TASS 2016: Workshop on Sentiment Analysis at SEPLN, septiembre 2016, pág. 23-28

Evaluación de Modelos de Representación del Texto con Vectores de Dimensión Reducida para Análisis de Sentimiento*

Evaluation of Reduced Dimension Vector Text Representation Models for Sentiment Analysis

Edgar Casasola Murillo
Universidad de Costa Rica
San José, Costa Rica
edgar.casasola@ucr.ac.cr

Gabriela Marín Raventós
Universidad de Costa Rica
San José, Costa Rica
gabriela.marin@ucr.ac.cr

Resumen: Se describe el sistema para análisis de sentimiento desarrollado por el Grupo de Análisis de Sentimiento GAS-UCR de la Universidad de Costa Rica para la tarea 1 del workshop TASS 2016. El sistema propuesto está basado en el uso de vectores de características de baja dimensión para representación del texto. Se propone un modelo simple fundamentado en la normalización de texto con identificación de marcadores de énfasis, el uso de modelos de lenguaje para representar las características locales y globales del texto, y características como emoticones y partículas de negación. Los primeros experimentos muestran las mejoras que se obtienen en la precisión al identificar la polaridad de textos completos conforme se van incorporando las características aquí mencionadas.
Palabras clave: análisis de sentimiento, clasificación de textos por polaridad, textos cortos

Abstract: The Sentiment Analisys System developed by GAS-UCR team of the University of Costa Rica for task 1 of TASS 2016 workshop is presented. Preliminar evaluation results of the proposed Sentiment Analysis System are presented. The system is based on low dimension feature vectors for text representation. The proposed model is based on text normalization with emphasis mark identification, the use of local and global language models, and other features like emoticons an negation terms. Initial experimentation shows that the introduction of the selected features have a positive impact on precision at the polarity classification task.
Keywords: sentiment analysis, polarity based text clasification, short texts.

1 Introducción

Este trabajo tiene como propósito describir el sistema utilizado por el grupo de investigación en análisis de sentimiento de la Universidad de Costa Rica en su participación en el taller TASS2016 (García-Cumbreras et al., 2016). El enfoque del trabajo del grupo ha sido el estudio de los factores que van incidiendo en las mejoras en la precisión obtenida al llevar a cabo la clasificación de la polaridad de *tweets* en idioma español. Nuestro sistema se fundamenta en tres elementos básicos que son: la normalización del texto en la etapa de preprocesamiento identificando los potenciales marcadores de énfasis presentes en el mismo, la creación de vectores de características de dimensión reducida para disminuir el efecto de la dispersión de los datos, y la exploración del impacto del uso de diccionarios de polaridad que se generan mediante la utilización de diferentes modelos de representación del lenguaje asociados tanto al contexto local como global de los datos. Para esto estamos utilizando una adaptación propia del algoritmo de Turney (Turney, 2002)sobre un corpus de 5 millones de *tweets* en español. Estos modelos se almacenan en forma de diccionarios con polaridad para su posterior reutilización. Nos interesa particularmente la investigación en este campo dado que si bien desde el año 2013 se identificó una brecha importante entre la cantidad de investigación y tecnología del lenguaje desarrollada para el

* Este trabajo se ha llevado a cabo gracias al apoyo económico de la Universidad de Costa Rica y el Gobierno de la República de Costa Rica a través del MICITT. Se agradece a los asistentes del grupo de investigación GAS-UCR por su trabajo

idioma inglés y el español (Cambria et al., 2013) (Melero et al., 2012), de la misma forma debemos tener presente que no necesariamente las soluciones para español peninsular van a tener los mismos resultados al aplicarse a variantes de español americano, por lo que los recursos y métodos que utilizamos tienen la intensión de aportar a la investigación en español y colaborar para su posterior aplicación en otros contextos de habla hispana.

