

TECNALIA-UNED @ TASS: Using a linguistic approach for sentiment analysis

TECNALIA-UNED en TASS: Uso de un enfoque lingüístico para el análisis de sentimientos

Esther Villar Rodríguez OPTIMA Unit TECNALIA esther.villar@tecnalia.com	Ana I. Torre Bastida OPTIMA Unit TECNALIA isabel.torre@tecnalia.com	Ana García-Serrano ETSI Informática UNED agarcia@lsi.uned.es	Marta González Rodríguez OPTIMA Unit TECNALIA marta.gonzalez@tecnalia.com
---	---	--	---

Resumen. En el marco de TASS, se ha desarrollado un framework, probado ya previamente con el corpus del 2012 de este workshop, con el objeto de detectar la polaridad de un tweet. Este entorno realiza un procesamiento lingüístico por fases para abordar temas complejos como la negación o el tratamiento de enfatizadores (permitiendo distinguir entre los distintos niveles de polaridad contemplados en el taller).

Palabras clave. Análisis de sentimiento, procesamiento lingüístico, negación, detección de relevancias.

Abstract. In the framework of TASS, a framework has been developed. It has previously tested with the 2012 TASS corpus, in order to detect polarities in a tweet. The methodology is to execute a stage by stage linguistic process trying to deal with complex issues such as negation detection and emphasisers treatment (aiming at distinguishing the range of polarity levels).

Keywords. Sentiment Analysis, Linguistic Processing, Negation, Relevance Detection.

1 Introduction

TASS es un workshop orientado al análisis de sentimientos y de reputación sobre tweets en español. Así, aborda 4 tareas:

- Análisis de Sentimiento a nivel global. Pretende clasificar el tweet de acuerdo a los cinco niveles típicos ($N+$, N , Neu , P , $P+$).
- Detección del topic. Los temas incluidos son *política*, *fútbol*, *literatura* y *entretenimiento*.
- Análisis de Sentimiento a nivel de entidad. Orientado a clasificar cada tweet del corpus de política de acuerdo a los tres niveles básicos (N , Neu , P) sobre el corpus de política. Las entidades implicadas son las siguientes: Partido Popular (PP), Partido Socialista Obrero Español (PSOE), Izquierda Unida (IU) y Unión, Progreso y Democracia (UPyD).

- Identificación de la orientación política. Entendiendo tendencia dentro de las siguientes categorías: *Derecha*, *Izquierda*, *Centro e indefinida*.

Existen dos corpus dependiendo de cuál sea la tarea a ejecutar:

- Corpus general (para las tareas 1, 2 y 4): Compuesto por 68000 tweets en español recogidos durante el 2012. Tratan sobre 150 entidades pertenecientes al mundo de la política, medios de comunicación, economía y cultura. El corpus está dividido en el conjunto de training (90%) y el de test (10%). El formato sigue los estándares del corpus publicado en el 2012 que permite abordar el análisis de sentimiento global además de un análisis a nivel de entidad (como se puede comprobar por las anotaciones del ejemplo siguiente).

```

<tweet>
  <tweetid>0000000001</tweetid>
  <user>usuario1</user>
  <content><![CDATA['UPyD conta-
rá casi seguro con grupo gracias al Foro As-
turias.]]></content>
  <date>2011-12-02T00:21:01</date>
  <lang>es</lang>
  <sentiments>
    <polari-
ty><value>P</value><type>AGREEMENT<
/type></polarity>
    <polari-
ty><entity>UPyD</entity><value>P</value>
</type>AGREEMENT</type></polarity>
    <polari-
ty><entity>Foro_Asturias</entity><value>P
</value><type>AGREEMENT</type></pola
rity>
  </sentiments>
  <topics>
    <topic>política</topic>
  </topics>
</tweet>

```

- Corpus de política (para la tarea 3): Compuesto por 2500 tweets recogidos en las pasadas elecciones generales de 2011. A diferencia del corpus anterior, el training viene anotado bajo tres niveles de sentimiento (N, Neu, P) junto con uno adicional None para indicar la no existencia de sentimiento. La diferencia entre el formato de estos tweets y el formato seguido por el corpus general radica en la existencia de un nuevo atributo dentro del tag destinado a la polaridad llamado *source*.

```

<tweet>
<tweetid>137231808990412800</tweetid>
  <user>marianarajay</user>
  <content>
<![CDATA['@marianorajoy Por favor, ami-
gosh, no me votéish que me lo he penshado
mejor con este tshunami que se me viene en-
cima.]]> </content>
  <date>2011-10-17T19:13:07</date>
  <lang>es</lang>
<sentiments>

```

```

  <polarity> <value>N</value>
<type>AGREEMENT</type></polarity>
  <polarity><entity source="PP">
@marianorajoy</entity><value>N</value><t
ype> AGREEMENT </type></polarity>
  </sentiments>
  <topics>
    <topic>política</topic>
  </topics>
</tweet>

```

Las métricas empleadas en la evaluación son las típicas para los sistemas de Information Retrieval (IR): precisión, recall y la medida F.

