rector.

Complutense University of Madrid	**Abstract of End-of-degree Project**

Author: Iván Arias Rodríguez

Title: Sentiment analysis as a tool for predicting election results.
A contribution for the Spanish language.

Date: June the 17th, 2015

Number of pages: 36

Department: Romance Philology, Slavic Philology and General Linguistics

Supervisors: Ana María Fernández-Pampillón Cesteros

Víctor Peinado Herencia

During the last years, social networks, and especially Twitter, have become a global debate forum where users discuss about current affair topics. Political parties and their candidates, as well as citizens in general, use Twitter as a tool to express freely their political views. Besides that, the so called *sentiment analysis* is a task that is gaining more and more importance in the scope of natural language processing, consisting in detecting subjective information within a text, such as its positivity or negativity.

In this work we will study whether it is possible to obtain any kind of correlation between the main topics of conversation or opinions (in favor or against) extracted from messages published in Twitter about some elections (in our case, the election of the dean of the Complutense University of Madrid, held on May the 12th, 2015), and the real results in the elections, with the intent to implement a system that could make a forecast about percentage of vote that the candidates will obtain

Keywords: sentiment analysis, polarity, natural language processing, NLP, Python, Twitter, elections, dean.

Índice

1. Introducción

En el presente trabajo se va a desarrollar una metodología para estudiar la relación entre el resultado de unas elecciones (en nuestro caso las elecciones a rector de la Universidad Complutense de Madrid -en adelante UCM- celebradas en mayo de 2015) y los mensajes publicados en la red social Twitter relacionados con el proceso electoral y sus candidatos. Para ello, se va a crear una herramienta que pueda detectar de forma automática (o al menos, semiautomática) las preocupaciones y opiniones (positivas o negativas) de una comunidad de electores y candidatos involucrados en un proceso electoral, utilizando como material de partida sus mensajes en las red social Twitter. Partiendo de los datos de opinión se hará una predicción de resultados y se comprobará si efectivamente existe una relación entre los resultados electorales y los mensajes de Twitter de partida.

Se ha decidido escoger Twitter por ser una red social con mensajes públicos donde los usuarios se expresan libremente sobre cualquier acontecimiento. Además, tiene el aliciente adicional de que los mensajes suelen contener códigos que agrupan mensajes relacionados con algún tema en concreto (llamados *hashtags*, una cadena de texto de la forma #*palabraclave*).

El llamado *análisis de opinión* es una tarea que está ganando cada vez más importancia dentro del ámbito del procesamiento del lenguaje natural. Se basa en la identificación y extracción de información subjetiva de textos en los que los usuarios evalúan distintos aspectos de algún producto o servicio, como puede ser una película, un restaurante o prácticamente cualquier cosa.

El análisis de lo que opinan los usuarios de internet sobre un cierto producto es una tarea de gran importancia para las empresas, puesto que una gran parte de los clientes de muchos servicios suelen utilizar dichas opiniones para decidirse en su compra. Por ello, desde principios de la década pasada se han empezado a hacer intentos por conseguir un analizador automático de opinión (polaridad) positivo, negativo o neutro de las opiniones expresadas en un cierto texto.

Los primeros trabajos en el área de análisis de opinión son obra de Turney (2002) y de Pang, Lee y Vaithyanathan (2002). Desde entonces se han ido aplicando diversas técnicas para analizar la polaridad de un texto y los sistemas actuales son capaces de extraer los temas más importantes y su polaridad con bastante precisión. El problema principal, es que el desarrollo de estas técnicas se ha hecho principalmente para la lengua inglesa, y su funcionamiento está aún bastante limitado en otros idiomas. Para el inglés se alcanzan precisiones superiores al 70%, lo cual es un valor bastante alto teniendo en cuenta que las valoraciones hechas por humanos normalmente alcanzan precisiones de en torno al 79% (Ogneva, 2010), con lo que bastaría alcanzar esta precisión para tener un sistema que se comportara como un humano.

Twitter es hoy en día una herramienta muy potente, utilizada por millones de usuarios[1]. Esta red social, creada en 2006[2], se fue popularizando y se ha convertido en un foro mundial donde sus usuarios opinan sobre algún tema de actualidad. Personalidades públicas, y cada vez más frecuentemente políticos e incluso partidos políticos en su conjunto, lo utilizan para contestar en tiempo real y de forma sencilla a sucesos de la actualidad que antes requerían la presencia de los medios para hacerse pública. Los partidos políticos hacen campaña en Twitter y exponen sus propuestas, y su uso es especialmente importante para candidatos sin el respaldo de grandes partidos y que tienen que dar a conocer su programa y opiniones de forma rápida y económica.

Siempre que hay alguna elección de relevancia hay un gran interés por conocer los resultados de los sondeos antes de que dichas elecciones se produzcan. Organizaciones como Metroscopia o el CIS, entre muchas, invierten una cantidad importante de recursos para preguntar a los electores sobre su intención de voto, en un proceso que requiere un importante gasto económico y de tiempo, con lo que normalmente se realizan cada varias semanas. Así que sería muy deseable contar con algún sistema que pudiera automatizar este proceso de consulta y que extrajera proyecciones de voto, que aunque tuvieran menor precisión que las encuestas tradicionales (que no siempre son precisas), sí que les sirvieran de apoyo debido a que virtualmente tendría un coste nulo y se podrían extraer los datos de forma casi instantánea en minutos u horas.

Así pues, en el presente trabajo se ha investigado si el análisis de lo que se comenta en Twitter sobre unas elecciones puede llevarnos a proyecciones de voto válidas o no. Para ello, se ha desarrollado una metodología, válida para analizar los mensajes cortos en lengua española de las redes sociales, con el fin de extraer los temas tratados y la opinión general sobre dichos temas. Esta metodología se ha aplicado a los mensajes de Twitter relacionados con las elecciones a rector de la UCM de mayo de 2015.

En la sección segunda de este trabajo se expondrán trabajos previos realizados en el mismo ámbito. Posteriormente, en la sección tercera se explicará la metodología desarrollada. En la sección cuarta se utilizarán los datos proporcionados por el sistema creado para hacer proyecciones de voto y se analizarán los resultados y su relación con los resultados reales. Tras ello, se presentará en la sección quinta las limitaciones de nuestra metodología y se propondrán posibles mejoras.

[1] Twitter cuenta con más de 300 millones de usuarios activos mensuales, más de 230 millones de ellos de fuera de los Estados Unidos como se indica en la propia red social: https://about.twitter.com/company.
[2] Twitter se creó en marzo de 2006 y finalmente se lanzó en julio del mismo año: https://es.wikipedia.org/wiki/Twitter

2. Antecedentes

El análisis de opinión es una tarea que lleva realizándose más de una década. A lo largo de este tiempo se han aplicado distintas técnicas (principalmente estadísticas) y se le ha intentado dar distintos usos, casi siempre en el ámbito de las críticas de películas, restaurantes y otro tipo de actividades comerciales que son frecuentemente evaluadas por los usuarios. Si bien en los últimos cinco años se han popularizado estudios que se salen de este ámbito.

Hay abundantes ejemplos de trabajos, como el de Asur y Huberman (2010), en los que se utiliza un procedimiento relativamente sencillo para calcular la polaridad de un texto. Se basa en la búsqueda de una serie de palabras que tienen un cierto peso (positivo, negativo o neutro) asignado previamente de forma manual, tras lo que se calcula la suma total de los pesos individuales. En el artículo se expone un método para utilizar un corpus de mensajes de Twitter (en adelante *tweets*) para predecir el éxito o fracaso en taquilla de los próximos estrenos cinematográficos basándose en la polaridad de los mensajes publicados sobre la película.

Existen intentos previos de utilizar los mensajes de Twitter (en adelante, *tweets*) para predecir el estado de ánimo general y su efecto en diversos ámbitos. En un trabajo de Bollen, Mao y Zeng (2011), se recopilaron tweets de usuarios estadounidenses durante un periodo de dos meses que expresaran estado de ánimo (tweets que incluyeran cadenas como *"I feel"*, *"I am feeling"* o *"makes me"*, entre otras). Posteriormente analizaron cada tweet de forma individual utilizando las herramientas OpinionFinder[3], un etiquetador de polaridad que funciona con el algoritmo estándar descrito anteriormente, además del uso de la estadística mediante un clasificador. Este sistema, que funciona únicamente en inglés, ha demostrado tener en torno a un 78%[4] de precisión. En este trabajo, también se analizaron los tweets usando su propia herramienta, GPOMS (*Google Profile of Mood States*), que calcula la opinión en función de seis dimensiones: *calmado*, *alerta*, *seguro*, *vital*, *amable* y *feliz*. Como OpinionFinder, este sistema funciona por el sistema de la suma de pesos. Así, se encontró una correlación entre la dimensión *calmado* y las variaciones del índice Dow Jones entre tres y cuatro días después.

En un trabajo similar (O'Connor, Balasubramanyan, Routledge y Smith, 2010), se encontró una importante correlación entre la opinión positiva o negativa de un corpus de tweets y el índice Gallup de confianza en la economía (que se confecciona a partir de encuestas), así como con el índice de aprobación del gobierno (Obama, 2009). Pero no se consiguió encontrar una relación

[3] Disponible en http://mpqa.cs.pitt.edu/opinionfinder/

[4] Según se indica en la propia documentación de la herramienta, disponible para su descarga en http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2/opinionfinder_2_0/opinionfinder_2_0_README.txt

estrecha entre la polaridad de los tweets y los resultados de las elecciones presidenciales estadounidenses de 2008.

Centrándonos más en la tarea de utilizar tweets para intentar predecir el resultado de unas elecciones, es destacable el trabajo de Choy, Cheong, Laik y Shung (2011). En este caso se recopilaron más de 16.000 tweets relacionados con las elecciones presidenciales de Singapur de 2011 y se hizo una estimación de voto para cada uno de los cuatro candidatos. Se tuvieron en cuenta los tweets procesados según su polaridad (de nuevo con un algoritmo similar de suma de pesos positivos y negativos de palabras) y el censo, además de proyecciones sobre población por edad, alfabetización y porcentaje de usuarios de Twitter. El sistema fue capaz de predecir una victoria por un margen estrecho entre los dos principales candidatos, si bien erró a la hora de decantarse por el ganador. Los otros dos candidatos obtuvieron el puesto que se esperaba según los datos de Twitter aunque el último de ellos recibió muchos menos votos de los calculados.

