UNIVERSITAT JAUME I
Departament de Llenguatges i Sistemes Informàtics

Modeling and analyzing opinions from customer reviews

Ph. D. dissertation
MSc. Lisette García Moya

Supervisor
Dr. Rafael Berlanga Llavori

Castellón, November 2015

To my parents.
To Henry, Sandra and Eric.

Acknowledgements

This thesis would not be possible without the support and help that several people provided me during the last years.

First of all, I would like to express my deepest gratitude to Rafael Berlanga and María José Aramburu, who have helped me a lot, not only on the academic but also on the personal whenever I needed it. They have offered me their unconditional friendship. I will never thank them enough.

Also, it has been a pleasure for me to share these years with my colleagues at TKBG. Thanks you all for letting me be part of this incredible group.

I would also like to specially acknowledge Aurora Pons Porrata (RIP) for introducing me into the research world.

Last but not least, I would like to thanks my family. To my parents, for their unmeasurable love and support during my whole life. To Sandra and Eric, who inspire me to be better every day. To Henry, thanks for being always there.

This work has been mainly supported by the PhD Fellowship Program of the Universitat Jaume I (PREDOC/2009/12), and by the research projects from the Ministerio de Economía y Competitividad (TIN2008-01825, TIN2011-24147) and Fundació Caixa Castelló P1-1B2008-43 and P1-1B2010-49.

Abstract

The main motivation behind this thesis is the problem of aspect-based sentiment summarization and its application to Business Intelligence (BI). Given a collection of opinion posts, aspect-based summarization has to do with extracting from the collection the most relevant opined aspects (also called features) along with their associated sentiment information (usually an opinion word and/or a polarity score that express the sentiment orientation of the opinion).

In the recent scenario of e-commerce, we presume that BI could rely on extracted knowledge from reviews available in the Web in order to analyze recent trends as well as the satisfaction and behavior of customers and to prepare strategic plans accordingly.

Specifically, this thesis proposes new methodologies to:

- model and extract the opinions and their respective targets (i.e., aspects or features) from collections of opinion posts, and
- integrate the extracted sentiment data into a traditional corporate data warehouse to enable BI.

The modeling of opinions and their targets takes place in the general framework of statistical language modeling. The hypothesis is that there exists a language model of opinion words able to model the opinion lexicon of a domain, and that there is also a language model of aspects that can be learned from the model of opinions.

Both the learning of the models and the extraction of the sentiment data (i.e., the tuples feature-opinion) are implemented using unsupervised approaches that do not need exhaustive natural language processing (except for POS-tagging/ lemmatization). The resulting methodologies can be applied to any language and domain given a seed set of general-domain opinion words.

For the integration of sentiment data with traditional corporate data two scenarios are considered: a static one in which both the data sources and the user requirements are static and known in advance, and dynamic one based on an open data infrastructure where BI data can be linked to external sources on demand, without being attached to predefined (rigid) data structures or multidimensional schemas.

We demonstrate our proposal on datasets of real opinions available in the Web. Results of the proposed method corroborate the thesis claims and show a good effectivity for their usage as a BI analysis tool.

Resumen, principales contribuciones y resultados

La principal motivación de esta tesis es el problema de la construcción de sumarios de opiniones basados en aspectos y la aplicación de éstos a la Inteligencia de Negocios (BI, por sus siglas en inglés). Dada una colección de opiniones sobre un producto o servicio, el problema de la construcción de sumarios de opiniones se centra en la extracción de los aspectos (o características) más relevantes sobre los que se emite alguna opinión en la colección, además de la opinión o información subjetiva asociada (usualmente una palabra de opinión que puede estar acompañada de la polaridad que expresa la orientación positiva o negativa de la opinión).

En el más amplio escenario actual de comercio electrónico, asumimos que BI puede utilizar la información extraída de los comentarios disponibles en la Web con el objetivo de analizar las tendencias recientes, así como la satisfacción y el comportamiento de los clientes, para, en consecuencia, preparar planes estratégicos.

Específicamente, esta tesis propone nuevas metodologías para:

- modelar y extraer las opiniones y sus respectivos objetivos (es decir, los aspectos o características de los que se opina) de las colecciones de opiniones, además de

- integrar los datos de opinión extraídos en un data warehouse corporativo para brindar soporte a BI.

En ellas, el modelado tanto de las opiniones como de los aspectos de opinión se basa en el marco general de modelos estadísticos de lenguajes. La hipótesis es que existe un modelo de lenguaje capaz de modelar el lenguaje de las palabras de opinión de un dominio, y que, además, existe un modelo del lenguaje de aspectos que se puede aprender a partir del modelo de opiniones.

El aprendizaje de los modelos y la extracción de los datos de opinión (es decir, de tuplas de la forma característica-opinión) se implementan usando enfoques no supervisados, que no necesitan de un procesamiento exhaustivo del lenguaje natural (a excepción de POS-tagging y lematización). Las metodologías resultantes se pueden aplicar a cualquier idioma y dominio dado un conjunto de semillas de palabras de opinión de dominio general.

Para la integración de los datos de sentimiento con los datos corporativos tradicionales se consideran dos escenarios: uno estático en el cual tanto las fuentes de datos y las necesidades de los usuarios son conocidas de antemano,

y uno dinámico basado en una infraestructura abierta de datos donde los datos de BI pueden enlazarse a fuentes externas, sin estar ligadas indisolublemente a estructuras de datos predefinidas o esquemas multidimensionales rígidos.

Se realiza una demostración de la propuesta sobre conjuntos de datos de opiniones reales que se encuentran disponibles en la Web. Los resultados experimentales obtenidos corroboran las hipótesis de partida de la tesis, además de mostrar la efectividad los métodos propuestos.

Principales contribuciones

Las principales contribuciones de la tesis son las siguientes:

1. Partiendo de la idea de que las características de un producto pueden ser modeladas por medio de un modelo estádistico de lenguaje, se propone una metodología preliminar para modelar el lenguaje de las características de un producto a partir de un conjunto de palabras de opinión. La principal novedad de la metodología es que se basa en un modelo estocástico de correspondencia entre las palabras que tiene como objetivo capturar la vinculación latente entre las opiniones y las características que éstas modifican en los textos. La metodología propuesta también es capaz de obtener un ránking de palabras de opinión a partir de una colección de opiniones

2. A partir de la metodología preliminar, y teniendo en cuenta nuestras dos principales hipótesis, se propone una nueva metodología independiente del dominio capaz de modelar y extraer de un conjunto de opiniones sobre un producto o servicio aquellas características del producto de las cuales se opina, así como las opiniones correspondientes.

3. Como parte de la metodología, se introduce un nuevo método para inducir un modelo de opiniones del producto que se basa en una función kernel entre distribuciones de palabras. El modelo de opiniones se emplea en el aprendizaje de un modelo refinado del lenguaje de las características del producto, a partir del cual se implementa un método para recuperar tanto las características como sus correspondientes opiniones.

4. Tomando como base la metodología para modelar y extraer las características de un producto, se introduce una nueva metodología para integrar los datos de opinión de los textos en un modelo de BI implementado en un almacén de datos corporativo.

5. Para llevar a cabo la integración, se propone un nuevo método de anotación semántica –totalmente automático y que no requiere supervisión– para unificar características en el proceso de poblado del almacén de datos. Además, se propone un método de estimación de la polaridad de las opiniones.

6. Porúltimo, se identifican nuevos problemas en la integración que tienen en cuenta principalmente el caracter dinámico de los datos. Estos problemas son tenidos en cuenta en el desarrollo de una novedosa infraestruc-

tura semántica de datos para BI llamada SLOD-BI. SLOD-BI se basa en los principios de la iniciativa *Linked Open Data* (LOD).

Es importante mencionar que las metodologías propuestas en este trabajo tienen un buen rendimiento en la recuperación de los datos de opinión a pesar de no requerir de un preprocesamiento profundo del lenguaje de los textos (principalmente se realiza un análisis morfológico de los mismos). Además, las metodologías propuestas se pueden aplicar a cualquier idioma y dominio dado un conjunto de semillas de palabras de opinión de dominio general del lenguaje como entrada.

Resultados

Esta tesis se conforma como un compendio de publicaciones. El conjunto de los resultados alcanzados en cada una de ellas corroboran las hipótesis que se plantean en esta tesis (ver sección 1.3). Además, los resultados permiten dar respuesta a las siguientes preguntas de investigación:

RQ1 *¿Existe algún modelo estadístico de lenguaje capaz de modelar un lexicón de palabras de opinión para un dominio o producto a partir de un conjunto de semillas de palabras de opinión de dominio general?*

Respuesta: Sıexiste tal modelo. En el Capítulo 3, se muestra como un modelo basado en un kernel se puede emplear para aprender con éxito un modelo estadístico del lenguaje de las opiniones a partir de un conjunto de semillas de palabras de opinión de dominio general. El modelo alcanza una precisión promedio que va desde 0,77 hasta 0,81 en los textos de opiniones en Inglés procesados, y de aproximadamente 0,85 para los textos con idioma español procesados.

RQ2 *De ser así, ¿podemos modelar el conjunto de características de las que se opina como otro modelo de lenguaje a partir del modelo del lexicón de opiniones? ¿Es posible definir tal modelo como una "traducción" del modelo de opiniones?*

Respuesta: Sí, también se puede modelar las características por medio de un modelo de lenguaje que se puede aprender mediante sucesivas transformaciones lineales del modelo de las palabras de opinión.

En el capítulo 2, se muestra que es posible aprender con éxito un modelo de lenguaje de las características a partir de un conjunto de palabras de opinión. Los experimentos demuestran como los modelos aprendidos en este capítulo obtienen valores de precisión entre 0,80 y 0,96 para las 50 primeras característocas recuperadas, y entre 0,78 y 0,95 para las primeras 100. En este caso, los valores de precisión se corresponden con los porcentajes de palabras que están inclídas en el nombre de alguna característica.

En el capítulo 3, se emplean los modelos de características para obtener un ranking de frases multipalabras para representar las características de productos. Los valores MAP de los rankings obtenidos son de aproximadamente 0,20, lo que indica que los modelos son lo suficientemente precisos para modelar las características del producto.

RQ3 *¿Podemos recuperar los datos subjetivos o de opinión (es decir, las estructuras con la forma característica-opinión) a partir de los modelos anteriores?*

Respuesta: Además de medir la precisión en la recuperación de las características multipalabras de productos en términos de MAP en el Capítulo 3, también evaluamos el desempeño de la recuperación de las opiniones que modifican a las características del producto. En este caso, los valores MAP obtenidos estuvieron, en general, por encima de 0,60. Esto corrobora que la recuperación general de los datos subjetivos (es decir, las características y las opiniones asociadas) puede llevarse a cabo de una manera eficaz. Esto nuevamente corrobora nuestras principales hipótesis y la utilidad de los modelos aprendidos para recuperar los datos de opinión.