Para abordar cualquier tarea sobre un corpus de tweets siguiendo el enfoque lingüístico, es necesario aplicar a los mensajes una normalización [4]. Este pre-procesamiento suele incluir la sustitución de hashtags, urls y targets, el remplazo de caracteres repetidos, la traducción de las abreviaturas e inclusive, la corrección ortográfica [2]. Algunos de estos fenómenos no son errores, sino que son los mecanismos empleados para enfatizar ideas y es por ello por lo que hay que prestarles especial atención.

Otro de los elementos relevantes son los emoticonos. En el artículo [4] toman esta característica típica del micro-blogging como característica binaria para el clasificador y reportan buenos resultados en el modelo. Otros trabajos como [8], en cambio, los emplean para recolectar nuevos mensajes subjetivos y aprender y entrenar los clasificadores con corpus más potentes. En los enfoques léxicos [1, 2], los emoticonos se utilizan como indicios tempranos de polaridad y conforman una característica más a tener en cuenta.

Para la obtención de características más complejas y menos evidentes, también se ejecutan la tokenización, la lematización y la extracción de la etiqueta gramatical (que ayuda además a distinguir qué componentes se deben procesar por los módulos) [4, 9].

Otros recursos casi imprescindibles para esta tarea, son los diccionarios de polaridad como: SentiWordnet [10], DAL [1], MPQA [4],... Estos términos junto con su polaridad pasan a formar parte del vector de características en los sistemas de Minería de Datos o se convierten en anotaciones relevantes en los enfoques léxicos.

En los enfoques basados en Minería [3, 4, 7], la detección de la negación, ha sido superficialmente tratada y es ahí donde los enfoques basados en lingüística pretenden demostrar su potencial [6, 8].

La segmentación es uno de los procesos más debatidos, sobre todo cuando se considera la posibilidad de que existan diferentes temáticas en un mismo texto o cuando coexisten emociones dispares. Según [8], se debería considerar cada palabra como mayor o menor portadora de polaridad según su nivel en el árbol sintáctico. En este entorno se ha trasladado este concepto a las proposiciones, fundamentando la teoría en que la proposición principal tiene mucha más relevancia en la determinación de la polaridad. Esto es además porque las subordinadas, por definición, no tienen personalidad individual propia, y su polaridad debe tener distinto valor o influencia en la global.

2 Descripción del entorno

El entorno usado para el *run* estudiado asigna polaridades a los tweets basándose en un enfoque lingüístico. Este enfoque sigue las premisas del procesamiento por fases abordando las tareas siguientes:

- **Preprocesamiento:** actividades para la preparación de textos no normalizados.
 1. Eliminación de caracteres repetidos. Tq mi AMOOOOOORR → Tq mi AMOR
 2. Corrección de mayúsculas. Tq mi AMOR → Tq mi amor
 3. Detección de abreviaturas mediante un diccionario de 43 entradas con las más populares empleadas en las redes sociales. Tq mi amor → Te quiero mi amor
- **Cálculo de la polaridad de cada elemento:** Por medio de diccionarios propios se asigna la polaridad previa a emoticonos (diccionario de 123 entradas en su mayoría extraídas de Wikipedia), interjecciones (diccionario con 75 expresiones regulares para comprender un mayor rango de variaciones) y términos in-

dependientes (gracias al diccionario SentiWordnet).

- **Segmentación:** permite definir las estructuras portadoras de semántica y sentimiento además de marcar los límites de ideas diferentes (ámbito). También facilita el tratamiento de la negación o de los modificadores permitiendo cercar la estructura sobre la que recae la acción dado que también se segmenta a nivel sintáctico de sintagmas. Como punto clave del trabajo, esta tarea ayuda a asignar relevancias dependiendo de la importancia de la expresión de la idea dentro del texto (su nivel sintáctico dentro de la oración). Una posible segmentación con sus correspondientes niveles podría ser la siguiente:

Prop N.1 (coord.)

Prop N.2 (subord.)

- Tokens
- Sintag Nom.
 - Prop N.3 (adjetiva)
- Sintag Prep.
- Modificadores
- Emoticonos

...

Prop N.1 (coord.)

...