2 Antecedentes

Entre los resultados obtenidos con sistemas con enfoques basados en aprendizaje máquina, el uso de **máquina de soporte vectorial (MSV)** ha ofrecido buenos resultados tanto en inglés (Kiritchenko, Zhu, y Mohammad, 2014) y (Batista y Ribeiro, 2013) como en español donde 9 de los 14 sistemas para el español presentados en TASS2015 (Villena-Román et al., 2015) hacían uso de este tipo de clasificador. Sin embargo, la dependencia del lenguaje hace que estos clasificadores dependan de los vectores de características con los que son representados los comentarios de texto. Esta extracción de características ha sido el foco de atención de múltiples trabajos como (Cabanlit y Junshean Espinosa, 2014) , (Feldman, 2013), (Guo y Wan, 2012), (Sharma y Dey, 2012) y (Wang et al., 2011). En trabajos recientes de análisis de sentimiento en español tales como el trabajo de (Martínez-Cámara et al., 2015) se utilizan varios diccionarios de polaridad y se representan utilizando un modelo de espacio vectorial MEV. El diccionario en sí se convierte en un modelo de lenguaje que sirve como recurso para lograr representaciones eficientes de los vectores utilizados para la clasificación.

En los últimos años la representación vectorial basada en modelos de lenguaje como unigramas y bigramas se movió hacia representaciones de características ya que la cantidad de términos introduce un problema asociado a su alta dispersión en el vector (Cambria et al., 2013). Si los vectores contienen un alto número de atributos diferentes, uno por término, los conjuntos de datos para entrenamiento deben contener una mayor cantidad de textos anotados que atributos para un buen entrenamiento de los clasificadores. Es por esto que los modelos de representación del lenguaje basados en unigramas, bigramas o bien skipgramas requiren de una representación vectorial eficiente. Trabajos recientes buscan la representación vectorial de las palabras en el espacio continuo como es el caso del uso de Word2Vect (Díaz-Galiano y Montejo-Ráez, 2015).

3 Descripción del sistema

Nuestro sistema se fundamenta en cuatro elementos que consideramos importantes de mencionar. Primero nos referiremos a la forma en que construimos nuestro diccionario con la polaridad de los términos y las razones para haber construido uno propio. Posteriormente nos referimos a nuestro proceso de preprocesamiento e identificación de potenciales marcadores de énfasis durante esta etapa inicial. En la siguiente subsección explicamos la forma en que construimos vectores de baja dimensión con información y hacemos uso del diccionario. Finalmente se menciona la forma en que se pretende capturar en los vectores de características aspectos locales con respecto a los datos de entrenamiento, y globales, a partir de modelos de representación del lenguaje general.

3.1 Creación del diccionario polarizado

Decidimos desarrollar diccionarios de polaridad propios, en lugar de utilizar los existentes, ya que consideramos que desde el punto de vista del procesamiento de lenguaje natural tradicional (Indurkhya y Damerau, 2010) estos diccionarios con polaridad pueden ser vistos cada uno, como un modelo de lenguaje particular. Por este motivo tratamos de desarrollar y evaluar una adaptación del tradicional método de generación de estos recursos lingüísticos de (Turney, 2002). La decisión anterior no se debió a la no existencia de diccionarios polarizados ya que claramente en trabajos como (Martínez-Cámara et al., 2015) se hace uso de varios de ellos, sino con el fin de incorporar la etapa de creación de diccionario dentro de la metodología de trabajo para que posteriores investigaciones en otros países de habla hispana puedan replicar el trabajo y disminuir la barrera inicial asociada a la falta de recursos lingüísticos propios y el efecto del uso del diccionario polarizado sobre la calidad de los resultados de clasificación.

El diccionario de polaridad creado utiliza un corpus recolectado durante el año 2013, con 5 millones de *tweets* en español. La variante con respecto al algoritmo propuesto

por Turney (Turney, 2002) es la siguiente. Para el cálculo de la **orientación semántica de un término**, tal y como lo define Turney en su artículo original, se utilizaron grupos de palabras semilla en lugar de un solo término, y en lugar de utilizar consultas a motores de búsqueda para obtener la cantidad de textos donde aparecen las palabras analizadas cerca de las palabras positivas o negativas se utilizó el motor de búsqueda implementado con el software libre Solr `http://lucene.apache.org/solr/`. Con el motor se indexaron los 5 millones de *tweets* por lo que las consultas se ejecutaron en forma local. Este método cuenta con la ventaja de que se puede calcular entonces la orientación semántica de un término directamente o bien almacenarlo en un diccionario. En nuestro caso precalculamos la polaridad y la almacenamos en forma de diccionario. Por el momento solo se han llevado a cabo los cálculos para términos individuales.

