Análisis de sentimientos basado en aspectos: un modelo para identificar la polaridad de críticas de usuario

Miguel Angel Rosales Quiroga, Darnes Vilariño Ayala, David Pinto, Mireya Tovar, Beatriz Beltrán

Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science, Puebla,
México

miguelrosales@gmail.com,
{dvilarinoayala,davideduardopinto,mireyatovar,bbeltranmtz}@gmail.com
http://www.lke.buap.mx/

Resumen. Con el crecimiento de los usuarios de internet ha aumentado la cantidad de datos generados en la red por lo que se hace importante desarrollar modelos que permitan obtener información importante a partir de dichos datos. La propuesta presentada en este artículo pretende detectar la polaridad de enunciados, párrafos o fragmentos de texto, mencionadas en reseñas de usuarios. El objetivo es identificar los aspectos y el sentimiento expresado para cada aspecto. Se plantea la creación de un modelo no supervisado para la identificación de las características léxicas sintácticas para la detección de aspectos y la clasificación del sentimiento expresado en las reseñas en tres categorías: positivo, negativo y neutro. Se analizaron características como el etiquetado gramatical de las palabras, la similitud semántica entre palabras, la co-ocurrencia de palabras en un conjunto de documentos.

Palabras clave: Análisis de sentimientos, análisis basado en aspectos, polaridad.

Aspect Based Sentiment Analysis: A Model for Identification of User Reviews Polarity

Abstract. The continued growth of users has generated a huge number of data in Internet, thus leading to the importance of developing new models for obtaining relevant information from this data. In this paper we aim to detect the polarity of user reviews. The purpose of this paper is to identify aspects and sentiments expressed for each aspect. We propose an unsupervised model for the identification of lexical and syntactic features for the detection of aspects and the classification of the sentiment expressed in user reviews into three features: positive,

negative and neutral. Some features as the following ones are employed: part-of-speech tags, semantic similarity among words and co-ocurrence.

Keywords: Sentiment analysis, aspect-based analysis, polarity.

1. Introducción

El Análisis de Sentimientos es una tarea de clasificación de textos dentro del área del Procesamiento del Lenguaje Natural, su objetivo es dado una opinión de usuario poder detectar la polaridad de ésta, ya sea positiva, negativa o neutra. El conocer la opinión que una persona tiene hacia un producto o servicio, es de gran ayuda para la toma de decisiones, ya que permite a otros posibles consumidores detectar la calidad del producto o servicio evaluado para utilizarlo. En este artículo se plantea un modelo no supervisado para resolver el problema del análisis de opiniones basado en aspectos. Para la detección de aspectos y la polaridad de las opiniones se realizan el análisis léxico de las reseñas, así como también la semántica de las mismas. Se trabajó con un conjunto de datos de entrenamiento conformado por reseñas sobre Restaurantes y Laptops, éstas reseñas fueron proporcionadas en el marco del Semeval 2016.

Uno de los primeras investigaciones sobre el análisis de sentimientos fue presentada por Pang y Lee [11]. Publicaron su trabajo sobre la clasificación de documentos en base al sentimiento expresado en éstos. Analizaron reseñas sobre películas, encontraron que las técnicas de Aprendizaje Automático mejoran el rendimiento de las líneas base generadas por los expertos humanos. Emplearon tres algoritmos de Aprendizaje Automático: Naive Bayes (NB), Máxima Entropía (ME) y Máquinas de Soporte Vectorial (SVM).

Uno de los primeros trabajos que introdujeron el término de Análisis de Sentimientos fue el presentado por Nasukawa y Yi [10]. En esta publicación definen esta tarea como encontrar expresiones de sentimientos para un sujeto dado y determinar la polaridad de los mismos. En las investigaciones anteriores a ésta, se realizaba el análisis de la polaridad general de un documento, sin embargo en este enfoque se trata de identificar la opinión de cada sujeto mencionado en el texto.