RQ4 *En cuanto a la cuestión de almacenar y publicar los datos de opinión, ¿podemos integrar los datos extraídos en un almacén de datos corporativo tradicional para permitir BI? ¿Cuáles son los principales desafíos para lograr esta integración? ¿Es esta la solución adecuada para escenarios dinámicos, donde tanto las fuentes de datos y las necesidades de los usuarios pueden cambiar con el tiempo?*

Respuesta: En el Capítulo 4, se propone una nueva metodología para integrar los datos de opinión, que se obtienen aplicando la metodología propuesta en el Capítulo 3, en un modelo de BI implementado por medio de un almacén de datos corporativo. Para llevar a cabo dicha integración, hemos tenido que hacer frente a varios desafíos; siendo la anotación semántica y la ponderación de los datos de opinión usando su polaridad los más importantes. Por lo tanto, en primer lugar, desarrollamos un método de anotación semántica para realizar la unificación de los datos de opinión, y luego se propuso una primera aproximación para medir la orientación semántica de las tuplas (datos) de opinión.

Además, en el capítulo 5 se identifican nuevos problemas en la integración de los datos de opinión en los modelos de BI que se deben principalmente al dinamismo de los esenarios VoC y VoM. Por tanto, se propone SLOD-BI como una infraestructura de datos abierta basada en LOD, donde los datos de BI pueden vincularse a las fuentes externas de manera dinámica, sin estar vinculados indisolublemente a estructuras de datos predefinidas o a un esquema multidimensional rígido.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The ever increasing availability of user-generated contents in the Web, such as in Internet forums, social networks, blogs and review sites –i.e., repositories in which users express their opinions and sentiments and make them available to everyone–, has led to a growing challenge for Information Systems to effectively manage and retrieve the subjective information comprised in these repositories.

Opinion Mining, also known as Opinion Mining and Sentiment Analysis (OMSA), has arisen as the technology field that provides computational methods and tools to analyze opinions, sentiments and subjectivity from unstructured texts. This field is well-suited to various kinds of intelligence applications such as Business Intelligence (BI), which in the recent scenario of e-commerce could rely on extracted knowledge from reviews available on the Web in order to analyze recent trends as well as the satisfaction and behavior of customers and to prepare strategic plans accordingly. Thus, the need for methods and tools for analyzing and summarizing customer opinions from large repositories of customer reviews.

One of the most relevant applications of OMSA is the *aspect-based summarization* (Carenini et al., 2006; Liu, 2012; Yu et al., 2011). Broadly speaking, given a collection or stream of opinion posts, this task is aimed at obtaining the most relevant opined aspects, also called features, along with their associated sentiment information (usually an opinion word and/or a polarity score that express the sentiment orientation of the opinion). For example, given a collection of opinions about digital cameras, some relevant aspects can be the battery life, the quality of the lenses, etc. Sentences like "the battery life is too short" addresses the battery life aspect/feature and the sentiment information includes the opinion that it is too short.

Existing aspect-based summarization techniques can be broadly classified into two major approaches: supervised and unsupervised ones.

The supervised approaches require a set of pre-annotated review sentences as training examples. A supervised learning method is then applied to construct an extraction model, which is able to identify the opinion aspects (i.e., product features) from new customer reviews. Different approaches such as Hidden Markov Models and Conditional Random Fields (Jin et al., 2009; Wong and Lam, 2005, 2008), Maximum Entropy Models (Somprasertsri and Lalitrojwong, 2008), Class Association Rules and Naive Bayes Classifier (Yang et al., 2009), and other Machine Learning approaches have been employed to address this task. Although the supervised techniques can achieve reasonable effectiveness, preparing training examples is time consuming. In addition, the effectiveness of the supervised techniques has been shown to greatly depend upon the representativeness of the training examples.

Thus, in this thesis we are mostly interested in unsupervised approaches, which aims to automatically extract the different aspects from opinion posts without involving training examples. The reason is that existing unsupervised approaches have several limitations. For example, in many cases they require from some manual tuning of various parameters, such as the number of aspects to be extracted as in the topic modeling-based approaches (Mei et al., 2007; Titov and McDonald, 2008). Other approaches directly rely on term frequency and thus tend to extract many non-aspect features and discard low frequency ones (Qiu et al., 2011). Other approaches such as (Zhang et al., 2010) explicitly rely on lexical relations (e.g., part-whole) to set up their extractors, but rarely generalize features. Instead, they just regard a ranking algorithm to extract the features from a set of candidates that have been already identified by means of some NLP-based heuristics.

In this context, this thesis addresses the following issues:

(1) How to effectively extract the user opinions and their specific targets from a collection of customer reviews about a product?

(2) How to store and publish these data for further analysis?

In particular, we are interested in providing new unsupervised methodologies for extracting user opinions and their targets (from collections of unstructured texts) that cope with the issues mentioned above. We also aim at providing multilingual solutions that cover as many source languages as possible (but a single language at once).

1.2 Goals

The main goal of this thesis is to develop new methodologies to effectively extract the main opinions and their targets from a collection of unstructured texts. Specifically, we aim to:

1. Contribute with a general methodology to address the problem of aspect-based summarization. The proposed methodology is required to:

    - be an unsupervised methodology that completely disregard manually labeled examples of domain features and opinions,
    - effectively extract opinions and their targets despite both the size of the opinion posts and the frequency of the opinion in the collection (i.e., both large and short customer reviews should be effectively processed, and both highly frequent and sparse opinions should be retrieved),

  - cover as many languages as possible.

2. Evaluate the performance of the proposed methodology using manually labeled collections of reviews of different source language (e.g., English and Spanish). Broadly, the evaluation should include (i) validating the performance of the methodology on each: the extraction of product features (i.e., the targeted aspect of the users opinions) and the extraction of their opinions, and (ii) comparing the performance of the methodology to state-of-the-art methods.

3. Provide solutions to store and publish retrievable structures composed of the extracted features and their respective opinions for further semantic analysis (e.g. to be applied to BI).

1.3 Main hypotheses

For a given opinion domain we regard the following hypotheses:

- **Hypothesis 1**: There is a statistical language model capable of modeling a lexicon of domain opinions, which can be successfully learned from a seed set of general-domain opinion words.

- **Hypothesis 2**: Since product features are the target of user opinions, there is also a language model of product features (from which product features can be successfully generated) that can be learned as a translation from the lexicon model above via a lexical mapping (namely, a lexical mapping between words).

1.4 Research questions

Aligned with the main goals and hypotheses of this thesis, we consider to answer the following research questions:

RQ1 Is there a (statistical) language model capable of modeling a lexicon of opinion words for a given opinion domain or product from a seed set of general domain opinion words?

RQ2 If so, can we model the collection of opinion targets as another language model from the lexicon model of opinions? Is it possible to define the language model of opinion targets as a translation from the lexicon model?

RQ3 Can we effectively retrieve the subjective data (i.e., structures with the form product feature-opinion) from the above models?

RQ4 Regarding the issue of storing and publishing sentiment data, can we integrate the extracted sentiment data into a traditional corporate data warehouse (DW) to enable BI? Which are the main challenges to achieve this integration? Is this solution suitable to dynamic scenarios where both the data sources and the user requirements may change over time?

1.5 Background

1.5.1 Statistical Language Modeling

Commonly, Statistical Language Modeling (SLM) constitutes a method to represent text documents as well as user queries in Information Retrieval. However, in this thesis we consider SLM to represent models of products features and opinions from which we then perform the extraction/retrieval of these subjective data.

Formally, a statistical language model is a function that defines a probability distribution over the elements in a language; where a language is a set of word sequences over a vocabulary.

In the document representation scheme, each document is represented by means of a statistical language model, that is often a model from which the document is a sample with high likelihood.

Usually, the language underlying the model is defined in terms of all possible sequences of fixed length composed over the vocabulary of the document collection. These sequences are called n-grams, and the statistical language model is referred to as n-gram language model; where n is the length of the sequences ($n \geq 1$).

In an n-gram language model, the probability distribution that represents a text document is estimated from all sequences of length n included in the document (i.e., the n-grams of the document). Some estimation methods of the probability distribution that defines a model for a document d_i are the following:

- *Maximum Likelihood Estimator* (MLE):

$$\mathrm{p}_{MLE}(s|d_i) = \frac{tf(s,d_i)}{\sum_{s' \in d_i} tf(s',d_i)} \tag{1.1}$$

- *Laplace or adding one smoothing*

$$\mathrm{p}(s|d_i) = \frac{tf(s,d_i) + 1}{\sum_{s' \in d_i} tf(s',d_i) + |S|} \tag{1.2}$$

- *Jelinek-Mercer smoothing*

$$\mathrm{p}(s|d_i) = (1-\lambda)\,\mathrm{p}_{MLE}(s|d_i) + \lambda\,\mathrm{p}(s|D) \tag{1.3}$$

- *Dirichlet smoothing*

$$\mathrm{p}(s|d_i) = \frac{tf(s,d_i) + \mu\,\mathrm{p}(s|D)}{\sum_{s' \in d_i} tf(s',d_i) + \mu} \tag{1.4}$$

where $s \in S$ is an n-gram over the vocabulary of the collection, $tf(s,d_i)$ accounts for the number of times n-gram s is included document d_i, $\mathrm{p}(s|D)$ is an estimated probability value for s under the document collection, and both λ and μ represent smoothing factors ($0 < \lambda < 1$, $\mu > 0$).

Commonly, an n-gram language model representing a document d_i is chosen to be a stochastic language model; i.e, a model $\{\mathrm{p}(s|d_i)\}_{s\in S}$ such that $\sum_{s\in S}\mathrm{p}(s|d_i) = 1$.

Particular cases of n-gram language models frequently used for representing text documents in practice are the *unigram* and *bigram* language models, which are defined by setting $n = 1$ and $n = 2$ respectively.