3.2 Normalizador de texto con marcadores de énfasis

Luego de un proceso de análisis de las características presentes en el texto desarrollamos un sistema para normalización del texto. Para este preprocesamiento se segmentan los términos potenciales, signos de puntuación y emoticones. Se lleva a cabo un marcado y conversión de los términos. El proceso que seguimos hace una eliminación de los términos que son identificados en el diccionario. Este proceso se muestra en la figura 1.

Las repeticiones de letras, repeticiones de sílabas y mayúsculas son identificadas y eliminadas pero estos términos se marcan como potenciales identificadores de énfasis. Ejemplos son: **EXCELENTE**, **graciassss**, **buenisísimo**. En esta fase se identifican los *tweets* que contienen palabras positivas con énfasis para su posterior uso.

3.3 Representación vectorial de baja dimensión

Dos características representadas en los vectores tienen que ver con la presencia y polaridad de los emoticones y con la presencia de partículas de negación. Además, al desarrollar esta investigación se pudo observar que los términos positivos con marcadores de énfasis son un potencial identificador de la polaridad positiva de los textos que los contienen, por lo tanto esta característica también fue incorporada. La presencia de marcadores de énfasis tales como repetición de caracteres, de sílabas, o mayúsculas sobre términos que aparecen como negativos en algún contexto son registrados como una característica importante en el vector.

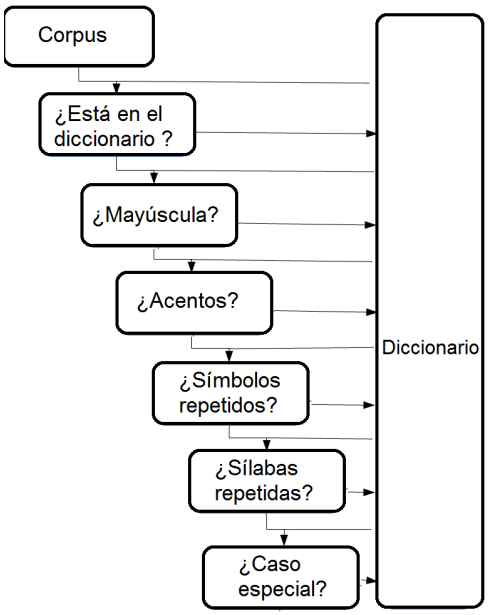

Figura 1: Proceso de normalizacion del texto

Los vectores generados utilizan la polaridad de los términos para determinar la posición en el vector de características creado. Cabe dejar claro que dependiendo del modelo de datos los términos pueden ser unigramas, bigramas o skipgramas. En el caso de los unigramas, por ejemplo, si se construye un vector con la frecuencia de los términos según su polaridad con valores de polaridad desde -1.0 hasta 1.0, el vector que se obtiene sería como el que se muestra en la figura 2. En este vector por ejemplo se muestran dos términos con polaridad, según diccionario, entre el -0.8 y -0.9, un término con polaridad entre 0.1 y 0.2, y otro con polaridad mayor a 0.9. En este caso, en nuestro diccionario, la polaridad se representa con valores distribuidos desde lo más negativo hasta lo positivo con valores entre -1.0 y 0 para los negativos y 0 a 1.0 para los positivos.

Para el taller TASS2016 quisimos evaluar inicialmente el uso de vectores con la menor dimensión posible, así que en lugar de vectores de 20 celdas utilizamos solo vectores de 5 celdas para cada grupo de características, en lugar de saltos de 0.1 el rango utilizado es de

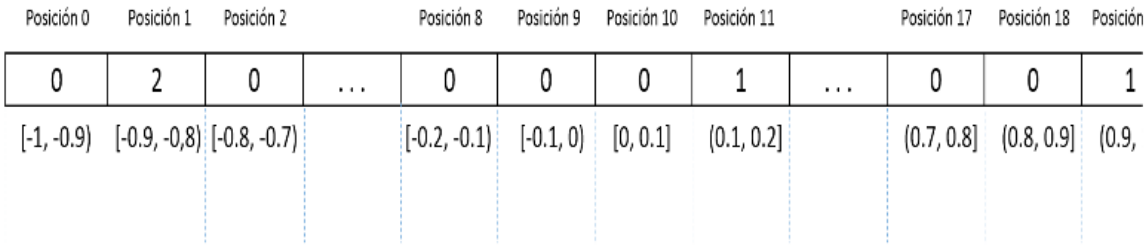

Figura 2: Vector de características

0.5.