Minqing Hu y Bing Liu [4], expusieron una propuesta para minar y resumir reseñas de consumidores. Los objetivos de este trabajo fueron encontrar las características a las cuales se hacían referencia en las críticas, identificar los enunciados que expresaban opiniones y polaridad sobre las mismas y resumir los resultados. Se propuso la creación de una pequeña lista de adjetivos "semilla" etiquetados manualmente dependiendo si expresan sentimiento positivo o negativo. Posteriormente esta lista es aumentada usando WordNet [8]. Para la detección de los aspectos se emplearon características de etiquetado de las partes del enunciado (Part of speech tagging). Se identifican las características frecuentes, aunque solo se analizan las que se presentan de manera explícita en los enunciados. A continuación, se realiza una extracción de palabras que expresan opinión, se tiene preferencia por los adjetivos cercanos a los aspectos

para tener enunciados de opinión. La identificación de la orientación de estas palabras se realiza mediante un análisis de sinónimos y antónimos.

Una de las propuestas más interesantes presentadas en el SemEval 2014 y que además ha obtenido los mejores resultados en esta tarea es la presentada por Kiritchenko, Zhu, Cherry y Mohammad [5], presentan técnicas como la creación de diccionarios para la detección de sentimientos, estos creados automáticamente utilizando fórmulas que analizan la información mutua [6]. Las palabras de negación son analizadas en un contexto diferente y se crearon diccionarios para estos casos. Adicional a estos se implementaron diccionarios sobre el dominio de laptops y restaurantes.

Otro trabajo que es importante destacar para este Foro de Competición es el desarrollado por Pavel Blinov y Eugeny Kotelnikov [3], que proponen un método para el Análisis de Sentimientos basado en Aspectos para un conjunto de opiniones sobre laptops y restaurantes. El método propuesto para la extracción de aspectos consiste en dos pasos: la selección de candidatos y la extracción de términos.

En la investigación desarrollada por Schouten, Frasincar y De Jong [13], presentan un enfoque basado en co-ocurrencias para la detección de categorías y uno basado en diccionarios de sentimientos para la clasificación. En particular los datos utilizados son los presentados en el Semeval 2016. A continuación se disctute me metodología propuesta.

2. Metodología

Para darle solución a este problema se propone un modelo compuesto por tres fases:

1. Fase de Pre-procesamiento.
2. Fase de Identificación de Aspectos.
3. Fase de Identificación de Polaridad.

Fase de preprocesamiento Durante esta fase se realiza el análisis de los datos de entrenamiento proporcionados por el SemEval para la tarea, además de reunir los datos y generar los archivos de entrada necesarios para el modelo propuesto. Los datos de entrenamiento se encuentran en formato XML. Estos datos contienen el conjunto de reseñas con las categorías correctas, el objeto al cual se hace referencia en la reseña y la polaridad expresada hacia éste. Para el dominio de Laptops sólo se proporciona la categoría correcta identificada en la reseña.

Del conjunto de datos de entrenamiento se obtienen varios elementos de entrada. Primero, se genera un diccionario con los aspectos que se encontraron en esas reseñas. Después, cada reseña del conjunto de datos de entrenamiento es tratada mediante la herramienta Clips Pattern[1] para obtener su etiqueta