Representing documents using statistical language models allows us to relate text documents in a variety of ways. For example, it can be estimated the probability of generating an arbitrary document or phrase (over the vocabulary of the collection) from the statistical language model representing a given document in a document collection. Besides, distance metrics between distributions such as the geodesic distance (Dillon et al., 2007), can be employed to set up a family of kernel-based density estimators. The geodesic distance between distributions $\mathrm{p}_i = \{\mathrm{p}_i(s)\}_{s\in S}$ and $\mathrm{p}_j = \{\mathrm{p}_j(s)\}_{s\in S}$ is defined as follows:

$$g(\mathrm{p}_i, \mathrm{p}_j) = 2\arccos\left(\sum_{s\in S}\sqrt{\mathrm{p}_i(s)}\sqrt{\mathrm{p}_j(s)}\right) \tag{1.5}$$

1.5.2 Business Intelligence

BI refers to the methodologies, architectures and technologies that transform raw data into meaningful and useful information to enable more effective decision-making in business. BI technologies provide historical, current and predictive views of business operations. Common functions of BI are reporting, online analytical processing (OLAP), data mining, complex event processing and text mining among others. Often BI applications use data gathered from a DW or a data mart. In fact, one of the most successful approaches to BI has been the combination of DW and OLAP (Codd et al., 1993).

Traditional BI follows a three-layered architecture consisting of the data sources layer, where all the potential data of any nature is gathered, the integration layer, which transforms and cleanses the data from the sources and stores them in a DW, and the analysis layer, where different tools exploit the integrated data to extract useful knowledge that is presented to the analyst as charts, reports, cubes, etc. For the integration layer, the multidimensional model (MD) is used, where factual data gathered from the data sources layer must be expressed in terms of numerical measures and categorical dimensions. The semantics of this model consists in representing any interesting observation of the domain (i.e., a measure such as particular sales, profits, etc.) at its context (dimensions). The typical processes in charge of translating data from the data sources layer to the integration layer are called ETL processes (extract, transform, and load).

In recent years, the massive availability of web-based social media related to business processes has become a valuable asset for the BI community. Thus, the integration of these external and heterogeneous data sources with corporate data would enable more insightful analysis and would bring new marketing opportunities so far unexplored.

The problem of how to exploit social data to extract sentiment data that could be useful for BI applications and how to integrate the extracted opinion data into the existing corporate DW is still an open issue, which is addressed in this thesis.

1.6 Related work

In what follows, we review some of the most important and recent approaches to the problem of aspect-based opinion mining. Specifically, we focus on unsupervised approaches, since the supervised ones require from a set of pre-annotated training examples, and their effectiveness greatly depends on the representativeness of the these examples.

Besides, we are mainly focused on multi-domain and multi-lingual solutions to the problem of aspect-based sentiment summarization, which can be hardly approached by means of supervised methods.

We have broadly classified the approaches into two classes: the class of frequency-based approaches and the class of approaches based on Probabilistic Topic Modeling.

1.6.1 Frequency-based approaches

Approaches in this class mainly base the extraction of product features or aspect on high-frequency noun phrases. Thus, the main limitations of these approaches is that they tend to discard features that occur with low frequency. But also, they currently extract many non-aspect nouns (i.e., false positives).

Additionally, these approaches require some manual tuning of various parameters in many cases, such as the frequency threshold used to filtering out non-aspect nouns. This makes it hard to effectively apply these approaches to collections of opinions of different sizes without some human supervision.

To overcome these limitations, several methods also rely on lexical relations to approach the relationship between aspects and sentiments in opinion contexts in order to obtain more accurate solutions. However, such a heuristic introduces false positives as well as generates false negatives.

Some approaches included in this subclass of unsupervised approaches are described in turn.

The work by Hu and Liu (2004a,b), called Feature-based Summarization (FBS), uses the Apriori algorithm (Agrawal et al., 1994) to extract frequent itemsets as explicit product features.Then, in order to remove wrong frequent features, two types of pruning criteria are used: compactness and redundancy pruning. The technique is efficient and does not require the use of training examples or predefined sets of domain-independent extraction patterns. However, the design principle of FBS is not anchored in a semantic perspective. As a result, it is ineffective in excluding non-aspect features as well as opinion-irrelevant features. The algorithm also neglects the order of the words in the frequent itemsets to define the product features. As stated in (Wei et al., 2010), all these limitations negatively impact on the performance of the approach.

Popescu and Etzioni proposed the system named OPINE (Popescu and Etzioni, 2007), which is built on top of KnowItAll, a Web-based, domain-independent information extraction system (Etzioni et al., 2005). Product features are considered to be "concepts" that exhibit some relationships with the product (for example, for a "scanner", its "size" is one of its properties, whereas its "cover" is one of its parts). First, all noun phrases with frequency greater than a threshold are extracted. Then, each noun phrase is evaluated by estimating the Pointwise Mutual Information (PMI) between the phrase and meronymy discriminators associated to the product class (e.g., "of scanner", "scanner has", "scanner comes with", etc. for the *scanner* class). Then, each sentence is annotated with syntactic dependencies, and, finally, potential opinion phrases are identified by applying a set of syntactic rules over these dependencies. A relaxation labeling technique is used to find the semantic orientation of the opinion phrases in the context of the product features and sentences. One limitation of OPINE system is that it might not generalize well across product categories.

The Opinion Observer system proposed by Liu et al. (2005) is a framework for analyzing and comparing consumer opinions of competing products. It is oriented to short comments, and assumes that each sentence segment contains at most one product feature where sentence segments are fragment of texts separated by ',', '.', 'and', or 'but'. The method relies on a supervised rule-based discovery algorithm to extract product features. Once a training dataset has been tagged using a POS tagger, all words indicating an aspect are replaced by the generic pattern "[feature]", even if it is an implicit aspect. This aims to represent general language patterns so that the system can be applied to any product. Then, the association mining system CBA (Liu et al., 1998) is employed to mine aspects. As not all of the generated rules are useful, some post-processing is needed to filter out patterns that are not aspect predictors. For sentiment orientation, they simply assume that sentiments appearing in a Pros section are positive, whereas those appearing in a Cons section are negative.

In (Qiu et al., 2011), a method based on bootstrapping is proposed. Specifically, it relies on the use of the double-propagation strategy (and therefore it is called *Double Propagation*) to allow the incremental identification of features and opinion words from a predefined initial set of words (usually, a lexicon of opinion words). The idea underlying the use of the double-propagation strategy is to iteratively apply a series of dependency-based patterns to incrementally identify new features and new opinion words until no new elements can be further identified. Double propagation assumes that features are nouns (namely, noun phrases) and opinion words are adjectives. The method works well for small and medium–size corpora. But for large corpora, the method tends to extract many nouns that are not features.

The method by Zhang et al. (2010) follows a similar idea to that of the Double Propagation approach. However, instead of incrementally identifying new features and opinions, it uses the algorithm Hyperlink-Induced Topic Search (HITS) (Kleinberg, 1999) to rank possible features from a candidate set. Thus, candidates in the bottom of the ranking are assumed to have low interactions with opinion words and therefore they are considered to be noise. Additionally, two improvements based on part-whole relation patterns and a "no" pattern are regarded to find features that double propagation cannot find. Different from (Qiu et al., 2011), the scoring scheme based on HITS is not used to discover new features and opinion words from customer reviews but just to rank candidate features already identified through NLP-based heuristics.

Relying on a similar bootstrapping principle to that of (Qiu et al., 2011) and (Zhang et al., 2010), Cruz (2011) proposes a method based on simple syntactic relations and morphosyntactic information to extract domain features from a seed set of opinion words. However, this method is interactively supervised by a human expert.

1.6.2 Approaches based on Probabilistic Topic Modeling

PTM is an unsupervised learning technique focused on inferring a set of probability distributions from a target collection to model each individual element in the collection (usually a text document represented by means of a bag of words) as a mixture of such distributions. Thus, the output of a PTM approach is typically a set of word distributions, each one deemed to represent the main themes that runs through the target collection and therefor these distributions are called *topics*.

The aim of adapting PTM to OMSA has not been to extract specific named features but to learn some broader aspects (i.e., topics) that can represent the different features and/or their corresponding opinions. Many adaptations have been performed (Jo and Oh, 2011; Lu et al., 2009; Mei et al., 2007; Titov and McDonald, 2008; Wang et al., 2010b; Zhao et al., 2010). However, the learned distributions from these adaptations in many cases do not correspond to actual features and/or opinions. They are simply vocabulary regularities found in the corpus that are difficult to interpret by ostensible end-users.

Last but very important, all these approaches need to know a priori the number of aspect to be modeled, which is not applicable to our unsupervised scenario. In practice, it is very difficult to predict such a number in an automatic manner.

1.7 Organization

This thesis is presented as a compendium of publications. Thus, the remainder of the manuscript is organized as follows:

Chapter 2 presents the work "Probabilistic ranking of product features from customer reviews"; where we propose to learn a probabilistic model of product features in an unsupervised manner from a collection of customer reviews. The novelty of this work relies on modeling features from a stochastic mapping model between words. The model is able to obtain also a ranking of corpus-based opinion words. One strong point of this method is that it does not need deep natural language processing except for lemmatization/POS tagging.

Chapter 3 presents a more consolidated work on modeling and extracting product features and their corresponding opinion; namely, "Retrieving product features and opinions from customer reviews". In this publication, a new methodology for the retrieval of product features from a collection of customer reviews about a product or service is presented. The proposed methodology does not require any training examples of product features nor domain specific opinion words. From a seed set of general-domain opinion words, the methodology firstly induces a stochastic model of opinions for the target product/service based on a kernel function between word distributions. Then, it learns a language model of product features from which we finally implement a method to retrieve both the features and their associated opinions.

Chapter 4 presents the article "Storing and analysing voice of the market data in the corporate data warehouse"; where a multidimensional data model is defined to integrate sentiment data extracted from opinion posts into a BI model implemented as a traditional corporate DW. A case study over a set of real opinions about digital devices is developed. This article also introduces a semantic annotation model to be applied to both corporate and sentiment data. Besides, new measures to assess the relevance of product features and opinion facts are proposed and implemented.

Chapter 5 presents the article "SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence"; where the integration of BI with sentiment data is extended to the recent scenario provided by linked open data (LOD). Specifically, it proposes an open and dynamic framework based on LOD, where data can be linked to external sources on demand, without being attached to predefined (rigid) data structures or multidimensional schema.

Chapter 6 concludes this thesis by summarizing the main contributions and results. In the summary, we stress the relationship between the articles from chapters 2 to 5. Finally, the chapter outlines and describes interesting lines for further research relying on the results of this thesis.

Chapter 2

Probabilistic ranking of product features from customer reviews

Lisette García-Moya, Henry Anaya-Sánchez, Rafael Berlanga, and María José Aramburu. In Pattern Recognition and Image Analysis, pp. 208-215. Springer Berlin Heidelberg, 2011.