3.4 Modelos locales y globales de representación del lenguaje

Nuestra propuesta pretende representar en los vectores de características información propia obtenida durante el proceso de entrenamiento, al igual que datos que representen información obtenida de modelos de lenguaje del español en general. En nuestro caso se utilizó inicialmente el diccionario generado a partir del corpus recolectado como insumo para obtener de él la información general del español. En el momento de entrenamiento, la polaridad de los términos en cada *tweet* son **conocidos** para ese conjunto de datos. La información global es la que se ha calculado previamente y se encuentra almacenada en forma de diccionarios. En nuestra propuesta lo que queremos hacer es representar en el vector las frecuencias de los términos de cada *tweet* distribuidos según su polaridad pero utilizar diferentes modelos de representación de lenguaje para llevar a cabo este cálculo. El diccionario utilizado en estos experimentos fue nuestra versión con unigramas. Se pretende utilizar representaciones con bigramas y una versión de skipgramas que incluye solo los términos anteriores a la palabra que se desea representar. Durante el entrenamiento, la polaridad obtenida en forma local es almacenada al igual que las frecuencias tomadas de diccionarios de polaridad global. Por lo tanto, los vectores cuentan con entradas para las distribuciones de polaridad local y las distribuciones de polaridad global. Aquí es donde incorporamos los diferentes modelos de lenguaje. Inicialmente trabajamos con **unigramas** para obtener resultados base para posteriores experimentos. Posteriormente, se genera un diccionario para **bigramas** y otro para lo que definimos como **skip-gramas previos**. Por el momento estas variantes no fueron enviadas como experimentos a TASS2016 sino solo las versiones iniciales.

4 Metodología

Utilizando el diccionario, el normalizador y el modelo de representación vectorial se procedió a crear vectores de respresentación con diferentes configuraciones. Primeramente se construyó una versión con vectores de dimensión 20 distribuyendo la polaridad de los términos según la polaridad almacenada para unigramas en el diccionario local. En este caso se pretende evaluar solamente el uso del diccionario y los marcadores de énfasis como repeticiones y mayúsculas. Este primer experimento es el denominado GASUCR-01. El segundo experimento consistió en evaluar un modelo un poco más robusto a nivel local con bigramas y la polaridad para el unigrama en el diccionario, si el bigrama no está presente durante el proceso de evaluación. En este caso se crearon vectores de menor dimensión para los datos locales, con solo cinco campos. Esta ejecución se idendificó como experimento GASUCR-01-noEMO-noPartNeg. Esta es la implementación base para luego evaluar el uso de bigramas tomados del contexto global. Esta versión base también fue enviada a la tarea de 4 categorías. En este caso, lo que se hizo fue unir las categorías +P y P en una sola, y la categoría +N con la N. El tercer experimento agregaba al anterior el uso de los emoticones, aparición de términos positivos con énfasis y las partículas negativas. En los resultados esta versión se identificó como GASUCR-04 En esta versión de TASS no nos dió tiempo de ejecutar las versiones con bigramas globales, ni skipgramas.

5 Resultados

Los resultados oficiales obtenidos para las ejecuciones antes mencionadas son los que se muestran en las Tablas 1 y 2. En estas figuras la columna **Ac.** muestra la *exactitud*, **P** se refiere a la **Macro Precisión**, **R** al **Macro Exhaustividad** y **F1** al **Macro F1**. En los resultados generales de TASS los resultados del grupo aparecen con el id indicado bajo el nombre del grupo GASUCR. En nuestro caso con el experimento 01 obtenemos los casos base para el uso de unigramas globales con vectores de dimensión 20 y los bigramas locales con dimensión 5. Es importante observar que los bigramas locales con dimensión 5 y las características de énfasis positivo, partículas de negación y emoticones producen un leve incremento pasando de 0.32 a 0.41. Otro aspecto que rescatamos es el aumento de la exactitud al pasar a la tarea de 3 categorías.

Tabla 1: Resultados Tarea 1 con 5 levels y corpus completo)

id	*Ac.*	*P*	*R*	*F1*
01	0.342	0.217	0.237	0.227
01-noEmNeg	0.326	0.334	0.258	0.291
04	0.410	0.268	0.242	0.254

Tabla 2: Resultados Tarea 1 con 3 niveles y corpus completo

id	*Ac.*	*P*	*R*	*F1*
01-noEmNeg	0.373	0.212	0.303	0.250

Estos casos se fueron seleccionando para ir evaluando en forma incremental cada uno de los aspectos relacionados a nuestra propuesta. Con cada característica nueva se trata de determinar su impacto sobre los valores de exactitud, precisión y exhaustividad.