En cualquier caso, son cada vez más populares los estudios centrados en el análisis de opinión que investigan métodos para predecir el resultado de unas elecciones usando un corpus formado por tweets. Y los resultados son diversos: en ocasiones se hace notar una aparente relación entre las predicciones hechas mediante el análisis del texto, y en otras el resultado no parece estar relacionado y la predicción es incorrecta. Hay incluso artículos que recopilan los resultados de este tipo de trabajos, como el de Gayo-Avello (2012), en el que se interpreta que la mayor parte de las correlaciones que encontraron dichos artículos entre algún aspecto del corpus de tweets y los resultados de las elecciones eran debidos puramente a la casualidad, al sesgo del corpus o a la manera de medir dichos aspectos para escoger la media más ventajosa. También da en su artículo una serie de recomendaciones a la hora de hacer este tipo de estudios, que se han tenido en cuenta en el presente trabajo.

Aunque los artículos relacionados con el objetivo de este trabajo son numerosos, la práctica totalidad de ellos analizan textos en inglés. Uno de los principales problemas existentes a la hora de abordar un trabajo similar en la lengua española es que aún no hay disponible una cantidad similar de corpus en español etiquetados con sus polaridades que puedan utilizarse con los algoritmos de análisis de opinión habituales.

Sin embargo, la necesidad por abordar esta misma tarea en otras lenguas es creciente. En un trabajo de Brooke, Tofiloski y Taboada (2009), se estudió la forma de reutilizar las herramientas en inglés ya existentes para poder hacer el análisis de opinión para textos en español. La base del trabajo era comprobar cuál de las tres siguientes soluciones al problema daba mejor resultado:

- Crear un analizador de polaridad similar a los existentes en inglés partiendo de la nada.
- Hacer una traducción automática del texto español al inglés para después utilizar las herramientas inglesas para analizar la polaridad.
- Utilizar un corpus etiquetado en español y un clasificador basado en aprendizaje automático con el método de la bolsa de palabras.

Si bien el corpus utilizado no era demasiado grande -lo que hizo que la tercera solución fuera la menos precisa-, el hecho de traducir el texto original del español al inglés también aumentaba la tasa de fallos, aunque únicamente en un 3-5%. El mejor resultado lo obtuvo la herramienta diseñada específicamente para el español, que tenía un funcionamiento análogo al de los analizadores en inglés. Este analizador no era excesivamente complejo, y por lo tanto queda aún espacio para la mejora, pero fue relativamente rápido de crear. Incluía análisis específicos tanto para las negaciones, como para la intensidad de los adjetivos usados y la identificación de expresiones hipotéticas.

Así pues, el análisis de opinión, y en concreto su uso conjunto con corpus extraídos de Twitter, es una cuestión que no está totalmente resuelta y que despierta el interés de los sectores científico, social y económico. El principal problema reside en la escasez de herramientas creadas expresamente para otros idiomas distintos al inglés y, como se ha visto (Brooke y otros, 2009), los sistemas que mejor resultado dan son los creados específicamente para el español. Debido a ello, en este trabajo se desarrollará una herramienta específica para el español basándose en métodos ya utilizados previamente, desarrollando además un método nuevo que se basa en un análisis sintáctico del texto de los tweets.

3. Metodología de trabajo

Para llevar a cabo el objetivo de este trabajo, que es el desarrollo de una herramienta que sea capaz de extraer proyecciones de voto válidas a través del análisis de los tweets que se publican en relación a los candidatos que concurren a unas elecciones, se ha seguido una metodología de desarrollo incremental en siete fases que se describirán en las subsecciones siguientes:

1. Construcción del corpus de tweets.
2. Creación de un lexicón etiquetado por polaridad.
3. Creación de un flexionador, un silabeador y un lematizador.
4. Creación de un segmentador y etiquetador sintáctico.
5. Creación de un analizador sintáctico.
6. Creación de un extractor de temas y palabras relacionadas.
7. Creación de un extractor de polaridad.

3.1. Construcción del corpus de tweets

El corpus se creó en tres partes. A continuación se mostrará en qué consistió cada una.

3.1.1. Creación de una tabla de términos asociados a los candidatos y a las elecciones

Tras hacer un análisis previo de los mensajes publicados en Twitter en relación a las elecciones a rector, se extrajeron manualmente una serie de palabras, usuarios de Twitter[5] y hashtags que resultaban relevantes para nuestra búsqueda de tweets relacionados con las elecciones. Todos los candidatos utilizaban al menos un usuario de Twitter para hablar sobre sus propuestas.

Para cada candidato se hizo una ficha que incluye su nombre, su usuario de Twitter y los hashtags relacionados, como los eslóganes utilizados en la campaña electoral. Se añadieron también otros hashtags relacionados que estaban incluidos en tweets que hablaban sobre el candidato (habitualmente hashtags de la forma *#nombrecandidato*). La Tabla 1 muestra estas cadenas de texto que posteriormente se buscarán en Twitter. Los candidatos aparecen de más a menos votados en la primera vuelta, una convención que se utilizará en el resto de tablas.

Candidato	Atributo	Valores
Carlos Andradas	"usuario"	"@carlosandradas"
	"nombre"	"carlos andradas"
	"eslogan"	"#mascomplutense", "#andradasrector"
	"relacionados"	"#carlosandradas"
José Carrillo	"usuario"	"@josecarrilloucm", "@pepcarrillo"
	"nombre"	"jose carrillo", "pep carrillo"
	"eslogan"	"#contigoucm"
	"relacionados"	"#josecarrillo"
Federico Morán	"usuario"	"@fmoranab"
	"nombre"	"federico moran"
	"eslogan"	"#eselmomento", "#votamoran"
	"relacionados"	"#federicomoran"
Rafael Calduch	"usuario"	"@rafaelcalduch"
	"nombre"	"rafael calduch"
	"eslogan"	"#sercomplutense"
	"relacionados"	"#rafaelcalduch"
Dámaso López	"usuario"	"@damasolopezucm"
	"nombre"	"damaso lopez"
	"eslogan"	"#lopezrector"
	"relacionados"	"#damasolopez"

Tabla 1: Fichas de textos en tweets asociados a los candidatos

[5] Un usuario de Twitter es una cadena de caracteres que identifica al emisor del tweet. Son de la forma *@nombreusuario*. En adelante nos referiremos a ellos simplemente como *usuarios*.

Además, se observó que había una serie de usuarios que eran foros de discusión sobre las elecciones. También se encontraron hashtags genéricos que se asociaban a discusión relacionada con estas elecciones. De la misma manera que se hizo la ficha para los cinco candidatos, se hicieron otras tres más, que se muestran en la Tabla 2.

Cuenta de Twitter	Atributo	Valores
La Complutense Decide	"usuario"	"@lacompludecide"
	"relacionados"	Ø
Universidad Complutense	"usuario"	"@unicomplutense"
	"relacionados"	Ø
Usuarios anónimos	"usuario"	Ø
	"relacionados"	"#eleccionesucm", "#debateucm", "#rectorucm", "#caraacaraucm"

Tabla 2: Fichas de textos en tweets asociados a las elecciones

Para desarrollar la herramienta se ha utilizado el lenguaje de programación *Python*. Utilizando la librería *tweepy* (Version 3.2.0; 2015) de Python[6] se descargaron los *timelines*[7] de los usuarios de la Tabla 1 y la Tabla 2. Por otro lado, se hizo una búsqueda de tweets que incluyeran en su texto bien una *mención*[8] a alguno de los usuarios, bien su nombre, eslogan o hashtags relacionados.

En este primer paso se descargaron 105.445 tweets, antes de proceder a su purga.

3.1.2. Purga de tweets duplicados

Los tweets que contienen más de una de las cadenas de texto de la Tabla 1 y/o la Tabla 2 se descargan más de una vez, con lo que se eliminan del corpus los tweets duplicados. Aunque podría ser relevante para el análisis, no se tiene en cuenta en el análisis el hecho de que un tweet haya podido ser reenviado o que haya sido marcado como favorito por otros usuarios.

Tras esta primera purga de duplicados quedaron 61.424 tweets y se pasó a identificar tweets distintos pero con el mismo contenido. En primer lugar, se buscaron *retweets*[9] y se reemplazaron por el mensaje original, manteniendo únicamente una copia. Tras este paso quedaron 42.315 tweets.

Cuando un tweet incluye alguna URL externa, el texto del tweet contiene una versión acortada de esa dirección creada dinámicamente, que es de la forma *http://t.co/ALFANÚMERO* y ocupa 22 caracteres. Esta sustitución plantea el problema de que puede haber mensajes realmente idénticos,

[6] Python es un software estándar, disponible en https://www.python.org.
[7] El timeline es el conjunto de tweets enviados por un usuario de Twitter en concreto.
[8] Una mención a un usuario en Twitter es un tweet que contiene en su texto el código de ese usuario.
[9] Un retweet consiste en un reenvío de un tweet escrito originalmente por otro usuario.

pero que únicamente se diferencien por la URL. Así que se identificaron y eliminaron dichas URL acortadas y se purgaron los tweets que contuvieran el mismo texto.

Finalmente se purgaron los tweets previos al ocho de abril de 2015 (fecha de proclamación definitiva de candidatos) y posteriores al doce de mayo de 2015 (víspera de la segunda votación).

3.1.3. División del corpus en secciones

Se dividió el corpus en once secciones para procesarlas por separado:

- Un total de cinco secciones *de candidato*, que incluyen los tweets enviados por cada candidato (es decir, su timeline).
- Otras cinco *sobre candidato*, que incluyen los tweets que contienen alguna de las cadenas de texto asociadas con cada candidato en particular.
- Otra *sobre elecciones*, que incluye los demás tweets.