2.1 Abstract

In this paper, we propose a methodology for obtaining a probabilistic ranking of product features from a customer review collection. Our approach mainly relies on an entailment model between opinion and feature words, and suggests that in a probabilistic opinion model of words learned from an opinion corpus, feature words must be the most probable words generated from that model (even more than opinion words themselves). In this paper, we also devise a new model for ranking corpus-based opinion words. We have evaluated our approach on a set of customer reviews of five products obtaining encouraging results.

2.2 Introduction

The Web has become an excellent way of expressing opinions about products. Thus, the number of Web sites containing such opinions is huge and it is constantly growing. In recent years, opinion mining and sentiment analysis has been an important research area in Natural Language Processing (Pang and Lee, 2008). Product featured extraction is a task of this area, and its goal is to discover which aspects of a specific product are the most liked or disliked by customers. A product has a set of components (or parts) and also a set of attributes (or properties). The word *features* is used to represent both components and attributes. For example, given the sentence, "The battery life of this camera is too short", the review is about the "battery life" feature and the opinion is negative.

This paper focuses on the feature extraction task. Specifically, given a set of customer reviews about a specific product, we address the problem of identifying all possible potential product features and ranking them according to their relevance. The basic idea of our ranking method is that if a potential feature is valid, it should be ranked high; otherwise it should be ranked low in the final result. We believe that ranking is also important for feature mining because ranking helps users to discover important features from the extracted hundreds of fine-grained potential features.

The remainder of the paper is organized as follows: Section 2.3 describes related work. In Section 2.4, we explain the proposed methodology. Section 2.5 presents and discusses the experimental results. Finally, in Section 2.6, we conclude with a summary and future research directions.

2.3 Related Work

Existing product feature extraction techniques can be broadly classified into two major approaches: supervised and unsupervised ones.

Supervised product feature extraction techniques require a set of preannotated review sentences as training examples. A supervised learning method is then applied to construct an extraction model, which is able to identify product features from new customer reviews. Different approaches such as Hidden Markov Models and Conditional Random Fields (Wong and Lam, 2005, 2008), Maximum Entropy (Somprasertsri and Lalitrojwong, 2008), Class Association Rules and Naive Bayes Classifier (Yang et al., 2009) and other ML approaches have been employed for this task.

Although the supervised techniques can achieve reasonable effectiveness, preparing training examples is time consuming. In addition, the effectiveness of the supervised techniques greatly depends on the representativeness of the training examples. In contrast, unsupervised approaches automatically extract product features from customer reviews without involving training examples. According to our review of existing product feature extraction techniques, the unsupervised approaches seem to be more flexible than the supervised ones for environments in which various and frequently expanding products get discussed in customer reviews.

Hu and Liu's work (Hu and Liu, 2004a,b) (PFE technique), uses association rule mining based on the Apriori algorithm (Agrawal and Srikant, 1994) to extract frequent itemsets as explicit product features. However, this algorithm neglects the position of sentence words. In order to remove wrong frequent features, two types of pruning criteria were used: compactness and redundancy pruning. The technique is efficient and does not require the use of training examples or predefined sets of domain-independent extraction patterns. However, the design principle of PFE technique is not anchored in a semantic perspective. As a result, it is ineffective in excluding non-product features and opinion-irrelevant product features. Such limitations greatly limit its effectiveness. Details about these limitations are presented in (Wei et al., 2009). To address these limitations, Wei et al. (2009) proposed a semantic-based product feature extraction technique (SPE) that exploits a list of positive and negative adjectives defined in the General Inquirer (Stone et al., 1966b) in order to recognize opinion words, and subsequently to extract product features expressed in customer reviews. Even when the SPE technique attains better results than previous approaches, both rely on mining frequent itemsets, with its commented limitations.

Qiu et al. (2009) proposed a double propagation method, which exploits certain syntactic relations between opinion words and features, and propagates them iteratively. A dependency grammar was adopted to describe relations between opinion words and features themselves. The extraction rules are designed based on these relations. This method works well for medium-size corpora. However, for large and small corpora, it can result in low precision and low recall (Zhang et al., 2010). To deal with these two problems, Zhang et al. (2010) introduce *part-whole* and *no* patterns to increase the recall. Finally, feature ranking is applied to the extracted feature candidate in order to improve the precision of the top-ranked candidates.

2.4 Methodology

In this section we propose a new methodology to extract features from opinion reviews. It firstly extracts a set of potential features. Then, it defines a translation model based on words entailments. The purpose is to obtain a probabilistic ranking of these potential features. Finally, a new model for ranking corpus-based opinion words is proposed.

2.4.1 Extraction of Potential Features and Construction of a Basic Opinion Words List

Potential Features: In this work, we consider a set of *potential features* defined as the word sequences that satisfy the following rules:

1. Sequences of nouns and adjectives (e.g. "battery life", "lcd screen").
2. When gerund and participle occur between nouns, they are considered as part of the feature (e.g. "battery charging system").
3. Let PF_1 and PF_2 be potential features extracted by applying any of the previous rules. Let also $connector_1 = (of, from, at, in, on)$, and $connector_2 = (the, this, these, that, those)$. If the pattern $PF_1 connector_1 [connector_2] PF_2$ occurs, then the phrase formed by PF_2 concatenated with PF_1 is extracted as a potential feature. For example, "quality of photos" → "photo quality".

Opinion Words: We construct our own list of basic *opinion words*. An initial list was created by the intersection of adjectives from the list proposed by Eguchi and Lavrenko (2006), the synsets of WordNet scored positive or negative in SentiWordNet (Esuli and Sebastiani, 2006) and the list of positive and negative words from the General Inquirer. Then, this initial list was extended with synonyms and antonyms from WordNet 3.0. Finally, the obtained list was manually checked, discarding those adjectives with context-dependent polarity. Additionally, some adverbs and verbs with context-independent polarity were added. The resulting list is formed by 1176 positive words and 1412 negative words.

2.4.2 Translation Model for Feature Ranking

In order to rank the set of potential features from customer reviews with vocabulary $V = \{w_1, \ldots, w_n\}$, we rely on the entailment relationship between words given by $\{p(w_i|w_j)\}_{w_i, w_j \in V}$, where $p(w_i|w_j)$ represents some posterior probability of w_i given w_j. In this work, we interpret $p(w_i|w_j)$ as the probability that $w_j \in V$ entails word $w_i \in V$.

In the context of customer reviews, opinion words usually express people sentiments about features, and therefore they can be seen as feature modifiers. Thus, in our work we consider that feature words can be successfully retrieved from the ranking given by the following conditional probability:

$$p(w_i|O) = \sum_{w \in V} p(w_i|w) \cdot p^*(w|O) \tag{2.1}$$

where $i \in \{1, \ldots, n\}$ and $\{p^*(w|O)\}_{w \in V}$ represents a basic language model of opinion words. The underlying idea is that in a probabilistic opinion model of words learned from an opinion corpus, feature words must be the most probable words generated from that model (even more than opinion words themselves), because of the entailment relationship between opinion and feature words. In this way, we regard that the probability of including a word w into the class of feature words F can be defined as:

$$p(F|w) \propto p(w|O). \tag{2.2}$$

Notice that if we estimate $\{p(w_i|w_j)\}_{w_i, w_j \in V}$ from customer reviews, the model $\{p(w_i|O)\}_{w_i \in V}$ can be seen as a corpus-based model of opinion words that is obtained by smoothing the basic model $\{p^*(w_i|O)\}_{w_i \in V}$ with the translation model $\{p(w_i|w_j)\}_{w_i, w_j \in V}$. Accordingly, we can also obtain a ranking of corpus-based opinion words by using Bayes formula on $\{p(w_i|O)\}_{w_i \in V}$. That is,

$$p(O|w_i) \propto \frac{p(w_i|O)}{p(w_i)}. \tag{2.3}$$

The same analysis can also be applied to features defined by multiword phrases (e.g. "battery life", "battery charging system" or "highest optical zoom picture"). Specifically, the probability of including a phrase $s = w_{i_1} \ldots w_{i_m}$ into the class of general features $\mathscr{F}$ can be defined as:

$$p(\mathscr{F}|s) \propto p(s|O) = p(w_{i_1} \ldots w_{i_m}|O). \tag{2.4}$$

2.4.3 Probability Density Estimation

The above probabilistic models for retrieving features and corpus-based opinion words depend on estimations for $\{p(w_i|w_j)\}_{w_i, w_j \in V}$, $\{p(w_i)\}_{w_i \in V}$ and the basic model of opinion words $\{p^*(w_i|O)\}_{w_i \in V}$.

For estimating $\{p(w_i|w_j)\}_{w_i, w_j \in V}$ we rely on a translation model like that presented in (Berger and Lafferty, 1999). Thus, we firstly compute an initial word posterior probability conditioned on the vocabulary words defined as:

$$p_1(w_i|w_j) = \frac{p_1(w_i,w_j)}{p_1(w_j)} \tag{2.5}$$

where

$$p_1(w_i,w_j) \propto \sum_{v \in W} p(w_i|v) \cdot p(w_j|v) \cdot p(v), \tag{2.6}$$

$$p_1(w_j) = \sum_{w_i \in V} p_1(w_j,w_i), \tag{2.7}$$

and W is the set of all possible word windows of size k that can be formed in each sentence from the customer reviews. In the experiment carried out in this paper the best performance is achieved using $k = 5$. These probabilities are estimated using $p(w_i|v) = |v|_{w_i}/|v|$ and $p(v) = |W|^{-1}$, where $|v|_{w_i}$ is the number of times w_i occurs in window v, $|v|$ is the length of v, and $|W|$ is the cardinal of W.

For all $w_i,w_j \in V$, the probability $p_1(w_i|w_j)$ can be seen as the probability of translating w_j into w_i in one translation step. Then, we define $p(w_i|w_j)$ as a smoothed version of $p_1(w_i|w_j)$ obtained by generating random Markov chains between words. Specifically, we define $p(w_i|w_j)$ as:

$$p(w_i|w_j) = \left((1-\alpha) \cdot (I - \alpha \cdot P_1)^{-1}\right)_{i,j} \tag{2.8}$$

where I is the $n \times n$ identity matrix, P_1 is a $n \times n$ matrix whose element P_{ij} is defined as $p_1(w_i|w_j)$, and α is a probability value that allows the generation of arbitrary Markov chains between words. In the experiment carried out in this paper we use $\alpha = 0.99$, which allows the generation of large chains, and thus a great smoothing.