[1] http://www.clips.ua.ac.be/

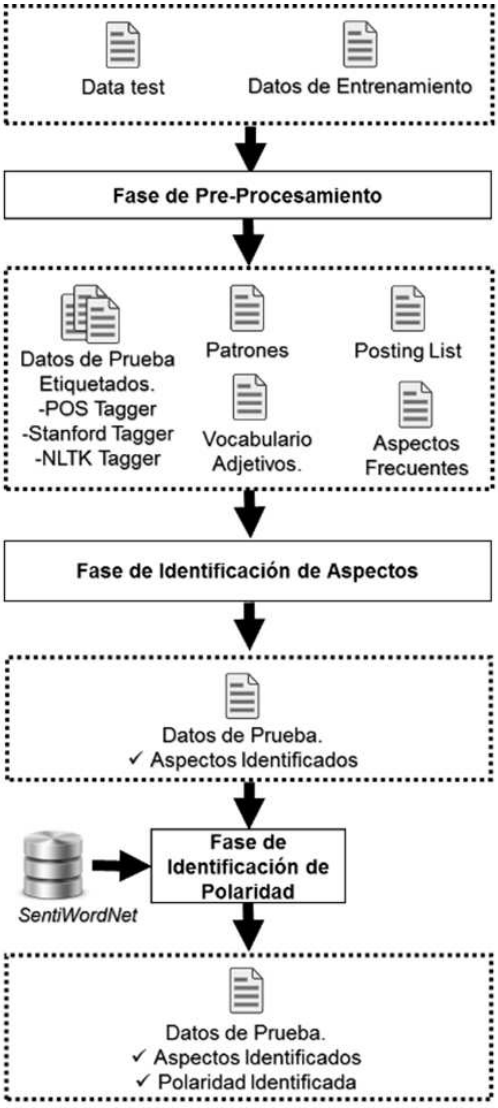

Fig. 1. Metodología basada en tres fases

POS (Part of Speech). Se extraen los patrones de cada aspecto y para cada aspecto se obtienen sus elementos gramaticales adyacentes. Esto para generar un diccionario con los patrones que cumple un aspecto, además de conocer que elementos gramaticales se encuentran generalmente junto a éste.

Una vez que se tienen el conjunto de aspectos de los datos de entrenamiento se genera un Crawler con estos términos para la generación de dos corpus con datos específicos relacionados a Laptops y Restaurantes. Este consiste en para cada

aspecto encontrar documentos relacionados con el mismo en artículos de internet. Es decir, para cada aspecto son extraídos un conjunto de párrafos relacionados. Para su implementación se utilizaron las bibliotecas de BeautifulSoup [9], que permite acceder a páginas web y obtener su contenido.

A continuación, teniendo estos corpus específicos se generó un Posting List para cada corpus. Esto consiste que para cada palabra se genera una lista indexada con el identificador del párrafo que la contiene. Por ejemplo, para el aspecto "Food" se genera una lista con el índice de los párrafos en los que aparece. Con esto, se generaron dos Posting List con 554,356 y 463,924 palabras para Restaurantes y Laptops respectivamente. Finalmente se genera un diccionario con los adjetivos, adverbios y verbos incluidos en los datos de entrenamiento. En este caso con su relación con la polaridad identificada. Como entrada para la siguiente fase, también se genera el etiquetado gramatical del conjunto de datos de prueba (Data test) con los tres etiquetadores propuestos (Stanford Parser [7], Clips Pattern y NLTK [2]).

Fase de identificación de aspectos En esta fase se realiza la detección de los aspectos mencionados en cada reseña. Primero, para cada crítica, se extraen los sustantivos o secuencia de sustantivos y son agregados a un diccionario de candidatos a aspectos indicado con un peso que cumple con esta característica. Posteriormente, bajo la hipótesis de que dos palabras que se encuentran en el mismo párrafo están relacionadas, se realiza una búsqueda de cada uno de los elementos que se encuentran en el diccionario de candidatos en el Posting List generado en la fase anterior. Una vez que es realizada la búsqueda, cada candidato tiene un conjunto de párrafos en los que se encuentra. A continuación se realiza una intersección de cada uno de los candidatos con el resto, y aquellos que se encuentran relacionados con al menos la mitad de candidatos más uno, se les aumenta un valor en su peso, para así indicar que cumplen con esta segunda característica. Otra característica más es el análisis de similitud semántica mediante la biblioteca gensim y el módulo Word2Vec [12]. Esta herramienta necesita un corpus de tamaño considerable para su funcionamiento. El corpus proporcionado a la herramienta fue el corpus generado mediante la implementación del Crawler en la fase anterior. De manera similar a lo realizado con el Posting List, cada candidato es comparado con el resto de elementos y aquellos candidatos que estén relacionados con los demás se aumentan su valor de peso. Además cada candidato es comparado también con el conjunto de entidades predefinidas por el SemEval. Cuando la medida de similitud encontrada por el modelo generado por Word2Vec entre el candidato y alguna entidad es grande, el valor de peso del candidato es aumentado, esto ya que si es muy similar a una entidad, lo más probable es que sea un aspecto.