La Tabla 3 muestra el contenido de cada una de ellas:

Candidato	De			Sobre		
	Tweets	Palabras	Caracteres	Tweets	Palabras	Caracteres
Carlos Andradas	825	15.171	45.513	2.383	40.646	121.938
José Carrillo	1.023	17.251	51.753	1.899	33.304	99.912
Federico Morán	542	10.491	31.473	1.527	26.282	326.436
Rafael Calduch	277	4.691	14.073	329	5.726	17.178
Dámaso López	190	3.048	9.144	211	3.690	11.070
Subtotal	**2.857**	**50.652**	**151.956**	**6.349**	**109.648**	**576.534**
Sobre elecciones	3.241	55.867	167.601	-	-	-
Total	**6.098**	**106.519**	**319.557**	**6.349**	**109.648**	**576.534**

Tabla 3: Tamaño de las secciones del corpus de y sobre candidatos y de la sección sobre elecciones

Hay un total de 12.447 tweets, si bien algunos hablan sobre más de un candidato, con lo que aparecen en varias de las secciones sobre candidatos. Por ello, se ha hecho una duodécima sección, la sección *de comentarios*, que incluye todos los tweets únicos enviados por usuarios distintos de los candidatos (es decir, la unión de la sección sobre elecciones y las cinco secciones sobre candidato, purgando duplicados). Se muestra en la Tabla 4 la cantidad final total de tweets, según su emisor:

Emisor	Tweets	Palabras	Caracteres
Carlos Andradas	825	15.171	45.513
José Carrillo	1.023	17.251	51.753
Federico Morán	542	10.491	31.473
Rafael Calduch	277	4.691	14.073
Dámaso López	190	3.048	9.144
Comentarios	7.322	126.613	804.010
Total	**10.179**	**177.265**	**955.966**

Tabla 4: Tamaño de las secciones del corpus según el emisor del tweet

3.2. Creación del lexicón etiquetado por polaridad

La mayoría de los análisis de opinión se basan bien en un corpus de textos de opinión previamente etiquetados según su polaridad (para posteriormente utilizar un clasificador basado en aprendizaje automático), bien en un lexicón etiquetado con información acerca de la polaridad de la palabra además de otros aspectos como su subjetividad o su intensidad. Sin embargo, los corpus etiquetados según polaridad en español son escasos y, hasta donde se ha podido encontrar, no tratan el ámbito de la política.

Para solventar este problema se ha utilizado la ayuda de SentiWordNet, obra de Esuli, A., y Sebastiani, F. (2006), una herramienta de análisis de polaridad, subjetividad e intensidad, que incluye lexicones etiquetados públicos. Concretamente se utilizó un lexicón etiquetado que contiene 2.918 adjetivos, nombres, verbos y adverbios en sus formas inglesa y neerlandesa, obra de Smedt, T., y Daelemans, W. (2012)[10]. Este lexicón se creó a partir de una lista de 1.100 adjetivos etiquetados manualmente, para cada uno de los sentidos de dichas palabras en WordNet. Posteriormente los autores aumentaron el lexicón utilizando las relaciones de sinonimia y antonimia que ofrece WordNet, y es por ello que 2.784 de las palabras de este lexicón incluyen su identificador único de WordNet.

Además de este lexicón, se ha utilizado un archivo con un listado de los lemas de WordNet en español junto con su identificador único[11], mediante el cual se tradujeron los lemas del lexicón etiquetado por polaridad del inglés al español. Al ser la versión española más limitada que la versión inglesa original, sólo se pudieron traducir 1.712 de dichos términos, si bien, como las traducciones se hacen según la palabra y su sentido, se considera que las traducciones de los términos ingleses son las mejores posibles. En adelante nos referiremos a este lexicón como el *Diccionario de Polaridad*.

3.3. Creación del flexionador, del silabeador y del lematizador

En este momento, el Diccionario de Polaridad solamente incluía lemas, algo que es insuficiente en español. Para ampliar este diccionario con las formas flexionadas se creó un flexionador de palabras en español que devuelve tanto la forma como su etiqueta sintáctica completa según el estándar del grupo EAGLES para la anotación morfosintáctica de lexicones y corpus[12].

El flexionador toma como entrada la forma del lema y su categoría sintáctica (extraída de WordNet) y genera una serie de términos flexionados:

---

[10] Disponible en https://github.com/clips/pattern/blob/master/pattern/text/en/en-sentiment.xml
[11] Disponible en http://compling.hss.ntu.edu.sg/omw/
[12] Disponible en http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

- En caso de tratarse de un nombre, se siguen las normas de la RAE para la formación del plural[13].
- Si se trata de un adjetivo, además se crean las formas femeninas.
- Para el caso de los verbos, se conjuga el verbo siguiendo el patrón regular de la conjugación a la que pertenezca (según la terminación del lema). Tiene en cuenta los clíticos que pueden añadirse a la forma verbal en infinitivo, gerundio e imperativo, así como las formas femeninas y plurales del participio. Por cada verbo genera 310 formas distintas (62 formas simples, 3 formas extra del participio, y 245 formas con clítico).

Como apoyo para el flexionador se ha creado un silabeador (implementado como una gramática independiente del contexto -GIC- en Python), que divide el lema de entrada en sílabas e identifica la sílaba tónica, para así cumplir las normas ortográficas del español sobre acentuación de palabras cuando se usa el flexionador.

Se amplió cada lema del Diccionario de Polaridad con todas sus formas flexionadas, terminando por contener 18.326 formas flexionadas.

Con el objetivo de hacer un recuento de palabras más frecuentes para hallar los temas principales, se necesita un lematizador que aglutine todas las formas de una palabra en su forma canónica, para contarlas conjuntamente. Así que se utilizó el flexionador para crear un lematizador, que toma la palabra de entrada y su etiqueta sintáctica, y busca posibles lemas (basándose en heurísticas y las normas de acentuación de la RAE) que, al flexionarse, generen la forma de entrada junto con su etiqueta sintáctica. El lematizador devuelve entonces una lista de posibles lemas (en ocasiones, hasta tres), ordenados de formas más cortas a más largas. En caso de igual longitud se escoge primero la forma masculina para nombres y adjetivos, y para el caso de verbos se devuelven antes posibles lemas de la primera conjugación (más comunes), seguidos de los de la segunda y la tercera. Se usa por defecto la primera opción posible y el resto de formas se descarta.

3.4. Creación del segmentador y etiquetador sintáctico

Se ha creado un segmentador que se basa en el algoritmo de Grefenstette, G., y Tapanainen, P. (1994), que divide los tweets en frases y dichas frases en palabras. Se tienen en cuenta signos de puntuación ambiguos como el punto, la coma o los dos puntos, que pueden formar parte de acrónimos, abreviaturas, números, fechas u horas, o pueden ser separadores de palabras.

[13] Disponibles en http://lema.rae.es/dpd/srv/search?id=Iwao8PGQ8D6QkHPn4i

Una vez segmentado el tweet, se deben etiquetar las palabras según por categoría sintáctica. Para ello se crea un etiquetador en cascada que aplica cinco etiquetadores consecutivos: etiquetador de trigramas, bigramas, unigramas, basado en expresiones regulares y por defecto (se utiliza la etiqueta para nombre común). Los etiquetadores de n-gramas se entrenan con el corpus español de CoNLL-2002[14], etiquetado siguiendo el estándar de EAGLES en su forma corta, y que está disponible en el paquete *nltk* (Bird, Klein y Loper, 2009) de Python. Para el etiquetador basado en patrones se crearon 134 expresiones regulares que cubren las 77 etiquetas de EAGLES, y además se crearon dos etiquetas nuevas para usuarios y hashtags (que en la práctica, se tratarán como nombres propios). El etiquetador en cascada tiene una precisión del 94,90% sobre el corpus de evaluación que incluye el propio paquete de CoNLL-2002.

3.5. Creación del analizador sintáctico

Para el tipo de análisis que se quiere hacer, se necesita un analizador sintáctico de frases. Por ello se programó otra GIC que tiene como terminales las 77 etiquetas del vocabulario de EAGLES y 204 producciones. Para hacer un análisis sintáctico, se etiqueta el texto y se introduce la lista de etiquetas a la GIC, la cual devuelve todos los árboles sintácticos que es capaz de encontrar. Este analizador se testeó con 100 tweets escogidos aleatoriamente y analizados manualmente, haciendo un análisis correcto en 56 de ellos, aunque más de la mitad tenían al menos dos interpretaciones debido a la ambigüedad de la frase (siendo habitualmente la primera interpretación la más lógica). En 44 de ellas no encontró un análisis sintáctico válido, debido a que los tweets suelen ser textos sin mucha estructura, llenos de hashtags y usuarios, y donde se relajan las normas de puntuación y de acentuación, lo cual también hace que el etiquetador sintáctico falle y esto impida que el analizador encuentre un árbol sintáctico válido.

3.6. Creación del extractor de temas y palabras relacionadas

El extractor de temas se aplica a cada una de las cinco secciones de candidato del corpus y también a la sección de comentarios. Previamente, dichos tweets se segmentan en frases y palabras, que se etiquetan con el etiquetador sintáctico y se lematizan automáticamente. El extractor de temas funciona en dos fases, que se describen a continuación.

---

[14] *The Conference on Computational Natural Language Learning* (CoNLL) es la conferencia en relación a una evaluación competitiva que se celebra periódicamente. La colección de datos incluida en el paquete CoNLL-2002, utilizada para evaluar sistemas de reconocimiento de entidades, se generó para la evaluación de 2002 titulada *Language-Independent Named Entity*. Más información sobre el evento en http://www.cnts.ua.ac.be/conll2002/ner/.

3.6.1. Identificación de los temas principales de conversación

De entre las palabras del corpus se identifican, vía su etiqueta sintáctica, los nombres (comunes y propios) además de los usuarios y los hastags, que tienen un número de apariciones de al menos un 0.1% del número total de palabras de la sección del corpus que se esté tratando. Estos lemas compondrán la lista de temas principales (según el corpus, habitualmente entre 20 y 30 términos).

3.6.2. Identificación de palabras relacionadas

Posteriormente se identifican las frases de la sección del corpus de entrada en las que aparece alguno de los temas principales. Para cada una de ellas se identifican automáticamente los *modificadores*: adjetivos, verbos y adverbios (ya lematizados) que también aparecen en ella.

Tras ello, se usa el analizador sintáctico para identificar el nombre común, nombre propio, usuario o hashtag de la frase con el que el modificador tiene una relación sintáctica más cercana, según el criterio que se explica más adelante. Si esa palabra resulta ser una de las palabras identificadas como tema principal, se asocia a dicho tema principal el modificador, llevando un recuento de cuántas veces dicha palabra aparece relacionada con el tema principal.

Para calcular el grado de relación sintáctica de dos palabras en una frase, se parte del árbol sintáctico que nos genera el analizador. En él, se halla el subárbol mínimo que incluya ambas palabras (que son hojas del subárbol). Si las dos palabras están al mismo nivel sintáctico (ambas palabras tienen un nodo padre común en el subárbol), se considera que las palabras tienen el grado de relación más estrecho, que tiene un valor de 1. En cualquier otro caso, el grado de relación sintáctica será la altura de ese subárbol (mayor que 1) desde la raíz hasta una de las dos palabras cuya relación estemos calculando (bien el nombre, bien el modificador), la que esté más alejada de la raíz.

Si el analizador sintáctico hace más de un posible análisis, se sigue el proceso descrito en el párrafo anterior para identificar el nombre, usuario o hashtag con mayor relación sintáctica con el modificador para cada uno de los árboles sintácticos. Posteriormente, se calculan las medias de los grados de relación sintáctica para cada pareja de palabras en cada uno de los análisis, y se utiliza ese valor medio como grado de cercanía.