Thus, the overall model of words $\{p(w_i)\}_{w_i \in V}$ can be estimated from the linear equation system given by the n variables $\{p(w_i)\}_{w_i \in V}$, and $n+1$ equations:

$$p(w_i) = \sum_{w_j \in V} p(w_i|w_j) \cdot p(w_j) \qquad (i \in \{1,\ldots,n\}) \tag{2.9}$$

$$\sum_{w_i \in V} p(w_i) = 1. \tag{2.10}$$

The basic model of opinion words considered in this work is estimated from the list of basic opinion words described in section 2.4.1. We consider $p^*(w_i|O)$ defined as the following non-smoothed model:

$$p^*(w_i|O) = \begin{cases} \frac{1}{|Q|} & \text{if } w_i \in Q \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

where Q is the set of basic opinion words and $|Q|$ is the size of Q.

Table 2.1 Summary of customer review data set

	Apex	Canon	Creative	Nikon	Nokia
Number of review sentences	738	600	1705	350	548

Table 2.2 Precision at top N

N	*Baseline*					*Our approach*				
	Apex	Canon	Creative	Nikon	Nokia	Apex	Canon	Creative	Nikon	Nokia
50	72.0	92.0	78.0	72.0	86.0	96.0	92.0	80.0	90.0	94.0
100	62.0	78.0	72.0	55.0	67.0	95.0	94.0	82.0	78.0	91.0
150	48.0	61.3	69.3	43.3	52.7	92.0	91.3	82.0	72.0	86.0
200	42.5	54.0	69.5	38.0	49.0	91.0	91.5	81.0	65.0	78.5
250	42.0	54.0	64.4	37.2	46.0	85.6	90.0	80.0	58.4	73.6

2.5 Experiments

In order to validate our methodology, we have conducted several experiments on the customer reviews from five products: Apex AD2600 Progressive-scan DVD player, Canon G3, Creative Labs Nomad Jukebox Zen Xtra 40GB, Nikon coolpix 4300 and Nokia 6610. The reviews were collected from Amazon.com and CNET.com. [1] Table 2.1 shows the number of review sentences for each product in the data set. Each review was annotated using the Stanford POS Tagger. [2]

Firstly, we propose to compare our ranking method with a version of the method proposed by Zhang et al. (2010). Zhang et al. considered that the importance of a feature is determined by its relevance and its frequency. In order to obtain the relevance score of a feature, they apply the HITS algorithm where *potential features* act as authorities and *feature indicators* act as hubs forming a directed bipartite graph. The basic idea is that if a potential feature has a high authority score, it must be a highly-relevant feature. If a feature indicator has a high hub score, it must be a good feature indicator. The final score function considering the feature frequency is:

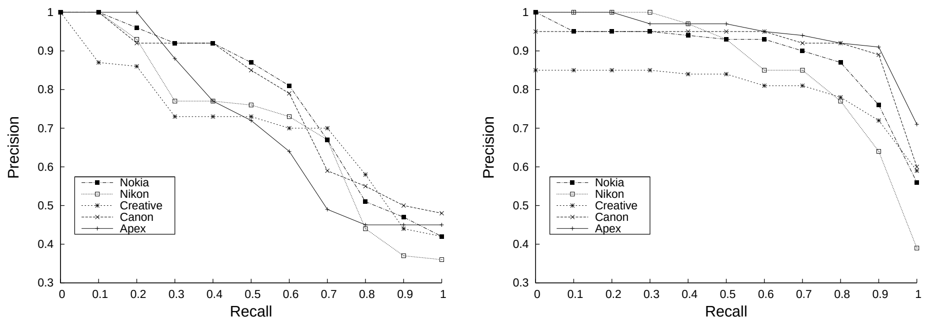
$$score(s) = A(s) \cdot \log(freq(s)) \tag{2.12}$$

where $freq(s)$ is the frequency of the potential feature s, and $A(s)$ is the authority score of the potential feature s. In our case, an opinion word o co-occurring with any word $w_i \in s$ in the same window v is considered as a feature indicator of s. We are going to consider this method as our *baseline*.

The performance of the methods is firstly evaluated in terms of the measure **precision@N**, defined as the percentage of valid features that are among the top N features in a ranked list. In Table 2.2, we show the obtained results for each $N \in \{50,100,150,200,250\}$. As it can be seen, our method consistently outperforms the baseline for each value of N. Also, it can be appreciated that different

[1] This customer review data set is available at `http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip`

[2] http://nlp.stanford.edu/software/tagger.shtml



(a) Baseline

(b) Our approach

Fig. 2.1 11-point Interpolated Recall-Precision curve for each product. The x-axis represents different recall levels while y-axis represents the interpolated precision at these levels.

Table 2.3 Fragment of the ranking of corpus-based opinion words obtained for some products.

Relevance	Creative	Relevance	Nokia	Relevance	Nikon
1.061	helpful	1.219	haggard	1.076	worse
1.057	**clock**	1.212	mad	1.070	**claim**
1.050	weighty	1.210	**gott**	1.069	kind
1.044	**biggie**	1.205	**junky**	1.063	**internal**
1.040	strange	1.202	**major**	1.059	**damage**
1.038	unlucky	1.197	bad	1.055	**refuse**
1.038	flashy	1.197	**duper**	1.052	**cover**
1.037	superfluous	1.197	**rad**	1.026	correct
1.037	evil	1.196	happy	1.023	**warranty**
1.036	smoothly	1.195	minus	1.022	**touchup**
1.036	user-friendly	1.193	negative	1.022	**alter**
1.036	**date**	1.190	brisk	1.022	redeye
1.036	**ounce**	1.189	significant	1.010	**cost**
1.036	ridiculous	1.188	**require**	1.009	outstanding
1.036	shoddy	1.188	**penny**	1.008	comfortable

from the baseline, the precision of our rankings do not decrease dramatically when N is greater than 100.

Secondly, we consider the 11-point interpolated average precision to evaluate the retrieval performance regarding the recall factor. Figure 2.1 compares the obtained curves. It can be seen that even considering the 11-point of recall scores, our approach outperforms the baseline, while also maintains a good precision through out all recall values.

Finally, as it was explained in Section 2.4.2, it is possible to obtain a ranking of corpus-based opinion words (see Equation 2.3). Table 2.3 shows the first 15 opinion words obtained for some products together with their relevance value (i.e., $p(w_i|O)/p(w_i)$). In this table, we bold-faced those words that are not included in our basic opinion list. The obtained ranking corroborates the usefulness of the proposal for also retrieving corpus-based opinion words.

2.6 Conclusions and Future Work

In this paper, a new methodology for obtaining a probabilistic ranking of product features from customer reviews has been proposed. The novelty of our approach relies on modeling feature words from a stochastic entailment model between opinion and feature words. In addition, a model for obtaining a ranking of corpus-based opinion words is also proposed. One strong point of our method is that it does not depend on any natural language processing except for POS tagging. The experimental results obtained over a set of customer reviews of five products validate the usefulness of our proposal. Our future work is oriented to design a method to cut the ranked list in order to remove those spurious features.

Acknowledgments.

This work has been partially funded by the Spanish Research Program project TIN2008-01825/TIN and the Fundacio Caixa Castelló project P1.1B2008-43. Lisette García-Moya has been supported by the PhD Fellowship Program of the Universitat Jaume I (PREDOC/2009/12).

Chapter 3

Retrieving product features and opinions from customer reviews

Lisette Garcia-Moya, Henry Anaya-Sanchez, and Rafael Berlanga-Llavori. IEEE Intelligent Systems 3, (2013): 19-27.

3.1 Abstract

In this article, we present a new methodology for the retrieval of product features from a collection of customer reviews about a product or service. The proposed methodology doesn't require any training set of product features, and the experiments carried out over several collections of customer reviews in English and Spanish have shown the proposal's usefulness for properly retrieving the product features and the opinions expressed about them, even from individual reviews. For future work, we plan to integrate our models into a probabilistic topic-modeling framework. The aim is to provide features and opinions with a topic-based description representing them. We also plan to extend our methodology to model the polarity of the opinion words ascribed to product features.

3.2 Introduction

With the increasing availability of user-generated contents, such as consumer opinion web sites, blogs, Internet forums and social networks, people have more opportunities to express their opinions and make them available to everyone. Publicly available opinions provide valuable information for decision-making processes based on a new collective intelligence paradigm referred to as *crowdsourcing*. This has inspired research in opinion mining and sentiment analysis to develop methods for automatically detecting emotions, opinions and other evaluations from texts.

One of the most relevant applications of opinion mining and sentiment analysis is aspect-based summarization (Carenini et al., 2006; Yu et al., 2011). Broadly speaking, given a collection of opinion posts, this task is aimed at obtaining relevant aspects (such as product features), along with associated sentiment information expressed by customers (usually an opinion word and/or a polarity score).

Aspect-based summarization is usually composed of three main tasks: aspect identification, sentiment classification, and aspect rating. Aspect identification is focused on extracting the set of aspects or product features from the source collection. The word *aspect* is intended to represent the opinion or sentiment targets, which are also referred to as *product features* (Qiu et al., 2011) when the collection of posts –typically, customer reviews– is about products or services. For example, given the sentence, "The bed was comfortable" in a review about a hotel room, the aspect being referred to is the "bed" and the opinion is positively expressed by means of the opinion word "comfortable". The sentiment classification task consists in determining the opinions about the aspects and/or their polarities, whereas aspect rating leverages the relevance of aspects and their opinions to properly present them to the users.

Here, we address the aspect-based summarization task by introducing a novel methodology for retrieving product features from a collection of free-text customer reviews about a product or service. The proposal relies on a language modeling framework that combines a probabilistic model of opinion words and a stochastic mapping model between words to approximate a language model of products.

Our work extends the preliminary approach introduced elsewhere (García-Moya et al., 2012) that addresses the modeling of a language of product features from customer reviews. Specifically, we propose here a more general methodology that effectively allows –for example– the use of grammatical dependency relations between words in modeling the language of features. We also provide a more formalized methodology for the retrieval of (multi-word) product features from the estimated language model of features, along with a more comprehensive evaluation.

As already shown in other work (García-Moya et al., 2012), one strong point of our proposal is that it can effectively retrieve product features without relying on natural language processing (NLP) techniques. This is the main difference with respect to most existing approaches on opinion mining and sentiment analysis for feature identification (Qiu et al., 2011; Wu et al., 2009), as they strongly rely on grammatical dependency analysis.

3.3 Related Work

Previous work on extracting product features from customer reviews has mainly relied on Natural Language Processing (NLP) (Liu, 2012). Part-of-speech (POS) tagging, shallow parsing techniques, and dependency grammars have been widely applied to identify both noun phrases that act as potential features and opinion words that affect them through syntactical dependencies. Using the double-propagation strategy (Qiu et al., 2011) allows the incremental identification of features and opinion words from a predefined initial set (usually a lexicon of opinion words). Generally, NLP-based approaches present good precision but low recall figures because they depend on the definition of extraction patterns, which are dependent on both the particular language and the reviews' application domain.