La siguiente característica analizada son los patrones encontrados en la primera fase. Estos patrones están conformados por las secuencias de etiquetas de POS (Part of Speech) de los aspectos en los datos de entrenamiento. Para obtener estos patrones cada reseña de los datos de entrenamiento es procesada mediante la herramienta de CLiPS para obtener la etiqueta POS (Part of Speech) de cada

palabra. Posteriormente se forman los patrones, estas son las secuencias de la forma POS_Izquierda + POS_Aspecto + POS_Derecha. Donde POS_Aspecto es la etiqueta o secuencia de etiquetas gramaticales del aspecto; POS_Izquierda y POS_Derecha son las etiquetas gramaticales de la palabra izquierda y derecha al aspecto respectivamente. Para cada reseña, se realiza la búsqueda de los posibles n-gramas que cumplan con el patrón de aspecto detectado. Una vez identificados estos patrones, también son estudiados sus elementos adyacentes izquierdo y derecho. Si cumplen con alguno de los patrones identificados en la fase de pre-procesamiento son agregados al diccionario de candidatos, si este elemento ya se encuentra, su valor de peso es aumentado. Finalmente, la última característica tomada en cuenta es la búsqueda de los candidatos encontrados hasta este momento en la lista de aspectos, del conjunto de datos de entrenamiento. Si es encontrado el candidato en esta lista se aumenta su valor de peso. Ya realizado el análisis de las características mencionadas (Identificación de sustantivos, análisis de contexto, identificación de patrones, identificación de aspectos frecuentes), el criterio de selección de los candidatos es que cumplan con tener un peso mayor o igual a 3, es decir, cumplen con 3 o más características de las mencionadas anteriormente. La metodologìa propuesta se puede ver en la figura:

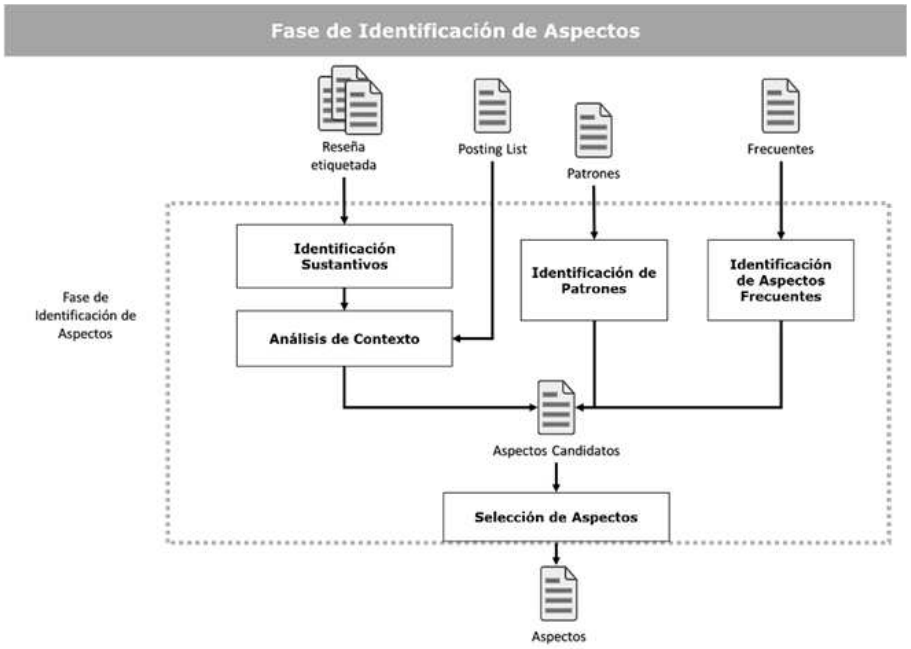

Fig. 2. Metodología basada en tres fases

2.1. Fase de identificación de polaridad

Teniendo ya los candidatos propuestos, las entidades y atributos identificados, la última parte consiste en identificar la polaridad del sentimiento expresado sobre cada aspecto. Para esto, se utilizó un enfoque basado en diccionario, para tener el valor de sentimiento para cada palabra. En esta propuesta se utilizó un diccionario creado en base a los datos de entrenamiento. Se creó mediante el procesamiento de adjetivos, adverbios y verbos en cada enunciado. Los datos de entrenamiento se encuentran etiquetados con su polaridad correcta, por tanto, para cada elemento encontrado en la sentencia se le da un valor de sentimiento en un rango de -1 a 1, donde -1 es muy negativo y 1 es muy positivo. Para el cálculo de este valor se realiza la división entre el número de ocasiones que aparece el elemento para cada polaridad, positiva, negativa y neutra. De estos resultados, el que tenga mayor valor es elegido como la polaridad predominante de esta palabra y el resultado de la división es asignado como valor de sentimiento. Para aquellas palabras que no se encuentren en el conjunto de palabras encontradas en los datos de entrenamiento se utilizó el diccionario SentiWordNet [1] que proporciona un valor numérico de sentimiento para cada palabra.