Como se ha visto, el analizador sintáctico es incapaz de encontrar interpretaciones sintácticas de la frase en casi la mitad de los casos. También puede ocurrir que la frase sea demasiada larga, en cuyo caso se opta por no hacer el análisis sintáctico[15]. En estos casos se opta por un plan alternativo para

[15] El límite máximo es de 20 palabras, porque al analizar frases más largas –de las que apenas hay una decena en el corpus- el proceso de análisis puede ser extremadamente largo y suele dar una cantidad desproporcionada de posibles interpretaciones.

calcular la palabra más relacionada. Si el modificador está en una posición X en la frase, se buscan las palabras anteriores a la posición X (hasta tres) y si es un nombre o término de Twitter, se asocia a él. Si no se encuentra ninguno, se buscan las palabras siguientes a la posición X (hasta dos) y se procede de igual manera. Si aun así no se consigue encontrar ninguna palabra candidata, no se relaciona el modificador con ninguna palabra. Esta es una solución de compromiso que se ha visto que asocia bien unas palabras con otras en la mayor parte de los casos.

Así pues, al final del análisis se tiene una lista de temas principales (lematizados), ordenados por número de apariciones y para cada uno de ellos se tiene una lista de modificadores (lematizados) que aparecen relacionados sintácticamente con ellos en el corpus, según número de apariciones.

3.7. Creación del extractor de polaridad

Para crear el extractor de polaridad se utilizó la técnica más utilizada en la bibliografía consultada (que denominaremos algoritmo *trivial*), que consiste en utilizar el Diccionario de Polaridad para identificar la polaridad de las palabras individuales, y asignar a la frase la suma de polaridades de sus palabras.

Se hacen tres cálculos distintos de la polaridad, siendo el segundo una variante del primero:

- Para cada una de las cinco secciones de corpus de tweets que tratan sobre un candidato concreto, se utiliza el algoritmo *trivial* para calcular el número tweets positivos, negativos y neutros, así como la media de polaridad por tweet. Los resultados aparecen en la Tabla 16.
- Posteriormente se usa un clasificador de máxima entropía (al que llamaremos algoritmo *reclasificador*). Inspirándonos en el trabajo de Blair-Goldensohn, Hannan, McDonald, Neylon, Reis y Reynar (2008), se utilizan los tweets ya identificados como positivos o negativos en el paso anterior para entrenar un clasificador usando el método de la bolsa de palabras, incluyendo únicamente adjetivos, verbos, adverbios y nombres que no sean referencias al candidato. Así, se reclasifican los tweets neutros (con al menos un elemento en su bolsa de palabras) y se obtienen los resultados de la Tabla 17

  Para medir la precisión de este sistema, se entrenó inicialmente al clasificador con el 90% de los tweets positivos y negativos y se evaluó con el 10% restante. Tras calcular la exactitud (que aparece también en la Tabla 17), se termina de entrenar el clasificador añadiendo ese 10% al corpus de entrenamiento.
- Por último, se utiliza la sección de comentarios del corpus. Se extraen de él los temas según se explicó en la sección 3.6, eliminando todos los temas que no sean referencias a los candidatos (sus nombres, usuarios o hashtags asociados) y se agrupan por candidato. Finalmente, como el extractor de temas devuelve una lista de modificadores relacionados

con los temas, se suman las polaridades de todos esos modificadores relacionados con cada uno de los candidatos. A este algoritmo lo denominamos *temático*, y el resultado aparece en la Tabla 18. De la Tabla 19 a la Tabla 23 se muestran las palabras asociadas a cada candidato y su polaridad.

Las tres tablas de resultados incluyen además una proyección de votos que toma los valores de polaridad media (por tweet o por palabra relacionada, según el caso) o de la diferencia entre el porcentaje de tweets positivos y negativos. Estos datos se discutirán en la próxima sección.

4. Discusión de resultados

En esta sección, primero se evalúan algunas partes del sistema creado y posteriormente se comparan las proyecciones de votos con los datos reales de las elecciones y se analizarán las relaciones existentes entre ambos datos.

4.1. Evaluación de nuestra herramienta

La Tabla 17 muestra datos sobre la precisión del clasificador del algoritmo reclasificador. El clasificador tiene una precisión superior al 80%, al menos para los tres candidatos cuya sección de corpus es mayor. Para los casos de Rafael Calduch y Dámaso López, la sección del corpus de partida es demasiado pequeña, lo que produce altas tasas de fallos. Ponderando la exactitud del clasificador por el número de tweets que tiene que clasificar en cada sección, se tendría una exactitud media teórica para las cinco secciones del corpus del 80,82%.

Por otra parte, se hizo una evaluación del algoritmo de extracción de opinión utilizando el corpus español de reseñas etiquetadas de Julian Brooke[16]. Este corpus consta de ocho categorías, y en cada una de ellas hay 50 reseñas etiquetadas según su positividad (puntuaciones 3 y 4) o negatividad (puntuaciones 1 o 2). Para evaluar el extractor de polaridad (el algoritmo trivial y el reclasificador) se utilizan las reseñas más positivas y más negativas de cada categoría, y los datos de precisión aparecen en la Tabla 24. Los resultados son modestos, de en torno al 60% de precisión, debido fundamentalmente a que se evalúa sobre un corpus formado por reseñas, habitualmente de varios párrafos, donde se mezclan aspectos positivos y negativos. En el corpus creado en este trabajo, formado por tweets que rara vez sobrepasan las 20 palabras, los resultados son distintos. Se escogieron aleatoriamente 100 tweets y se evaluaron manualmente como positivos, negativos o neutros, y posteriormente se evaluaron con el extractor de polaridad (algoritmo trivial). La precisión fue del 82% tal y como puede verse en la Tabla 25.

[16] Disponible para su descarga en la dirección http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html

4.2. Comparación de proyecciones de voto y resultados reales

En la Tabla 5 se muestran los resultados de la primera ronda de las elecciones a rector, celebrada el 5 de mayo de 2015, desglosada según el voto de los distintos grupos que componían el censo electoral. La Tabla 6 es similar a la anterior y muestra los resultados de la segunda ronda, en la que sólo participaron los candidatos Carlos Andradas y José Carrillo.

Candidato	Número de votos					% de voto			
	PD	PDI-DC	PDI-DP	EST	PAS	PD	EST	PAS	Total
Carlos Andradas	818	261	151	3197	728	37,78%	36,13%	26,10%	34,61%
José Carrillo	502	256	155	1819	1340	28,04%	20,56%	48,05%	27,34%
Federico Morán	547	114	131	1737	466	24,32%	19,63%	16,71%	20,11%
Rafael Calduch	113	39	19	1380	123	5,25%	15,59%	4,41%	11,24%
Dámaso López	91	27	32	716	132	4,61%	8,09%	4,73%	6,70%
Votos en blanco	79	19	15	331	233				
Votos nulos	14	6	0	195	117				
Votos emitidos	2164	722	503	9375	3139				
Censo	2633	1077	2151	77187	3810				

Tabla 5: Resultados de la primera vuelta de las elecciones (5 de mayo de 2015)

Candidato	Grupo de votantes					% de voto			
	PD	PDI-DC	PDI-DP	EST	PAS	PD	EST	PAS	Total
Carlos Andradas	1154	335	237	5147	1099	59,89%	68,03%	40,42%	60,55%
José Carrillo	694	290	172	2419	1620	40,11%	31,97%	59,58%	39,45%
Votos en blanco	210	33	19	447	268				
Votos nulos	17	4	3	266	128				
Votos emitidos	2075	662	431	8279	3115				
Censo	2633	1077	2151	77187	3810				

Tabla 6: Resultados de la segunda vuelta de las elecciones (13 de mayo de 2015)

Observando ambas tablas se puede considerar tres grupos de votantes: estudiantes, personal docente y personal administrativo. El orden de candidatos más votados en los dos primeros grupos fue el mismo que el de los resultados finales. Sin embargo, el personal administrativo intercambió las posiciones de los dos primeros (Andradas y Carrillo) en las dos vueltas y las de los dos últimos (Calduch y López) en la primera vuelta.

Teniendo en cuenta esta diferencia, el primer análisis que se ha hecho ha sido identificar los temas principales de cada candidato con los tres grupos de votantes. De la Tabla 26 a la Tabla 30 aparecen los listados de estos temas principales. Se han eliminado temas irrelevantes, o todos aquellos relacionados con la propia candidatura, y se han añadido temas con un porcentaje de aparición menor al 0,1% que resultaban relevantes. Se agrupan estos temas por categorías:

- Relacionado con los estudiantes (Tabla 7):
    - Estudiantes: incluye los temas *estudiante* o a *alumno*. Son los temas marcados en rojo en las tablas de temas (Tabla 26 a Tabla 30).
    - Relacionados: incluye temas *grado*, *tasa* y *3+2*. Aparecen en magenta en las tablas de temas.
- Relacionado con el personal docente (Tabla 8):
    - PDI: incluye el tema *PDI*. Este tema aparece en azul oscuro en las tablas de temas.
    - Investigación: incluye los temas *investigación*, *investigador*, *doctor*, *innovación* y *platinvestucm*. Estos temas aparecen en morado.
    - Profesorado: incluye los temas *docencia*, *profesor*, *profesorado* y *prof*. Utiliza el color cian.
- Relacionado con el personal de administración y servicios (Tabla 9):
    - PAS: incluye el tema *PAS*. Aparece en verde claro.
    - Relacionados: incluye los temas *trabajo*, *empleo* y *servicio*. Aparecen en verde oscuro.

Candidato	Estudiantes	%	Relacionados	%	Total	Proyec. de votos
Carlos Andradas	Estudiante (0,64%) estudiantesucm (0,09%) Alumno (0,07%)	0,80%	Tasa (0,06%)	0,06%	0,86%	31,39%
José Carrillo	Estudiante (0,40%)	0,40%	Grado (0,10%) Tasa (0,09%) 3+2, (0,09%)	0,28%	0,68%	24,82%
Federico Morán	Estudiante (0,23%) Alumno (0,07%)	0,30%	Grado (0,10%) Tasa (0,06%)	0,16%	0,46%	16,79%
Rafael Calduch	Alumno (0,06%) Estudiante (0,04%)	0,10%	Tasa (0,06%) 3+2 (0,04%)	0,10%	0,20%	7,30%
Dámaso López	Estudiante (0,30%)	0,30%	Grado (0,10%) 3+2 (0,07%) Tasa (0,07%)	0,24%	0,54%	19,71%

Tabla 7: Temas relacionados con los estudiantes

La Tabla 7 muestra el porcentaje de aparición, para cada uno de los candidatos, de temas relacionados con estudiantes. Puede apreciarse que los candidatos que más han hablado de los estudiantes son los que más votos han recibido de ellos excepto por el caso de Dámaso Alonso, que aparece dos puestos por encima (tercero en vez de quinto).