Another limitation of NLP-based approaches is that they don't account for feature relevance. Thus, an additional process is required for scoring the identified features. Most approaches just apply simple statistics such as word counts to rank the features (Hu and Liu, 2004b). A recent approach (Zhang et al., 2010) applies the Hyperlink-Induced Topic Search (HITS) (Kleinberg, 1999) algorithm to score the identified features according to their interaction with opinion words. In contrast to our proposal, these scoring schemes aren't used to discover new features and opinion words from customer reviews, but only to rank features already identified through some NLP-based method.

Other recent approaches propose to extract sentiment and aspect words from corpora (Cambria et al., 2013; Titov and McDonald, 2008; Wang et al., 2010b, 2011; Yu et al., 2011). In these approaches, the objective isn't to find specific product features, but some predefined broader aspects. Usually, these approaches state the problem as a particular case of statistical inference such as Latent Dirichlet Allocation (LDA), where latent topics are intended to represent the aspects and/or sentiments.

The main limitation of all such approaches is that they need to fix a number of latent topics that aren't known a priori. Furthermore, even if an optimal number of topics is found, topics aren't ensured to represent true aspects.

In this paper we attempt to combine the best of these approaches into a single statistical framework. This framework's goal is to discover new features and opinion words starting from a minimal knowledge source, to rank them according to the same statistical principles, and eventually to infer and rank broader aspects that can serve to summarize the discovered features.

3.4 A Language Model of Product Features

Statistical mappings between words have been proposed to model semantic relationships between language models in several fields related to text processing. For example, they've been directly applied to define stochastic translations between words from different source languages in machine translation (Brown et al., 1993), or to reduce the vocabulary gap between documents and queries in ad-hoc information retrieval (Karimzadehgan and Zhai, 2012). Also, in the field of NLP, lexical entailment models between words have been built based on word-to-word statistical mapping models (Glickman et al., 2006).

Following a similar idea, in this paper we propose statistical mapping models between words to retrieve product features and opinions from customer reviews. Specifically, we consider stochastic mappings between words to estimate a unigram language model of product features (for example, a probabilistic model that assigns higher probability values to words defining product features) from a probabilistic model of opinion words (for example, a model that assigns higher probability values to words defining customer opinions, such as "excellent", "awesome", "bad", and "terrible"). Then, we base the retrieval of product features on the estimated unigram model of features.

3.4.1 Mapping Opinions to Product Features

Formally, given a collection of customer reviews $\mathscr{C}$ about one or more products, with vocabulary $\mathscr{V} = \{w_1, \ldots, w_n\}$, and a document D, which can be either an individual customer review or a subcollection of reviews from $\mathscr{C}$ (that is, $D \subseteq \mathscr{C}$), we define our problem to be the retrieval of product features from D, as well as the opinion words ascribed to them.

To address this problem, we assume that reviews provide a latent mapping between words that relates opinions to product features. The rationale is that opinion words are utilized to express sentiments about the features (that is, the opinion targets).

Then, since such a mapping is hidden in the reviews, we consider both

- a stochastic vector $Q = \langle Q(w_1), \ldots, Q(w_n) \rangle^\top$ representing a probability model of opinion words from D; and
- an n-by-n stochastic matrix $\mathscr{T} = \{\mathrm{p}(w_i|w_j)\}_{1 \le i \le n, 1 \le j \le n}$ representing a statistical mapping between words (based on word co-occurrences from local contexts in D) to model a unigram language model of product features as follows:

$$P(w_i) \propto \left(\begin{pmatrix} \mathrm{p}(w_1|w_1) & \cdots & \mathrm{p}(w_1|w_n) \\ \vdots & \ddots & \vdots \\ \mathrm{p}(w_n|w_1) & \cdots & \mathrm{p}(w_n|w_n) \end{pmatrix}^k \cdot \begin{pmatrix} Q(w_1) \\ \vdots \\ Q(w_n) \end{pmatrix} \right)_i \tag{3.1}$$

$$= \left(\mathscr{T}^k \cdot Q \right)_i \tag{3.2}$$

where $P(w_i)$ is the probability that word w_i refers to or is part of a description or name of a relevant aspect (such as a product feature) from D, and $k > 0$ is the number of times mapping $\mathscr{T}$ is applied to the opinion model Q. The idea underlying this formulation is that, by successively applying the co-occurrence based mapping $\mathscr{T}$ to Q ($\mathscr{T}$ is actually a conditional distribution of words), we can capture the application of the hidden latent mapping to define a model of feature words from the opinion model Q.

3.4.2 Refining the Language Model of Product Features

In addition, we consider refining the unigram model P to avoid the assignment of high probability values to meaningless words such as prepositions and conjunctions. The aim is to improve the quality of the generative model of product features as these common words (called *stop words* in Information Retrieval) can bias the retrieval towards features containing them. The refined unigram language model P' is obtained by means of an Expectation-Maximization process aimed at minimizing the cross entropy:

$$-\sum_{i=1}^{n} P(w_i) \log(\lambda P'(w_i) + (1 - \lambda) P_{bg}(w_i)) \tag{3.3}$$

where P_{bg} is a background language model of the source language of the reviews (for example, English).

Thus, starting from initial values of λ and P' –that is λ_0 and P'_0, where $\lambda_0 = 0.5$ and P'_0 is randomly chosen from a small stochastic perturbation of P–, the Expectation-Maximization process iteratively approximates the values of λ and P' until convergence by using the following update equations in the kth iteration:

$$P'_k(w_i) = \frac{P(w_i)Z_{w_i}}{\sum_{j=1}^{n} P(w_j)Z_{w_j}} \tag{3.4}$$

$$\lambda_k = \sum_{i=1}^{n} P(w_i)Z_{w_i} \tag{3.5}$$

where Z_{w_i} is:

$$Z_{w_i} = \frac{\lambda_{k-1}P'_{k-1}(w_i)}{\lambda_{k-1}P'_{k-1}(w_i) + (1-\lambda_{k-1})P_{bg}(w_i)} \tag{3.6}$$

As a result, words with high frequency in the background decrease their probability in the final unigram model P'.

In the experiments described later, we estimate a version of P_{bg} from the Corpus of Contemprorary American English (COCA) (Davies, 2011) to analyze reviews in English, and estimate P_{bg} from word frequencies taken from "SkyDrive de Hermit Dave/WordLists" for analyzing reviews in Spanish.[1]

3.5 Retrieval of Product Features and Opinions

To retrieve the product features from D, we firstly consider generating a set of candidate product features $\mathscr{P} \subseteq \mathscr{V}^+$ by means of an iterative process that follows a bottom-up strategy to include in $\mathscr{P}$ potential features represented by word sequences of arbitrary length ($\mathscr{V}^+$ denotes the set of all possible sequences –that is, phrases– of words from $\mathscr{V}$ with at least one word).

In the ith iteration, the process initially considers as candidate features all possible word sequences with the form $w \cdot s$, where $w \in \mathscr{V}$, s is a word sequence in $\mathscr{P}$ with length $i-1$, and $w \cdot s$ denotes the operation of concatenating word w with word sequence s to form a new word sequence. In the first iteration, when $i = 1$, all words in $\mathscr{V}$ are initially considered as candidate features. Then, it prunes the candidates that fall below a threshold according to functions $H : \mathscr{V}^+ \to \mathbb{R}$ and $G : \mathscr{V}^+ \to \mathbb{R}$, defined as follows:

$$H(w \cdot s) = \eta(w \cdot s)\left(\log P'(w) - \log Q(w)\right) \tag{3.7}$$

$$G(w \cdot s) = \eta(w \cdot s)\left(\log \eta(w \cdot s) - \log\left(\eta(w)\eta(s)\right)\right) \tag{3.8}$$

where $w \in \mathscr{V}$, $s \in \{\epsilon\} \cup \mathscr{V}^+$ (ϵ is the empty sequence), and η is a measure of the likelihood of a sequence of r words $w_{i_r} \cdots w_{i_1}$ in $\mathscr{V}^*$ that we define from the

[1] https://skydrive.live.com/?cid=3732e80b128d016f&id=3732E80B128D016F%213584

mapping $\mathscr{T}$ for all $r \geq 1$ as follows:

$$\eta(w_{i_r} \cdots w_{i_1}) = \mathrm{p}(w_{i_r}) \prod_{j=1}^{r-1} \mathrm{p}(w_{i_j} | w_{i_r}) \tag{3.9}$$

For $r = 0$, we consider $\eta(\epsilon) = 1$. The value of $\mathrm{p}(w_{i_r})$ can be estimated from $\mathscr{T}$ using the following system of $n+1$ linear equations with variables $\{\mathrm{p}(w_1), \ldots, \mathrm{p}(w_n)\}$:

$$\sum_{j=1}^{n} \mathrm{p}(w_i | w_j) \mathrm{p}(w_j) = \mathrm{p}(w_i) \quad (i = 1..n) \tag{3.10}$$

$$\sum_{j=1}^{n} \mathrm{p}(w_j) = 1 \tag{3.11}$$

The aim of H and G is to jointly assess the possibility of combining together a word with a word sequence to compose a new sequence that potentially represents a product feature. Applying H to $w \cdot s$ ($w \in \mathscr{V}, s \in \{\epsilon\} \cup \mathscr{V}^+$) is intended to weight $\eta(w \cdot s)$ with a value that gives an idea of how much word w can be drawn from the unigram model of features P' instead of the opinion model Q. Similarly, $G(w \cdot s)$ weights $\eta(w \cdot s)$ with a value indicating whether the sequence $w \cdot s$ represents a phrase from D.

The iterative process terminates when a predefined number of iterations is reached or when no candidate sequences can be added to $\mathscr{P}$. In the experiments carried out in this work, we consider generating word sequences with at most two words to represent product features. We also consider setting the pruning thresholds for both H and G to 0.

Finally, we retrieve the product features from $\mathscr{P}$ using the following ranking function:

$$F(w_{i_r} \cdots w_{i_1}) = P'(w_{i_1}) \prod_{j=2}^{r} \mathrm{p}(w_{i_j} | w_{i_1})^{\frac{1}{r-1}} \tag{3.12}$$

which assumes that w_{i_1} is the head of the candidate phrase $w_{i_r} \cdots w_{i_1} \in \mathscr{P}$, and $w_{i_r}, \ldots, w_{i_2}$ are modifiers. Function F ranks candidates by combining the probability of their head under the refined language model of product features together with a normalized probability value of the sequence of modifiers conditioned on the head (for simplicity, independence between words is assumed).