Posteriormente para detectar la polaridad de cada aspecto encontrado, se realiza el promedio de las polaridades de las palabras de la frase donde el aspecto se encuentra. Se analizan palabras que invierten el valor de polaridad como son “NOT” Cuando este tipo de palabras aparecen, el valor de polaridad de las siguientes palabras de la sentencia es invertido. Palabras como “TOO”, “VERY” entre otras también causan un efecto en las palabras, estas aumentan el valor de polaridad de los siguientes elementos de la sentencia. Al finalizar, si el valor promedio encontrado es positivo y mayor a un rango establecido, la sentencia es clasificada como positiva. De lo contrario si es negativo y menor al rango, es clasificada como negativa. Si el valor promedio es igual a cero o si está dentro del rango establecido es marcada como neutra. El rango mencionado se establece de manera manual, en donde el valor de polaridad identificado es mínimo, lo que implica que el sentimiento expresado sobre un aspecto no es relevante para ser clasificado como positivo o negativo. En adición al promedio de las polaridades, se estudia la polaridad individual de las palabras adyacentes al aspecto identificado. Se analiza el valor de polaridad más alto, además del total de palabras positivas, negativas y neutras. La salida generada por esta fase es el conjunto de aspectos, entidades y atributos con su polaridad identificada. Es to puede observarse en la figura

3. Resultados obtenidos

Una vez desarrollado el modelo, se generaron las salidas necesarias para participar en el Foro de Competencia del SemEval 2016. Para la tarea de la identificación de aspectos, se obtuvo un 50.25

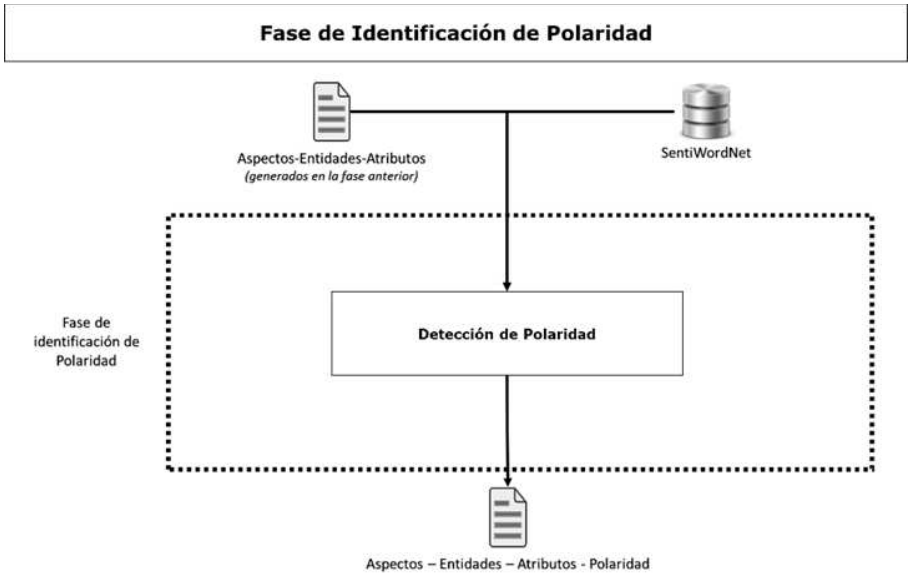

Fig. 3. Metodología basada en tres fases

Tabla 1. Comparación de resultados obtenidos en la identificación de aspectos: Dominio de Restaurantes

Propuestas	**Exactitud**
Mejor propuesta	72.34 %
Nuestra propuesta	50.25 %
Base line	44.07 %

4. Conclusiones obtenidas y trabajo a futuro

Evaluado el modelo propuesto para resolver la tarea del análisis de sentimientos basado en aspectos propuesto por el SemEval, se llegó a las siguientes conclusiones. Se diseñó e implementó un modelo para detectar los aspectos en las reseñas proporcionadas y se detectó la polaridad del sentimiento expresado en la reseña para cada aspecto. Con la implementación del Crawler con los aspectos de entrenamiento en la fase de pre-procesamiento se generaron dos corpus específicos, uno con información relacionada con restaurantes, y otro con datos sobre laptops. Estos corpus son una aportación de gran utilidad para futuras tareas que necesiten trabajar bajo estos contextos o similares. Una vez analizados los resultados obtenidos por el modelo se llegaron a las siguientes conclusiones:

- El modelo se comporta bien en la detección de aspectos, sería de gran utilidad generar un diccionario para mejorar la identificación de aspectos. Este diccionario podría incluir, por ejemplo, nombres de platillos para el caso

Tabla 2. Comparación de resultados obtenidos en la identificación de polaridad: Dominio de Restaurantes

Propuestas	**Exactitud**
Mejor propuesta	88.12 %
Nuestra propuesta	60.88 %
Base line	76.48 %

Tabla 3. Comparación de resultados obtenidos en la identificación de polaridad: Dominio de Laptops

Propuestas	**Exactitud**
Mejor propuesta	82.72 %
Nuestra propuesta	62.79 %
Base line	70.03 %

de restaurantes o nombres de aplicaciones y componentes para el dominio de laptops. Esto debido a que los etiquetadores gramaticales son etiquetadores generales y en ocasiones existen palabras de los contextos de (restaurantes y laptops) que no son etiquetados correctamente.

- Dado que los diccionarios de sentimiento como el utilizado SentiWordNet son diccionarios generales, es necesario generar un diccionario de sentimientos específico para cada contexto, ya que existen palabras como "hot" o "cold" que bajo el contexto general tienen una polaridad y bajo el contexto de comida y restaurantes tienen otro valor totalmente opuesto.
- Analizando los resultados obtenidos al utilizar las funciones brindadas por CLiPS, se concluye que esta herramienta es de gran utilidad, aunque mejora los resultados al apoyarse de otros etiquetadores como los utilizados en el modelo propuesto.
- Se debe mejorar en la tarea de la detección de la polaridad. Analizar problemáticas como el uso del sarcásmo en las críticas que invierten la polaridad del sentimiento expresado y estudiar la manera de mejorar los resultados del modelo.
- La inclusión de la herramienta Word2Vec fue de gran utilidad para mejorar los resultados, ya que proporciona una medida de similitud entre palabras basadas en un contexto dado. Esto mediante el análisis de un corpus especifico. Para mejorar los resultados se podría aumentar el tamaño del corpus de entrada. Finalmente, se pretende continuar con el trabajo en este modelo para futuras participaciones en las tareas del SemEval. Como trabajo futuro se propone lo siguiente:
- La creación de un diccionario de platillos para el dominio de restaurantes y de aplicaciones y componentes para el dominio de laptops.
- La inclusión de los datos de entrenamiento y pruebas del SemEval 2015 y 2016 para mejorar los diccionarios para la identificación de atributos y entidades y la extracción de aspectos.
- Implementar nuevas características que permitan obtener mejores resultados. Estas pueden ser el uso de aprendizaje automático y similitud entre frases.

Agradecimientos. El trabajo realizado con apoyo parcial de CONACYT.

Referencias

1. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proc. of LREC (2010)
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Inc., 1st edn. (2009)
3. Blinov, P., Kotelnikov, E.: Blinov: Distributed representations of words for aspect-based sentiment analysis at semeval 2014. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 140–144. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (August 2014)
4. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 168–177. KDD '04, ACM, New York, NY, USA (2004)
5. Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, S.: Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 437–442. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (August 2014), http://www.aclweb.org/anthology/S14-2076
6. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA (1999)
7. de Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proc.of LREC. pp. 449–454 (2006)
8. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Wordnet: An on-line lexical database. International Journal of Lexicography 3, 235–244 (1990)
9. Nair, V.G.: Getting Started with Beautiful Soup. Packt Publishing (2014)
10. Nasukawa, T., Yi, J.: Sentiment analysis: Capturing favorability using natural language processing. In: Proceedings of the 2Nd International Conference on Knowledge Capture. pp. 70–77. K-CAP '03, ACM, New York, NY, USA (2003)
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. pp. 79–86. EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002)
12. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010)
13. Schouten, K., Frasincar, F., de Jong, F.: Commit-p1wp3: A co-occurrence based approach to aspect-level sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 203–207. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (August 2014)