6 Conclusiones y trabajo futuro

El marco de evaluación de TASS es provechoso para los grupos que inician la investigación en análisis de sentimiento en español con el fin de extenderla a otras latitudes. En nuestro caso pudimos evaluar y comparar la calidad de los resultados de los primeros casos base de nuestro trabajo. Observamos los primeros resultados con un sistema que utiliza un método de normalización con identificación de potenciales marcadores de énfasis, un modelo de representación basado en vectores de baja dimensión, y modelos de representación del texto con características locales y globales. El trabajo además hace uso de características comunes con otros como los son el uso de emoticones y partículas negativas. Como trabajo futuro tenemos pendiente la evaluación usando 3 categorías de los datos que hacen uso de contexto local con bigramas y características adicionales como uso de emoticones, palabras positivas con énfasis, y partículas de negación. Esperamos que los mejores resultados sean obtenidos al incorporar los nuevos modelos de lenguaje que estamos calculando para bigramas y skipgramas previos al unirlo con nuestro método de representación en vectores de baja dimensión. Se desea estudiar el efecto de la reducción del tamaño del vector al igual que técnicas de extrapolación de la polaridad en los modelos para los términos que no aparecen en los datos de entrenamiento.

Bibliografía

Batista, F. y R. Ribeiro. 2013. Sentiment analysis and topic classification based on binary maximum entropy classifiers. *Procesamiento de Lenguaje Natural*, 50:77–84.

Cabanlit, M. A. y K. Junshean Espinosa. 2014. Optimizing n-gram based text feature selection in sentiment analysis for commercial products in twitter through polarity lexicons. En *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on*, páginas 94–97. IEEE.

Cambria, E., B. Schuller, Y. Xia, y C. Havasi. 2013. New avenues in opinion mining and sentiment analysis. *Intelligent Systems, IEEE*, PP(99):1–1.

Díaz-Galiano, M. y A. Montejo-Ráez. 2015. Participación de sinai dw2vec en tass 2015. En *Proceedings del Taller TASS 2015 en Análisis de Sentimiento de la XXXI Conferencia SEPLN 2015*, páginas 59–64.

Feldman, R. 2013. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89, Abril.

García-Cumbreras, M., J. Villena-Román, E. Martínez Cámara, M. C. Díaz-Galiano, M. T. Martín Valdivia, y L. A. Ureña López. 2016. Overview of

tass 2016. En *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016)*, Salamanca, Spain, September.

Guo, L. y X. Wan. 2012. Exploiting syntactic and semantic relationships between terms for opinion retrieval. *Journal of the american society for information science and technology*, 63(11):2269–2282, Noviembre.

Indurkhya, N. y F. J. Damerau. 2010. *Handbook of natural language processing*, volumen 2. CRC Press.

Kiritchenko, S., X. Zhu, y S. M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, páginas 723–762.

Martínez-Cámara, E., M. Á. García-Cumbreras, M. T. Martín-Valdivia, y L. A. Ureña-L'opez. 2015. Sinai-emma: Vectores de palabras para el análisis de opiniones en twitter. En *Proceedings del Taller TASS 2015 en Análisis de Sentimiento de la XXXI Conferencia SEPLN 2015*, páginas 41–46.

Melero, M., A.-B. Cardús, A. Moreno, G. Rehm, K. de Smedt, y H. Uszkoreit. 2012. *The Spanish language in the digital age*. Springer.

Sharma, A. y S. Dey. 2012. A comparative study of feature selection and machine learning techniques for sentiment analysis. En *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, páginas 1–7. ACM.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th annual meeting on association for computational linguistics*, páginas 417–424. Association for Computational Linguistics.

Villena-Román, J., J. García Morera, M. Á. García-Cumbreras, E. M. Cámara, M. T. M. Valdivia, y L. A. U. López. 2015. Overview of tass 2015. En *Proceedings del Taller TASS 2015 en Análisis de Sentimiento de la XXXI Conferencia SEPLN 2015*, páginas 13–21.

Wang, X., F. Wei, X. Liu, M. Zhou, y M. Zhang. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. En *Proceedings of the 20th ACM international conference on Information and knowledge management*, páginas 1031–1040. ACM.