We rank sentiment words with respect to the sequence of words $w_{i_r} \cdots w_{i_1}$ using the following scoring function:

$$R(w) = \prod_{t=1}^{r} \mathrm{p}(w_{i_t} | w)^{\frac{1}{r}} Q(w). \tag{3.13}$$

The estimation of both the stochastic mapping $\mathscr{T}$ and the opinion model Q are described next.

3.6 Probabilistic Mapping Model Between Words

To estimate the statistical mapping model $\mathscr{T}$ based on word co-occurrences from local contexts in D, we consider defining $\mathrm{p}(w_i|w_j)$ $(i,j \in \{1,\ldots,n\})$ to be proportional to the number of times word w_i occurs in a local context of words from D containing an occurrence of w_j. Thus,

$$\mathrm{p}(w_i|w_j) = \frac{\mathrm{p}(w_i,w_j)}{\mathrm{p}(w_j)} \tag{3.14}$$

where:

$$\mathrm{p}(w_i,w_j) \propto \sum_{l\in\mathscr{L}} \mathrm{p}(w_i|l)\cdot\mathrm{p}(w_j|l)\cdot\mathrm{p}(l) \tag{3.15}$$

$$\mathrm{p}(w_j) = \sum_{w_i\in\mathscr{V}} \mathrm{p}(w_i,w_j) \tag{3.16}$$

$\mathscr{L}$ is the set of all local contexts of words contained in D, $\mathrm{p}(w_i|l) = |l|_{w_i}/|l|$ and $\mathrm{p}(l) = |\mathscr{L}|^{-1}$ (here, $|l|_{w_i}$ is the number of times w_i occurs in l, and $|l|$ is the number of words contained in l).

We consider two alternatives for defining local contexts. The first one defines local contexts as N-grams occurring in the sentences of D. Our second alternative defines local contexts as the word tuples of D_{Rel}, where D_{Rel} consists of the bag of grammatical dependency relations observed among word occurrences in D.

3.7 A Kernel-Based Model of Opinion Words

We rely on a kernel-based density estimation approach to learn the model of opinion words Q from a (minimal) knowledge source of sentiments or opinions $\mathscr{U} = \{u_1,\ldots,u_m\}$; for example, a seed set of (general-domain) opinion words or concepts from SenticNet (Cambria et al., 2012). Specifically, we estimate Q as follows:

$$Q(w) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{K}(w,u_j) \tag{3.17}$$

where $w \in \mathscr{V}$, and $\mathrm{K}(w,u_j)$ represents the Gaussian kernel:

$$\mathrm{K}(w,u_j) = \exp\left(-0.5\cdot h(g(w),g(u_j))^2/\sigma^2\right) \tag{3.18}$$

such that $g(w)$ denotes the posterior distribution of words conditioned on w $\{\mathrm{p}(w_i|w)\}_{1\leq i\leq n}$ (defined from Equation 3.14). Similarly, for all $j \in \{1,\ldots m\}$, $g(u_j)$ represents a posterior distribution of words conditioned on u_j $\{\hat{\mathrm{p}}(w_i|u_j)\}_{1\leq i\leq n}$. Function h represents a metric between distributions, and σ is a predetermined distribution width. Our hypothesis is that opinion words entail similar conditional distributions of words.

Since h is applied to distributions in the $(n-1)$–simplex, we define h according to a geodesic in that manifold as follows (Dillon et al., 2007):

$$h(g(w),g(u_j)) = \arccos\left(\sum_{i=1}^{n}\sqrt{\mathrm{p}(w_i|w)\hat{\mathrm{p}}(w_i|u_j)}\right) \tag{3.19}$$

Notice that if $\mathscr{U}$ is a seed set of opinion words, we can define $\hat{\mathrm{p}}(w_i|u_j) = \mathrm{p}(w_i|u_j)$. Otherwise, $\hat{\mathrm{p}}(w_i|u_j)$ can be calculated from an estimation of the joint likelihood between words in the vocabulary $\mathscr{V}$ and elements in $\mathscr{U}$. For example, if $\mathscr{U}$ is a seed set of concepts expressing sentiments or emotions, we can define the joint likelihood between word $w_i \in \mathscr{V}$ and concept $u_j \in \mathscr{U}$ to be proportional to the number of times word w_i and concept u_j are "observed" in a local context from D, namely, $\ell(w_i,u_j)$. Thus, we can define $\hat{\mathrm{p}}(w_i|u_j)$ as the ratio $\ell(w_i,u_j)/\sum_{w\in\mathscr{V}}\ell(w,u_j)$.

To define both $g(w)$ and $g(u_j)$, we consider statistics from the whole collection of reviews $\mathscr{C}$, instead of only using D.

3.8 Evaluation

To evaluate our approach, we consider three different datasets that comprise several review collections of products. The first dataset (namely, **10-Products**) contains 10 review collections about ten different products, including digital cameras, phones, routers, and so on. The dataset includes test collections of products (Ding et al., 2008; Hu and Liu, 2004b) and review collections from `http://www2.cs.uic.edu/~liub/FBS/sentiment-analysis.html`.

The second dataset is the Taxonomy-Based Opinion Dataset (**TBOD**) (Cruz et al., 2010), which consists of three review collections about cars, headphones and hotels.

The source language of both 10-Products and TBOD is English. All products in 10-Products are related to the domain of electronics; whereas products in TBOD are from three different industry domains (the automotive, electronics and tourism industry). Both datasets provide manual annotations about opined features.

The third dataset consists of two broad collections of customer reviews about hotels and restaurants from **Hopinion**, which is a Spanish collection of reviews containing more than 18,000 texts retrieved from TripAdvisor. This dataset is available at `http://clic.ub.edu/corpus/es/node/106`. Unfortunately, there is no gold-standard for this dataset that indicates the features and opinion words referred to in the reviews.

Table 3.1 summarizes the details of the three datasets.

To learn the model of opinion words Q for the review collections in the English datasets, we regard as knowledge source the publicly available seed set of around 6800 opinion words based on the lexicon originally proposed by Hu and Liu (2004b). [2] In the case of Spanish, we use the fullStrengthLexicon (a dataset of around 1300 words) (Perez-Rosas et al., 2012). Due to space constraints, and

[2] `http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar`

Table 3.1 Summary of dataset statistics.

Dataset	Review collection	No. of reviews	No. of sentences
10-Products	canon g3	45	641
	canon sd500	1	241
	canon s100	36	248
	nikon coolpix	34	380
	apex ad2600	99	834
	jukebox zen xtra	95	1798
	nokia 6600	49	569
	nokia 6610	40	586
	hitachi router	31	314
	linksys router	48	568
TBOD	cars	943	28213
	headphones	580	8721
	hotels	982	35656
Hopinion	hotels	12412	85210
	restaurants	1779	8415

also because we lack a semantic knowledge source of sentiments or opinions in Spanish, we leave the issue of learning model Q from concepts for future work.

Because our methodology provides rankings of product features and opinion words, we adopt the traditional measure of *Average Precision* (AveP) (Turpin and Scholer, 2006) to evaluate the quality of the retrieval of product features and opinion words. This measure combines both precision and recall factors at each position in a ranking. The higher the measure's values, the better the retrieval is.

3.8.1 Evaluating the Model of Opinion Words

Our first experiment focuses on evaluating the quality of the proposed method to learn the model of opinion words Q, which is the component of our framework that requires training samples.

For this purpose, we consider modeling opinion words from each review collection about a product in 10-Products dataset and each category of products in TBOD (cars, headphones and hotels) and Hopinion (hotels and restaurants) datasets separately. Specifically, for each review collection $\mathscr{C}$ on a product or category, we first obtain the set of opinion words that simultaneously belong to both the opinion lexicon $\mathscr{U}$ and the collection's vocabulary. Let us denote by $\mathscr{U}_{\mathscr{C}}$ this set of opinion words. Then, we randomly sample a uniform-size partition $\mathscr{U}_{\mathscr{C}} = \mathscr{U}_{\mathscr{C},1} \cup \ldots \cup \mathscr{U}_{\mathscr{C},5}$. For each $i \in \{1,\ldots,5\}$, we learn a model Q using $\mathscr{U}_{\mathscr{C},i}$ as seed set, and consider the retrieval of all words in $\mathscr{U}_{\mathscr{C}}$ to evaluate the model.

Table 3.2 shows the averaged values of AveP obtained for each review collection. The label "w5" represents a model obtained by using 5-grams as local contexts of words to build the mapping model $\mathscr{T}$; whereas the labels "drAll" and "drSelected" refer to models that define the local contexts based on gram-

Table 3.2 Values of AveP obtained for different opinion models.

Dataset	Review collection	w5	drAll	drSelected
10-Products	canon g3	0.77	0.66	0.52
	canon sd500	0.81	0.64	0.48
	canon s100	0.75	0.55	0.41
	nikon coolpix	0.77	0.65	0.50
	apex ad2600	0.79	0.61	0.52
	jukebox zen xtra	0.77	0.73	0.49
	nokia 6600	0.76	0.54	0.48
	nokia 6610	0.75	0.61	0.49
	hitachi router	0.77	0.61	0.47
	linksys router	0.81	0.62	0.41
TBOD	cars	0.75	0.68	0.62
	headphones	0.76	0.68	0.60
	hotels	0.77	0.66	0.59
Hopinion	hotels	0.85	0.54	0.46
	restaurants	0.84	0.75	0.32

matical dependency relations. In particular, a model "drAll" is built using all the dependency relations obtained with the Stanford dependency parser (De Marneffe et al., 2006), while "drSelected" considers only the set of relations {"nn", "acomp", "advmod", "amod", "det", "dobj", "infmod", "iobj", "measure", "nsubj", "nsubjpass", "partmod", "prep", "rcmod", "xcomp", "xsubj"}. These results correspond to the models that produce the best overall averaged values (over all collections) obtained by varying the distribution width σ in the interval $[0.1, 0.5]$ with step 0.05. It is worth mentioning that results are stable with respect to σ.

The best performance is achieved using the models based on 5-grams; in particular, the models learned from the Spanish datasets about hotels and restaurants obtained the best values of AveP. It seems that the dependency-based contexts proposed are not sufficient to properly model opinions.

From these results, we assume that opinion words can be successfully modeled using our approach. It's also worth noting that we use a unique seed set of opinion words to learn Q for the different review collections (about products or categories of products from different industry domains) in the same language. This corroborates our claim that the proposed methodology can be applied to collections of reviews in any domain and language.