En cuanto a los resultados de la Tabla 8, similares a los de la tabla anterior pero para temas relacionados con la docencia y la investigación, las proyecciones de votos dieron ganador a Carrillo seguido de López, Andradas, Calduch y Morán. Aparte del error que supone asignar a Carrillo más votos que a Andradas, los dos candidatos con menor número de tweets, Calduch y Dámaso, son precisamente los que aparecen más desordenados en nuestra predicción (en cuarto y segundo lugar, respectivamente), en gran parte porque al ser su corpus mucho más pequeño, cada mención está desproporcionadamente representada sobre el total de palabras.

Candidato	PDI	%	Investigación	%	Profesorado	%	Total	Proyec. de votos
Carlos Andradas	PDI (0,14%)	0,14%	Investigación (0,24%) Ciencia (0,12%) Investigador (0,03%) platinvestucm (0,03%)	0,42%	Docencia (0,10%) Profesor (0,07%) Profesorado (0,03%)	0,20%	0,76%	24,68%
José Carrillo	PDI (0,14%)	0,14%	Investigación (0,26%) platinvestucm (0,12%) Ciencia (0,09%) Doctor (0,08%) Investigador (0,08%)	0,55%	Docencia (0,10%) Profesor (0,08%)	0,18%	0,87%	28,25%
Federico Morán	-	-	Innovación (0,11%)	0,11%	Docencia (0,06%) Profesor (0,06%)	0,12%	0,23%	7,47%
Rafael Calduch	-	-	Ciencia (0,23%) Investigación (0,04%)	0,27%	Profesor (0,09%) Prof (0,04%)	0,13%	0,40%	12,99%
Dámaso López	PDI (0,10%)	0,10%	platinvestucm (0,59%) Ciencia (0,13%)	0,72%	-	-	0,82%	26,62%

Tabla 8: Temas relacionados con el personal docente

Candidato	PAS	%	Relacionados	%	Total	Proyección de votos
Carlos Andradas	PAS (0,27%)	0,27%	Trabajo (0,14%) Servicio (0,10%)	0,24%	0,51%	30,36%
José Carrillo	PAS (0,15%)	0,15%	Trabajo (0,12%) Empleo (0,12%) Servicio (0,09%)	0,33%	0,48%	28,57%
Federico Morán	PAS (0,17%)	0,17%	Servicio (0,19%)	0,19%	0,36%	21,43%
Rafael Calduch	PAS (0,06%)	0,06%	Servicio (0,04%)	0,04%	0,10%	5,95%
Dámaso López	PAS (0,23%)	0,23%	-	-	0,23%	13,69%

Tabla 9: Temas relacionados con el personal de administración y servicios

Respecto a los temas en relación al personal administrativo y de servicio, la Tabla 9 muestra que la situación es parecida a la de los estudiantes. Según las proyecciones habría ganado Andradas, por delante de Carrillo y no a la inversa, si bien el resto de posiciones se predice correctamente, además de que la diferencia entre sendas puntuaciones es bastante baja (muy inferior a la diferencia entre la puntuación del segundo y el tercer clasificado, que es Morán).

La Tabla 10 resume estas tres proyecciones de voto que se acaban de hacer junto con los resultados reales de cada grupo de electores, y se observa que estos primeros resultados en cuanto a los temas tratados son bastante modestos. Pero queda responder a la cuestión de si la polaridad de los mensajes sobre los candidatos puede ser un indicador de los votos que recibirán.

Candidato	Proyección según temas "estudiantes"	% de voto real	Proyección según temas "profesorado"	% de voto real	Proyección según temas "PAS"	% de voto real
Carlos Andradas	31,39%	36,13%	24,68%	37,78%	30,36%	26,10%
José Carrillo	24,82%	20,56%	28,25%	28,04%	28,57%	48,05%
Federico Morán	16,79%	19,63%	7,47%	24,32%	21,43%	16,71%
Rafael Calduch	7,30%	15,59%	12,99%	5,25%	5,95%	4,41%
Dámaso López	19,71%	8,09%	26,62%	4,61%	13,69%	4,73%

Tabla 10: Proyecciones de voto por tema y resultados reales de la primera vuelta

La Tabla 11 resume las cuatro distintas proyecciones de voto que se han hecho en cuanto a la polaridad de los mensajes. La segunda columna son las proyecciones de voto según el algoritmo trivial y teniendo en cuenta la diferencia entre el porcentaje de tweets positivos y el porcentaje de tweets negativos. En la tercera, el algoritmo es el mismo pero teniendo en cuenta la polaridad media por tweet para hacer la proyección. Las proyecciones de votos no se corresponden con los resultados reales. Tan sólo se predice correctamente que Andradas quedaría por delante de Carrillo.

Candidato	Proyección según algoritmo trivial (P)-(N)	Proyección según algoritmo trivial (M)	Proyección según algoritmo reclasificador (P)+(p)-(n)-(N)	Proyección según algoritmo temático	% de voto real
Carlos Andradas	19,44%	21,67%	18,77%	25,22%	34,61%
José Carrillo	12,69%	13,79%	16,54%	18,29%	27,34%
Federico Morán	22,96%	24,88%	19,49%	17,24%	20,11%
Rafael Calduch	22,69%	21,18%	21,79%	18,20%	11,24%
Dámaso López	22,22%	18,47%	23,41%	21,05%	6,70%

Tabla 11: Proyecciones de voto según algoritmo de extracción de polaridad y resultados reales de la primera vuelta

En la cuarta columna, las proyecciones se obtienen según el algoritmo reclasificador y haciendo una proyección de voto sobre la diferencia entre el número de tweets positivos y negativos de cada

candidato. De nuevo, la predicción que se hace está alejada de la realidad y es prácticamente inversa al orden que se dio en las elecciones.

Por último, en la quinta columna aparecen las proyecciones de voto usando el algoritmo temático. De nuevo, aunque sí hay relación entre este orden y el real, tanto Dámaso López como Rafael Calduch aparecen mucho mejor valorados de lo que lo fueron en las urnas.

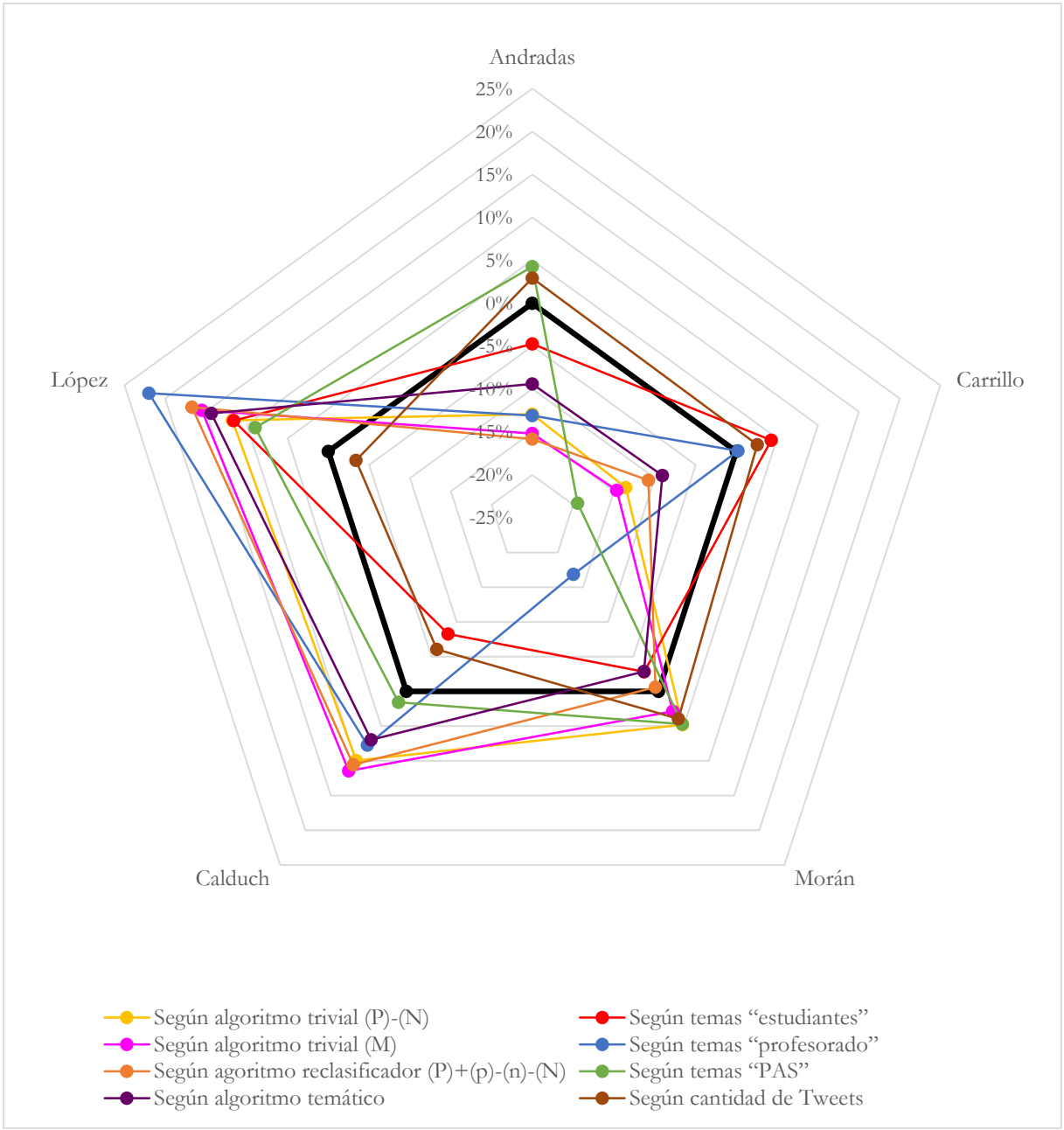

Figura 1: Diferencias en % de votos entre predicción y resultados reales según candidato y análisis empleado

La Figura 1 muestra gráficamente los errores en el porcentaje de votos entre la proyección y el resultado real según el método utilizado. En él puede apreciarse que la predicción hecha según el

análisis de los temas relacionados con los estudiantes es la que mejor se adapta al pentágono negro que representa los resultados reales (0% de error). Pero para intentar dar una medida objetiva de cuál de los sistemas propuestos sería un mejor predictor de las elecciones, para cada una de las proyecciones de voto hechas se ha calculado el vector de diferencias entre el porcentaje de votos predicho y el real para cada candidato. Se utiliza el módulo de este vector como medida de la diferencia entre la predicción y la realidad, y los resultados se muestran en la Tabla 12, según los cuales la predicción más correcta es la obtenida vía el análisis de temas relacionados con los estudiantes (como ya se había apreciado), seguida de la que usa el algoritmo temático (que es el mejor dentro de los algoritmos basados en polaridad). Se observa también cómo el algoritmo reclasificador se comporta peor que el algoritmo trivial.