Para esta tarea se utiliza la herramienta FreeLing aunque se han realizado algunos refinamientos para evitar los errores de las pruebas.

- **Tratamiento de la negación:** con diccionarios propios que contienen las formas posibles en el idioma español de negar un hecho (tokens independientes, conjunciones, locuciones, sintagmas preposicionales) y aplicando diferentes tratamientos dependiendo del caso y del ámbito que este marca. En la lengua castellana, existen varios mecanismos para negar un hecho. Los contemplados en este entorno son los siguientes:
 - Anteposición del adverbio “no” al verbo principal de la oración: *No vine*.

- Mediante el uso de otros adverbios de negación (“tampoco”, “nunca”, “jamás”): *Nunca dije que vendría.*
- Con la utilización de conjunciones negativas (“ni”): *Ni vendré hoy ni vendré mañana.*
- A través de locuciones adverbiales o combinaciones enfáticas (“nunca jamás”, “en mi vida”): *Nunca jamás volveré.*
- Con un sintagma preposicional (SP) (“en vez de”, “sin”, “a falta de”): *En vez de volver, prefiero ir a otro sitio.*

Para identificar este fenómeno son por tanto estrictamente necesarias las etiquetas gramaticales que actúan como desambiguadores y otorgan la certeza de encontrarnos ante uno de los ejemplos anteriores. Para determinar el ámbito de dicha negación, se emplean dos distinciones:

1. Ante uno de los 4 primeros casos, el ámbito estaría formado por la proposición simple que contenga la estructura de negación.

(1) [*Ni vendré hoy*] [*ni vendré mañana*].

Existen dos elementos de negación “ni” que afectan individualmente a una de las dos proposiciones coordinadas copulativas.

2. Si se da el caso 5, el ámbito estaría formado por los elementos contenidos dentro de ese sintagma.

(2) [*En vez de volver*],

- **Aplicación de modificadores:** dado que las categorías contemplan sentimientos en varios grados, este módulo se ha preparado para atender dicha enfatización. Así la repetición de letras, uso de mayúsculas, u aparición de adverbios u otros elementos potenciadores (muy, terriblemente,...) se tiene en cuenta calculando la polaridad general para el conjunto en base a unas reglas que permiten con-

siderar la aparición de varios modificadores juntos.

(3) *Terriblemente poco agraciada.*

El foco del primer adverbio es el segundo ya que se agrupa de izquierda a derecha, la función del primer modificador (en este caso, aumentativo) se le aplicará al siguiente (reductor), provocando que en su conjunción “terriblemente poco” sea un reductor de doble grado cuyo impacto es similar al de una negación.

- **Diferenciación del estilo directo e indirecto.** Es un intento de bloquear la asignación de polaridad a aquellos fragmentos que constituyan citas o sentimientos de personas ajenas al emisor del tweet. Un caso típico lo constituye el uso del estilo indirecto en la narrativa:

(4) *Ella dijo que el perro era agresivo.*

- **Calculo de la polaridad final del tweet:** permite de forma recursiva aglutinar las polaridades de niveles sintácticos inferiores y asignar el sentimiento final a todo el texto. Para ello se hace uso de las relevancias comentadas en el punto de la segmentación.

2.1 Preparación del entorno

Para poner en marcha este entorno es necesario instalar la librería Freeling que es la encargada de todo el procesamiento lingüístico. La versión instalada sobre una distribución de Linux Debian es la 3.0. El tiempo normal en dicho proceso es de 1-2 horas si no ocurre ningún percance, aunque los tiempos pueden sufrir demoras si la distribución es otra o se acusan dependencias de otras librerías.

Sobre esta versión se han realizado diversos cambios en algunos de sus diccionarios dadas las erratas o faltas halladas en las pruebas (archivo de reglas de dependencias, diccionario de locuciones, chunker,...).

3 Resultados

Aunque el entorno está preparado para comprobar individualmente el valor añadido de cada uno de sus módulos por separado, se estudia un único ejemplo (run) que mantiene activadas todas las funcionalidades arriba descritas. Los resultados se evalúan siguiendo las métricas de precisión, recall y la medida F. En el cuadro siguiente se incluyen los resultados para las dos categorizaciones (3 y 5 niveles).

	Precisión	Recall	Medida F
3 niveles	0.490	0.989	0.655
5 niveles	0.344	0.989	0.510

Tabla 1: Resultados

El entorno descartó 662 de los 60798 tweets en su ejecución. Esto es debido a la fuerte dependencia que tiene respecto a la librería Freeling. Dado que el procesamiento base es la segmentación y sobre este se asientan el resto de fases, cuando la librería no es capaz de realizarla con éxito el resultado se devuelve vacío.