Predicción	Módulo del error
Según temas "estudiantes"	15,89%
Según temas "profesorado"	31,63%
Según temas "PAS"	22,41%
Según algoritmo trivial (P)-(N)	28,72%
Según algoritmo trivial (M)	24,72%
Según algoritmo reclasificador (P)+(p)-(n)-(N)	27,54%
Según algoritmo temático	20,80%
Según cantidad de Tweets	8,88%

Tabla 12: Errores de las predicciones con respecto a los resultados de la primera vuelta

Como se ha dicho, Dámaso López y Rafael Calduch tienen una cantidad de tweets asociados muy bajo. Con lo que se podría excluir a estos candidatos de todos los análisis y hacer predicciones únicamente para los tres primeros candidatos. Así, la Tabla 13 compara las proyecciones de voto hechas para los análisis por temas y el resultado real (equivalente a la Tabla 10) pero considerando únicamente los tres candidatos con mayor número de tweets relacionados. Puede observarse que la predicción basada en el tema "estudiantes" es muy buena, y las otras dos proyecciones de votos descartan claramente al tercer candidato, que aparece muy alejado de los porcentajes del primero y del segundo (que aparecen en orden en inverso, pero a una distancia cercana).

Candidato	Proyección según temas "estudiantes"	% de voto real	Proyección según temas "profesorado"	% de voto real	Proyección según temas "PAS"	% de voto real
Carlos Andradas	43,00%	47,34%	40,86%	41,91%	37,78%	28,73%
José Carrillo	34,00%	26,94%	46,77%	31,11%	35,56%	52,88%
Federico Morán	23,00%	25,72%	12,37%	26,98%	26,67%	18,39%

Tabla 13: Proyecciones de voto por tema y resultados reales de la primera vuelta (sólo tres candidatos)

Candidato	Proyección según algoritmo trivial (P)-(N)	Proyección según algoritmo trivial (M)	Proyección según algoritmo reclasificador (P)+(p)-(n)-(N)	Proyección según algoritmo temático	% de voto real
Carlos Andradas	35,92%	35,29%	34,25%	41,52%	42,18%
José Carrillo	22,86%	23,03%	30,19%	30,11%	33,32%
Federico Morán	41,22%	41,68%	35,56%	28,38%	24,50%

Tabla 14: Proyecciones de voto según algoritmo de extracción de polaridad y resultados reales de la primera vuelta (sólo tres candidatos)

La Tabla 14 es análoga a la Tabla 11 pero considerando una elección de tres candidatos. De nuevo el algoritmo temático es el que da mejores resultados, en este caso una predicción muy cercana al resultado real. Para terminar este análisis sobre los tres primeros candidatos, se ha creado la Figura 2, análoga a la Figura 1 pero considerando sólo tres candidatos. Se aprecia que la predicción hecha utilizando el algoritmo temático es la que más se acerca a la realidad.

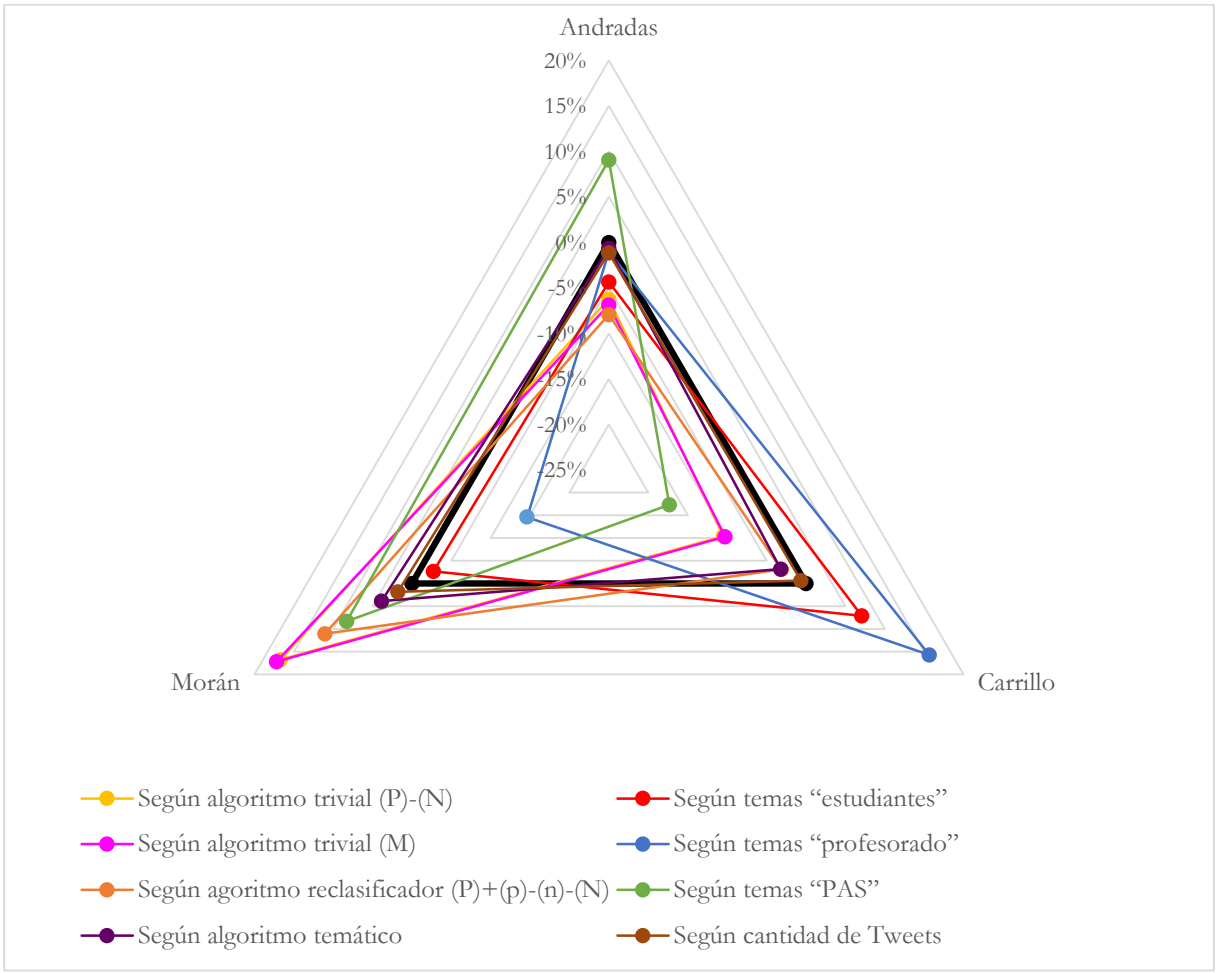

Figura 2: Diferencias en % de votos entre predicción y resultados reales según candidato y análisis empleado (sólo tres candidatos)

Por último la Tabla 15 utiliza de nuevo la métrica del módulo del vector de error en el porcentaje de voto para volver a mostrar que el algoritmo temático es el que mejor resultados produce.

Predicción	Módulo del error
Según temas "estudiantes"	8,73%
Según temas "profesorado"	21,45%
Según temas "PAS"	21,23%
Según algoritmo trivial (P)-(N)	20,69%
Según algoritmo trivial (M)	21,17%
Según algoritmo reclasificador (P)+(p)-(n)-(N)	13,96%
Según algoritmo temático	5,07%
Según cantidad de Tweets	2,21%

Tabla 15: Errores de las predicciones con respecto a los resultados de la primera vuelta (sólo tres candidatos)

5. Conclusiones y líneas de trabajo futuro

Según se ha visto en la sección anterior, el análisis de opinión de los tweets para el caso de las elecciones a Rector de la UCM no parece ser una medida demasiado útil para predecir los resultados. Este resultado se corresponde con los resultados obtenidos por Choy (Choy y otros, 2011) y por Gayo-Avello (2012). Sin embargo, pueden hacerse algunas matizaciones de interés.

En primer lugar, el desequilibrio en el tamaño de las secciones del corpus referidas a cada candidato parece influir en la fiabilidad de los resultados: en los casos de Rafael Calduch y Dámaso López, se dispone de una cantidad muy inferior de tweets relacionados. Debido a ello se observa que los resultados obtenidos para estos dos candidatos son mucho menos precisos, y son habitualmente estos dos candidatos los que aparecen desordenados en nuestras predicciones. Como se ha visto, eliminar del análisis a los candidatos con un número de tweets asociados muy pequeño mejora notablemente los resultados.

En segundo lugar, se ha llegado a la conclusión de que, en nuestro caso, el número de tweets escritos sobre cierto candidato es la medida que tiene mayor correlación con los resultados de las elecciones. Se observa en la Tabla 31 que las proyecciones hechas basándose en este valor son muy acertadas, llegando a predecir no sólo el orden de candidatos más votado, sino incluso la diferencia en porcentaje de votos entre los tres primeros. Tanto en la Figura 1 como sobre todo en la Figura 2 se aprecia claramente que la coincidencia es casi exacta. No se tiene constancia de que este valor se haya tenido en cuenta en otros trabajos similares, pero a la luz de los datos este valor podría ser un indicio importante sobre el resultado final de las elecciones y merece ser estudiado en más profundidad.

Este trabajo, además de corroborar estudios anteriores sobre la poca eficacia del análisis de opinión de los tweets como mecanismo para predecir resultados electorales, ha contribuido a crear nuevos recursos lingüísticos para el análisis de opinión en lengua española que serán de utilidad en nuevos trabajos de investigación en este campo. En concreto se ha desarrollado un corpus etiquetado, un lexicón de polaridad, un lematizador, un flexionador, un silabeador y un analizador sintáctico para extraer los modificadores y el núcleo de los sintagmas nominales y verbales de oraciones no siempre bien construidas como es el caso de los mensajes cortos.

Asimismo, se ha desarrollado un nuevo método y herramienta para extraer la opinión en textos cortos basada en el analizador sintáctico desarrollado y el lexicón de polaridad, con una eficacia similar a los clasificadores estadísticos tradicionales basados en corpus de polaridad. Esta herramienta puede aplicarse en los análisis de opinión en lenguas para las que no existen o son escasos los corpus de polaridad.

Como trabajo futuro, se plantean las siguientes mejoras del sistema de análisis de opinión desarrollado:

1. La mejora del analizador sintáctico aumentando su robustez frente a frases agramaticales. Esta mejora tiene un interés significativo porque ayudaría al procesamiento de los millones de mensajes cortos que circulan en internet diariamente.
2. Completar el lexicón de polaridad desarrollado. Se ha optado por una vía rápida para crear el Diccionario de Polaridad, pero se pierden muchos lemas en nuestra traducción. En futuros trabajos, sería muy recomendable crear un propio Diccionario de Polaridad etiquetado manualmente añadiendo más términos a los ya utilizados.
3. Efectuar un análisis específico para los adverbios, en especial los de negación y los intensificadores. Algo que sería muy deseable ya que el Diccionario de Polaridad ya incluye información sobre la intensidad de los adverbios, pero no se utiliza.
4. Análisis específico de las conjunciones y preposiciones como *pero*, *aunque*, *sin* o *contra*. Son términos habituales que no se han tratado de ninguna forma especial cuando en realidad modifican sustancialmente la polaridad de los sintagmas que les siguen o preceden. Es probable que teniéndolos en cuenta se obtengan valores más precisos de la polaridad de las frases.
5. Implementar un identificador de expresiones hipotéticas. Una expresión hipotética suaviza el valor de la polaridad, puesto que no expresa la realidad, sino que se teoriza sobre algo. Se podrían identificar por la aparición de modos subjuntivo o condicional, conjunciones

como *si*, o signos de admiración o interrogación. También podrían buscarse expresiones "gritadas" escritas únicamente con mayúsculas.

6. Análisis específico de las oraciones de predicado nominal. Si bien los verbos *ser* y *estar* ya se identifican y se utilizan para marcar una oración como de predicado nominal, nuestra manera de analizar la frase en árbol hace que usualmente el atributo quede relativamente lejos del sintagma nominal al que se refiere. En realidad, el grado de relación entre el núcleo del sujeto y el del atributo debería ser la misma que la existente entre el núcleo de un sintagma y su modificador, pero el sistema actual no lo hace así.
7. Identificación de *smileis*, tipo :), :-(o ;-P. Estos símbolos son una de las formas más inmediatas de reconocer si una frase tiene ánimo de ser tomada como positiva (o humorística/irónica), o si se trata de algo serio o desagradable.
8. Por último, sería muy interesante tener en cuenta el número de retweets de cada tweets, y el número de veces que ha podido ser marcado como favorito por el resto de usuarios. En este trabajo todos los tweets únicos se tratan con el mismo peso sobre el corpus, pero es evidente que no es igual de importante un tweet que quizá nunca nadie haya leído, que un tweet que ha sido reenviado o se haya marcado como favorito por cientos o miles de usuarios.

Agradecimientos

Me gustaría dar las gracias a las personas que han contribuido a que dentro de unas semanas pueda referirme a mí mismo como lingüista. En primer lugar me gustaría agradecer a Ana María Fernández-Pampillón su trabajo como directora de este Trabajo de Fin de Grado. Junto a Víctor Peinado han conseguido que la lingüística computacional pasara de asignatura a hobby y me han ayudado mucho dirigiéndome y aportándome información y sugerencias muy valiosas. Se nota en sus clases que hacen todo lo que está en su mano (que es mucho) para trasladarnos su interés por la matera y hacernos aprender (incluso aunque no quisiéramos). Ojalá hubiera más como ellos…

… y de hecho los hay. Me gustaría agradecer también a todos aquellos profesores que a lo largo de los últimos cuatro años han conseguido que asistir a clase fuera motivo de alegría y me han convencido con sus interesantes clases de que lo mío son las lenguas y el lenguaje. En especial me gustaría mencionar a Raquel Hidalgo, Светлана Малявина, Carlos Cid, Ana Matesanz, Françoise Baudet, Bautista Horcajada y a Таня Лалева. Muchas gracias a todos. Me habéis enseñado mucho.

Y por último, querría agradecer a mis amigos y a mi familia su comprensión cuando desaparecía durante exámenes, y por saber alegrarme en los días grises. ¡De aquí a tener el sillón de la Ñ en la RAE sólo hay un paso!

Bibliografía

Asur, S., y Huberman, B. (2010). Predicting the future with social media. *International Conference on Web Intelligence and Intelligent Agent Technology (WIIAT)*, *1*, 492-499.

Bird, S., Klein, E., y Loper, E. (2009). Natural language processing with Python. O'Reilly Media, Inc.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., y Reynar, J. (2008). Building a sentiment summarizer for local service reviews. *Workshop on NLP in the Information Explosion Era*, 14.

Bollen, J., Mao, H. y Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*, 1-8.

Brooke, J., Tofiloski, M., y Taboada, M. (2009). Cross-Linguistic Sentiment Analysis: From English to Spanish. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 50-54.

Choy, M., Cheong, M. L., Laik, M. N., y Shung, K. P. (2011). A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. *arXiv preprint arXiv:1108.5520*.

De Smedt, T., y Daelemans, W. (2012). "Vreselijk mooi!" (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives. *Proceedings of the Language Resources and Evaluation Conference* (LREC), *8*, 3568-3572.

Esuli, A., y Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of the Language Resources and Evaluation Conference* (LREC), *6*, 417-422.

Gayo-Avello, D. (2012). "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" - A Balanced Survey on Election Prediction using Twitter Data. *arXiv preprint arXiv:1204.6441*.

Grefenstette, G., y Tapanainen, P. (1994). What is a word, what is a sentence? Problems of Tokenisation. *International Conference on Computational Lexicography*, *3*, 79–87.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., y Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *International Conference on Weblogs and Social Media (WSM)*, *11*, 122-129.

Ogneva, M. (2010). How Companies Can Use Sentiment Analysis to Improve Their Business. *Mashable*. Recuperado de http://mashable.com/2010/04/19/sentiment-analysis/

Pang, B., Lee, L. y Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, *10*, 79–86.

Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the Association for Computational Linguistics*, *40*, 417–424.

Tweepy [Software]. (2015). Recuperado de http://tweepy.readthedocs.org/en/v3.2.0/

Apéndices

Candidato	Tweets	Positivos (P)	Neutros	Negativos (N)	Polaridad total	Polaridad media (M)	Proyección votos (P)-(N)	Proyec. votos (M)
Carlos Andradas	2.383	398 (16,70%)	1854 (77,80%)	131 (5,50%)	209,95	0,088	19,44%	21,67%
José Carrillo	1.899	235 (12,37%)	1.568 (82,57%)	96 (5,06%)	106,05	0,056	12,69%	13,79%
Federico Morán	1.527	288 (18,86%)	1.153 (75,51%)	86 (5,63%)	154,4	0,101	22,96%	24,88%
Rafael Calduch	329	52 (15,81%)	268 (81,46%)	9 (2,74%)	28,3	0,086	22,69%	21,18%
Dámaso López	211	33 (15,64%)	172 (81,52%)	6 (2,84%)	15,8	0,075	22,22%	18,47%
Total	**6.349**	**1.006 (15,85%)**	**5.023 (79,11%)**	**328 (5,27%)**	**517,9**	**0,082**		

Tabla 16: Recuentos de tweets positivos, negativos y neutros, medias de polaridad y proyecciones de voto con el algoritmo trivial

Candidato	Tweets	Positivos (P)	Neutros antes de reclasificar			Negativos (N)	Exactitud clasificador	Proyección de votos (P)+(p)-(n)-(N)
			⇨ Positivos (p)	⇨ Neutros	⇨ Negativos (n)			
Carlos Andradas	2.383	398 (16,70%)	1854 (77,80%)			131 (5,50%)	80,37%	18,77%
			1.540 (64,62%)	73 (3,06%)	241 (10,11%)			
José Carrillo	1.899	235 (12,37%)	1.568 (82,57%)			96 (5,06%)	92,54%	16,54%
			1.254 (66,03%)	21 (1,11%)	293 (15,43%)			
Federico Morán	1.527	288 (18,86%)	1.153 (75,51%)			86 (5,63%)	76,32%	19,49%
			956 (62,61%)	81 (5,30%)	116 (7,60%)			
Rafael Calduch	329	52 (15,81%)	268 (81,46%)			9 (2,74%)	46,15%	21,79%
			230 (69,91%)	16 (4,86%)	22 (6,69%)			
Dámaso López	211	33 (15,64%)	172 (81,52%)			6 (2,84%)	66,67%	23,41%
			152 (72,04%)	14 (6,64%)	6 (2,84%)			
Total	**6.349**	**1.006 (15,85%)**	**5.023 (79,11%)**			**328 (5,27%)**	**80,82%**	
			4.132 (65,08%)	**205 (3,23%)**	**678 (10,68%)**			

Tabla 17: Recuentos de tweets positivos, negativos y neutros, medias de polaridad y proyecciones de voto con el algoritmo reclasificador

Candidato	Nº Palabras asociadas	Palabras positivas	Palabras negativas	Polaridad total	Polaridad por palabra asociada	Proyección de votos
Carlos Andradas	95	87	8	54,7	0,575	25,22%
José Carrillo	45	33	12	18,8	0,417	21,05%
Federico Morán	58	47	11	22,8	0,393	18,29%
Rafael Calduch	13	11	2	5,4	0,415	18,20%
Dámaso López	5	5	0	2,4	0,480	17,24%
Total	**216**	**183**	**33**	**104,1**	**0,482**	

Tabla 18: Polaridades asociadas a los términos relacionados con los candidatos y proyecciones de voto con el algoritmo temático

Palabra asociada	Apariciones	Polaridad total
Ganar	41	32,8
Querer	16	8,0
Mejor	8	8,0
Bueno	4	2,0
Gustar	3	1,2
Fácil	2	0,8
Lleno	2	0,6
Crítico	2	0,4
Querido	1	0,7
Valiente	1	0,7
Sabio	1	0,7
Seguro	1	0,5
Interesante	1	0,5
Igual	1	0,5
Preciso	1	0,4
Vivo	1	0,3
Cercano	1	0,1
Arrancar	1	-0,2
Activo	1	-0,2
Mínimo	1	-0,3
Acabarse	1	-0,5
Culpable	1	-0,5
Serio	1	-0,5
Peligroso	1	-0,6
Malo	1	-0,7
Total	**95**	**54,7**
Polaridad por palabra		**0,575**