Existe una disminución en el valor de las métricas en el caso de la categorización a 5 niveles. Esto puede indicar dos causas:

- El módulo de tratamiento de modificadores que es el más vinculado a esta diseminación no se adapta a los usos de enfatización empleados en el corpus o existen otros síntomas no detectados ni analizados que también afectan a la intensidad del sentimiento.
- El cómputo general de la polaridad que se obtiene agregando los sentimientos de los distintos fragmentos o segmentos es en alguna medida poco acertado y acaban disipando los sentimientos extremos.

El valor de la precisión en ambos casos responde a la falta de adecuación del sistema a fenómenos lingüísticos complejos como la ironía, extensamente empleada en la red social. Además de este uso, también se han hallado invenciones de términos obviamente no considerados parte del léxico del idioma como *descombacante* o *empleabilidad*, además de otros cuyo motivo reside en las erratas ortográficas (*recitificado* o *antiipados*). Estas palabras no pueden ser evalua-

das por el sistema por no encontrarse en ninguno de los diccionarios del entorno.

Por otro lado, se ha observado que el estilo indirecto es igualmente necesario a la hora de asignar sentimiento a un tweet. Como en [5] que afirma esta misma conclusión, asegurando que al menos en el mundo de la política, entre un 62% y un 77% de usuarios (con mayor tendencia los de *derechas*) tienen a citar comentarios de usuarios de la ideología contraria. Esta afirmación indica que aunque el texto no sea emitido por el propio tweetero, su divulgación se corresponde con un deseo de plasmar crítica o disconformidad en la mayoría de los casos.

4 Conclusiones y trabajo futuro

Existen varias líneas que son necesarias rediseñar o más bien refinar. Este refinamiento vendría dado por una extensión de los diccionarios propios empleados además del estudio y análisis de otras herramientas para la segmentación y otros procesos gramaticales. Sería positivo reducir considerablemente el número de términos no hallados en los diccionarios bien por una ampliación de la jerga o bien añadiendo una etapa de corrección ortográfica donde se pudiera aplicar por ejemplo, el algoritmo de Levenshtein y su noción de distancia entre términos.

Incluso mejorando el entorno existiría una vía que resultaría ineludible si se pretenden obtener resultados mejores. A pesar de realizar procesamientos lingüísticos óptimos, el contexto y otros condicionantes pragmáticos involucrados en la comunicación alteran los significados de los términos y en último lugar de los fragmentos. Dentro de este conjunto se encuentra el contexto así como las relaciones del escritor con el mundo que le rodea. Es decir, la asignación de una polaridad previa obtenida por un diccionario de sentimiento no debe ser un valor tan crucial y determinante. Es necesario extraer un contexto general de la situación que rodea el comentario (contexto socio-cultural) además de las tendencias del escritor. Estas últimas podrían ser recogidas a partir del estudio de sus tweets publicados que conformarían su ideología y una visión de su carácter.

También deben considerarse las herramientas disponibles de la plataforma como el uso de

hashtags que en este entorno no son tenidos en cuenta pero en ocasiones son el único síntoma de sentimiento presente en el tweet: #horror, #Tentacionesirresistibles, #malditosgenios,...

Bibliografía

- [1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. Sentiment analysis of twitter data. In Proceedings of the ACL 2011 Workshop on Languages in Social Media (2011), pp. 30–38.
- [2] Balbachan, F., Dell’Era, D. Análisis Automatizado de Sentimiento en Textos Breves de la Plataforma Twitter Resumen.
- [3] Go, A., R. Bhayani, and L. Huang. Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.
- [4] Kouloumpis, E., Wilson, T., Moore, J. Twitter Sentiment analysis: The good the bad and the OMG! In Proceedings of the ICWSM (2011).
- [5] Mullen, T., Malouf, R. A preliminary investigation into sentiment analysis of informal political discourse. In Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs, 159–162. (2006)
- [6] Taboada, M., Brooke, J., Tološki, M., Voll, K., Stede, M. Lexicon-based Methods for Sentiment Analysis. Journal of Computational Linguistics, 2010.
- [7] Pang, B., Lee, L., Vaithyanathan, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. En EMNLP-CoNLL.
- [8] Urizar, XS. Fundazioa, Elhuyar. TASS: Detecting Sentiments in Spanish Tweets.
- [9] Villena Román, J., Lana-serrano, S., Moreno García, C. DAEDALUS at RepLab 2012: Polarity Classification and Filtering on Twitter Data.
- [10] Yang, C., Bhattacharya, S., Srinivasan, P. Lexical and Machine Learning approaches toward Online Reputation Management.