Tabla 19: Palabras asociadas a Carlos Andradas y sus polaridades

Palabra asociada	Apariciones	Polaridad total
Ganar	14	11,2
Mejor	6	6,0
Querer	5	2,5
Interesante	2	1,0
Suspender	2	-1,0
Malo	2	-1,4
Total	1	1,0
Sano	1	0,5
Igual	1	0,5
Original	1	0,5
Capaz	1	0,5
Gustar	1	0,4
Curioso	1	-0,1
Desgraciadamente	1	-0,1
Arrancar	1	-0,2
Pobre	1	-0,4
Debido	1	-0,5
Traicionar	1	-0,5
Serio	1	-0,5
Peligroso	1	-0,6
Total	**45**	**18,8**
Polaridad por palabra		**0,417**

Tabla 20: Palabras asociadas a José Carrillo y sus polaridades

Palabra asociada	Apariciones	Polaridad total
Gustar	9	3,6
Mejor	7	7,0
Ganar	7	5,6
Querer	5	2,5
Triunfar	2	1,6
Bueno	2	1,0
Feliz	2	1,0
Seguro	2	1,0
Interesante	2	1,0
Fácil	2	0,8
Vivo	2	0,6
Serio	2	-1,0
Bello	1	0,7
Amar	1	0,5
Realista	1	0,5
Disfrutar	1	0,4
Importante	1	0,1
Activo	1	-0,2

Palabra asociada	Apariciones	Polaridad total
Mínimo	1	-0,3
Peor	1	-0,4
Menudo	1	-0,5
Culpable	1	-0,5
Duro	1	-0,5
Abandonar	1	-0,5
Sencillo	1	-0,5
Enfadar	1	-0,7
Total	**58**	**22,8**
Polaridad por palabra		**0,393**

Tabla 21: Palabras asociadas a Federico Morán y sus polaridades

Palabra asociada	Apariciones	Polaridad total
Ganar	2	1,6
Acabarse	2	-1,0
Impresionante	1	1,0
Mejor	1	1,0
Seguro	1	0,5
Positiva	1	0,5
Bueno	1	0,5
Realista	1	0,5
Querer	1	0,5
Rápido	1	0,2
Importante	1	0,1
Total	**13**	**5,4**
Polaridad por palabra		**0,415**

Tabla 22: Palabras asociadas a Rafael Calduch y sus polaridades

Palabra asociada	Apariciones	Polaridad total
Querer	2	1,0
Seguro	1	0,5
Encantar	1	0,5
Gustar	1	0,4
Total	**5**	**2,4**
Polaridad por palabra		**0,48**

Tabla 23: Palabras asociadas a Dámaso López y sus polaridades

Categoría	Nº de opiniones	Exactitud algoritmo trivial	Exactitud algoritmo reclasificador
Coches	23	56,52%	56,52%
Hoteles	21	71,43%	71,43%
Lavadoras	33	69,70%	72,72%
Libros	25	52,00%	52,00%
Móviles	19	57,89%	57,89%
Música	24	62,50%	66,67%
Ordenadores	21	38,10%	42,86%
Películas	23	60,87%	60,87%

Categoría	Nº de opiniones	Exactitud algoritmo trivial	Exactitud algoritmo reclasificador
Total ponderado	**189**	**59,29%**	**60,85%**

Tabla 24: Evaluación del extractor de polaridad con el corpus de reseñas etiquetadas de Julian Brooke

Etiqueta manual	Etiqueta automática			**Total**
	Positivo	Neutro	Negativo	
Positivo	**5**	5	0	**10**
Neutro	6	**75**	0	**81**
Negativo	0	7	2	**9**
Total	**11**	**87**	**2**	**100**

Tabla 25: Precisión del extractor de polaridad (algoritmo trivial) en 100 tweets escogidos aleatoriamente

Tema	Nº de apariciones	% del total
Estudiante	**97**	**0.64%**
Proyecto	60	0.40%
Gracias	58	0.38%
Pas	**41**	**0.27%**
Investigación	**37**	**0.24%**
Apoyo	30	0.20%
Plan	28	0.18%
Propuesta	27	0.18%
Calidad	23	0.15%
Debate	23	0.15%
Pdi	**21**	**0.14%**
Formación	21	0.14%
Trabajo	**21**	**0.14%**
Programa	19	0.13%
Ciencia	**18**	**0.12%**
Centro	17	0.11%
Cambio	16	0.11%
Ilusión	16	0.11%
Comunidad	16	0.11%
Encuentro	16	0.11%
Voto	15	0.10%
Docencia	**15**	**0.10%**
Agenda	15	0.10%
Servicio	**15**	**0.10%**
Estudiantesucm	**14**	**0.09%**
Profesor	**11**	**0.07%**
Alumno	**10**	**0.07%**
Tasa	**9**	**0.06%**
Profesorado	**5**	**0.03%**
Investigador	**5**	**0.03%**
Platinvestucm	**4**	**0.03%**

Tabla 26: Temas principales de los tweets de Carlos Andradas

Tema	Nº de apariciones	% del total
Proyecto	99	0.57%
Gracias	97	0.56%
Defiendelopúblicoucm	97	0.56%
Estudiante	**69**	**0.40%**
Programa	58	0.34%
Voto	55	0.32%
Gestión	47	0.27%
Investigación	**45**	**0.26%**
Deuda	44	0.26%
Ganar	43	0.25%
Apoyo	27	0.16%
Pas	**26**	**0.15%**
Calidad	25	0.14%
Pdi	**24**	**0.14%**
Consulta	23	0.13%
Recorte	23	0.13%
Platinvestucm	**21**	**0.12%**
Empleo	**20**	**0.12%**
Trabajo	**20**	**0.12%**
Calendario	20	0.12%
Comunidad	19	0.11%
Política	19	0.11%
Plan	18	0.10%
Estudio	18	0.10%
Madrid	18	0.10%
Grado	**18**	**0.10%**
Compromiso	18	0.10%
Docencia	**18**	**0.10%**
Tasa	**16**	**0.09%**
3+2	**15**	**0.09%**
Servicio	**15**	**0.09%**
Ciencia	**15**	**0.09%**
Doctor	**13**	**0.08%**
Profesor	**13**	**0.08%**
Investigador	**12**	**0.08%**

Tabla 27: Temas principales de los tweets de José Carrillo

Tema	Nº de apariciones	% del total
Cambio	31	0.30%
Estudiante	**24**	**0.23%**
Investigación	**24**	**0.23%**
Gracias	20	0.19%
Servicio	**20**	**0.19%**
Programa	19	0.18%

Tema	Nº de apariciones	% del total
Pas	**18**	**0.17%**
Propuesta	16	0.15%
Plan	15	0.14%
Conocimiento	15	0.14%
Apoyo	14	0.13%
Investigador	**15**	**0.14%**
Cuento	12	0.11%
Innovación	**12**	**0.11%**
Prestigio	12	0.11%
Grado	**11**	**0.10%**
Campus	11	0.10%
Calidad	11	0.10%
Comunidad	11	0.10%
Problema	10	0.10%
Voto	10	0.10%
Persona	10	0.10%
Acceso	10	0.10%
Gestión	10	0.10%
Ilusión	10	0.10%
Derecho	10	0.10%
Actividad	10	0.10%
Alumno	**7**	**0.07%**
Docencia	**6**	**0.06%**
Profesor	**6**	**0.06%**
Tasa	**6**	**0.06%**

Tabla 28: Temas principales de los tweets de Federico Morán

Tema	Nº de apariciones	% del total
Modelo	21	0.45%
Comunidad	18	0.38%
Programa	15	0.32%
Ucmccinf	15	0.32%
Entrevista	15	0.32%
Oportunidad	14	0.30%
Visita	14	0.30%
Situación	12	0.26%
Cambio	11	0.23%
Decisión	11	0.23%
Ciencia	**11**	**0.23%**
Reto	10	0.21%
Proponer	9	0.19%
Propuesta	8	0.17%
Orgullo	7	0.15%
Voto	7	0.15%
Sala	7	0.15%
Inforadio_Ucm	7	0.15%

Tema	Nº de apariciones	% del total
Centro	7	0.15%
Debate	6	0.13%
Estatuto	6	0.13%
Información	6	0.13%
Base	5	0.11%
Colegio	5	0.11%
Medicina	5	0.11%
Futuro	5	0.11%
Grupo	5	0.11%
Gobierno	5	0.11%
Ucm_Derecho	5	0.11%
Apoyo	5	0.11%
Aceptar	5	0.11%
Profesor	**4**	**0.09%**
Alumno	**3**	**0.06%**
Pas	**3**	**0.06%**
Tasa	**3**	**0.06%**
3+2	**2**	**0.04%**
Servicio	**2**	**0.04%**
Investigación	**2**	**0.04%**
Prof	**2**	**0.04%**
Estudiante	**2**	**0.04%**

Tabla 29: Temas principales de los tweets de Rafael Calduch

Tema	Nº de apariciones	% del total
Platinvestucm	**18**	**0.59%**
Programa	15	0.49%
Estudiante	**9**	**0.30%**
Presentar	8	0.26%
Medio	7	0.23%
Pas	**7**	**0.23%**
Presentación	7	0.23%
Reflexión	7	0.23%
Noticia	7	0.23%
Excelencia	7	0.23%
Gracias	6	0.20%
Debate	6	0.20%
Lugar	5	0.16%
Promoción	5	0.16%
Palabra	5	0.16%
País	5	0.16%
Forma	5	0.16%
Educación	4	0.13%
Problema	4	0.13%
Estudio	4	0.13%
Ciencia	**4**	**0.13%**

Tema	Nº de apariciones	% del total
Duda	4	0.13%
Objetivo	4	0.13%
Comunidad	4	0.13%
Esfuerzo	4	0.13%
Trabajo	**4**	**0.13%**
Caso	4	0.13%
Ranking	4	0.13%
Vida	3	0.10%
Pdi	**3**	**0.10%**
Movilidad	3	0.10%
Modelo	3	0.10%
Posibilidad	3	0.10%
Institución	3	0.10%
Filólogo	3	0.10%
Cuenta	3	0.10%
Financiación	3	0.10%
Centro	3	0.10%
Grado	**3**	**0.10%**
Historia	3	0.10%
Plantilla	3	0.10%
Comunicación	3	0.10%
Mundo	3	0.10%
Condición	3	0.10%
Filología	3	0.10%
3+2	**2**	**0.07%**
Tasa	**2**	**0.07%**

Tabla 30: Temas principales de los tweets de Dámaso López

Candidato	Proyección según nº de tweets sobre candidato	% de voto real
Carlos Andradas	37,53%	34,61%
José Carrillo	29,91%	27,34%
Federico Morán	24,05%	20,11%
Rafael Calduch	5,18%	11,24%
Dámaso López	3,32%	6,70%

Tabla 31: Proyecciones de voto según número de tweets sobre candidato y resultados reales de la primera vuelta