3.8.2 Evaluating the Retrieval of Product Features

The second experiment is aimed at evaluating the quality of the retrieval of product features. For this purpose, we compare our proposal to both the double-propagation strategy (DP-based) (Qiu et al., 2011) and the approach based on the Hyperklink-Induced Topic Search (HITS-based) algorithm (Zhang et al., 2010). For the former, we rely on our own implementation of the method. For our approach, we consider w5 to define the statistical mapping model; and we

Table 3.3 Values of AveP obtained on the retrieval of product features.

Review collection	**DP-based**	**HITS-based**	**w5**
	collection / review	collection / review	collection / review
canon g3	0.12 / 0.12	0.19 / 0.09	**0.21** / **0.31**
canon sd500	0.09 / 0.09	0.15 / 0.15	**0.18** / **0.18**
canon s100	0.08 / 0.03	0.13 / 0.06	**0.16** / **0.25**
nikon coolpix	0.08 / 0.14	0.16 / 0.12	**0.18** / **0.32**
apex ad2600	0.11 / 0.05	0.18 / 0.01	**0.19** / **0.25**
jukebox zen xtra	0.14 / 0.03	**0.24** / 0.02	0.13 / **0.19**
nokia 6600	0.10 / 0.08	**0.19** / 0.06	0.14 / **0.26**
nokia 6610	0.25 / 0.14	**0.31** / 0.09	0.20 / **0.35**
hitachi router	0.06 / 0.05	0.09 / 0.05	**0.14** / **0.24**
linksys router	0.09 / 0.01	**0.16** / 0.01	0.12 / **0.14**
cars	0.14 / **0.27**	**0.25** / **0.27**	0.18 / 0.26
headphones	0.19 / **0.34**	0.24 / 0.23	**0.25** / 0.26
hotels	0.16 / 0.21	**0.30** / **0.27**	0.23 / 0.24

use a large value to define k ($k = 20$).

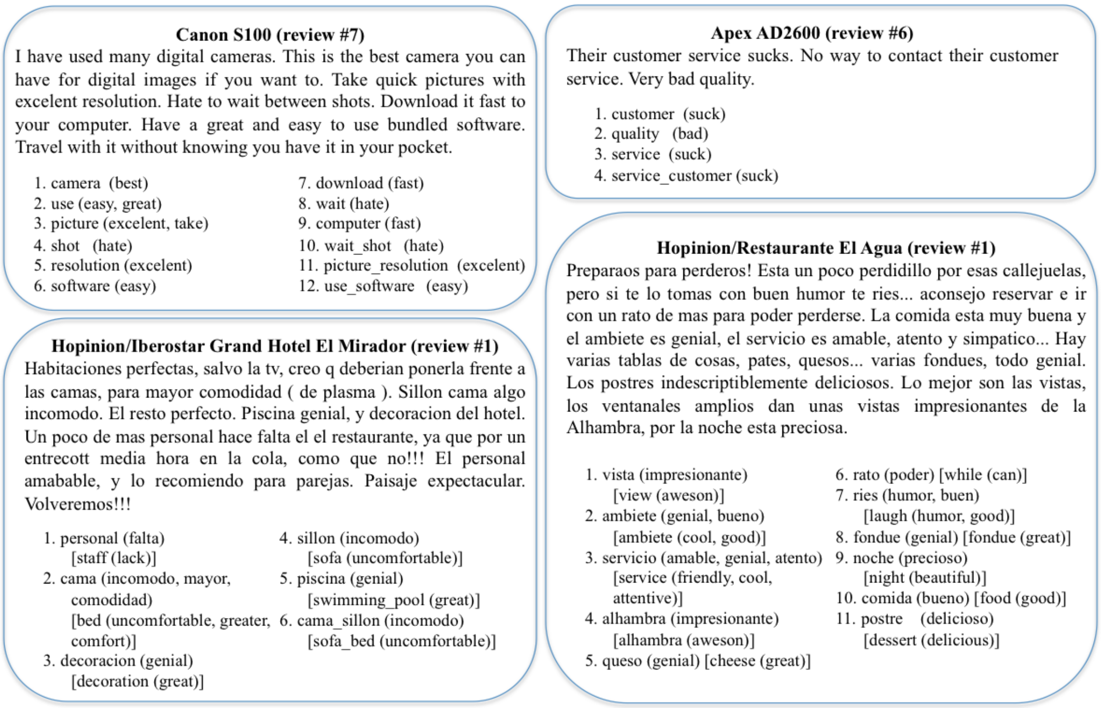
Table 3.3 shows the comparison between the different approaches. The columns labeled "collection" report the values of AveP, considering the retrieval of product features from the whole collection of reviews of each product; whereas the columns labeled "review" average the values obtained in the retrieval of features from the individual customer reviews. We didn't include results from Hopinion, because it doesn't provide a gold-standard.

As can be seen, our proposal clearly outperforms both DP-based and HITS-based in the retrieval from individual customer reviews in the 10-Product dataset. In other cases, our approach competed with HITS to achieve the best performance. Overall, the values of AveP obtained when considering the entire review collection are smaller than those obtained by performing the retrieval over individual customer reviews. This is probably because the gold standard has been annotated at the sentence level, and specific product features in the gold-standard have poor statistical significance in the whole collection of reviews about a product.

We manually inspected our system's results and found that most of the features it retrieves are correct even though not annotated in the gold-standard. Figure 3.1 shows the top-ranked features retrieved using w5 from some customer reviews in the datasets, together with their top-scoring opinion words. As the figure shows, our method can identify the most relevant features and related opinions from both positive and negative opinion reviews in either English or Spanish.

3.8.3 Evaluating the Retrieval of Opinions for Product Features

The third experiment was focused on evaluating the quality of the score function R to retrieve the opinions for each product feature. Thus, we consider retrieving opinions for each manually labeled product feature in TBOD dataset



Canon S100 (review #7)

I have used many digital cameras. This is the best camera you can have for digital images if you want to. Take quick pictures with excelent resolution. Hate to wait between shots. Download it fast to your computer. Have a great and easy to use bundled software. Travel with it without knowing you have it in your pocket.

1. camera (best)
2. use (easy, great)
3. picture (excelent, take)
4. shot (hate)
5. resolution (excelent)
6. software (easy)
7. download (fast)
8. wait (hate)
9. computer (fast)
10. wait_shot (hate)
11. picture_resolution (excelent)
12. use_software (easy)

Hopinion/Iberostar Grand Hotel El Mirador (review #1)

Habitaciones perfectas, salvo la tv, creo q deberian ponerla frente a las camas, para mayor comodidad (de plasma). Sillon cama algo incomodo. El resto perfecto. Piscina genial, y decoracion del hotel. Un poco de mas personal hace falta el el restaurante, ya que por un entrecott media hora en la cola, como que no!!! El personal amabable, y lo recomiendo para parejas. Paisaje expectacular. Volveremos!!!

1. personal (falta) [staff (lack)]
2. cama (incomodo, mayor, comodidad) [bed (uncomfortable, greater, comfort)]
3. decoracion (genial) [decoration (great)]
4. sillon (incomodo) [sofa (uncomfortable)]
5. piscina (genial) [swimming_pool (great)]
6. cama_sillon (incomodo) [sofa_bed (uncomfortable)]

Apex AD2600 (review #6)

Their customer service sucks. No way to contact their customer service. Very bad quality.

1. customer (suck)
2. quality (bad)
3. service (suck)
4. service_customer (suck)

Hopinion/Restaurante El Agua (review #1)

Preparaos para perderos! Esta un poco perdidillo por esas callejuelas, pero si te lo tomas con buen humor te ries... aconsejo reservar e ir con un rato de mas para poder perderse. La comida esta muy buena y el ambiete es genial, el servicio es amable, atento y simpatico... Hay varias tablas de cosas, pates, quesos... varias fondues, todo genial. Los postres indescriptiblemente deliciosos. Lo mejor son las vistas, los ventanales amplios dan unas vistas impresionantes de la Alhambra, por la noche esta preciosa.

1. vista (impresionante) [view (aweson)]
2. ambiete (genial, bueno) [ambiete (cool, good)]
3. servicio (amable, genial, atento) [service (friendly, cool, attentive)]
4. alhambra (impresionante) [alhambra (aweson)]
5. queso (genial) [cheese (great)]
6. rato (poder) [while (can)]
7. ries (humor, buen) [laugh (humor, good)]
8. fondue (genial) [fondue (great)]
9. noche (precioso) [night (beautiful)]
10. comida (bueno) [food (good)]
11. postre (delicioso) [dessert (delicious)]

Fig. 3.1 Top ranked product features and their opinion words retrieved from some reviews in the datasets using model w5.

Table 3.4 Average values of AveP obtained on the opinion retrieval for individual features.

Review collection	**w5**	**drAll**	**drSelected**
	collection / review	collection / review	collection / review
cars	0.65 / 0.53	0.68 / 0.55	**0.69** / **0.56**
headphones	**0.64** / **0.57**	**0.64** / **0.57**	**0.64** / 0.56
hotels	0.63 / 0.56	0.72 / **0.64**	**0.73** / **0.64**

(it's the only one of the three datasets annotated with the opinion words associated to each feature). We compare the obtained opinion ranking to the set of manually annotated opinions for each feature in relation to AveP.

In Table 3.4, we report the averaged results obtained for each model (w5, drAll and drSelected) when applied to a product's entire review collection and individual customer reviews.

In this case, the models based on grammatical dependency relations outperform the model based on 5-grams, despite the fact that the latter achieves better results in the modeling of opinion words. This suggests that grammatical dependency relations better capture the entailment from features to opinion words. Our results also show that the retrieval of opinions is better when using a selected set of dependency relations rather than using all of the dependency relations.

Finally, with respect to speed performance, w5 is faster than drAll and drSelected, since it doesn't need any further text processing.

3.9 Conclusion

In this article, we present a new methodology for the retrieval of product features from a collection of customer reviews about a product or service. The proposed methodology doesn't require any training set of product features, and the experiments carried out over several collections of customer reviews in English and Spanish have shown the proposal's uselfulness for properly retrieving the product features and the opinions expressed about them, even from individual reviews. For future work, we plan to integrate our models into a probabilistic topic modeling framework. The aim is to provide features and opinions with a topic-based description representing them. We also plan to extend our methodology to model the polarity of the opinion words ascribed to product features.

Acknowledgment

This work has been partially funded by the "Ministerio de Economía y Competitividad" with contract number TIN2011-24147 and by the Fundació Caixa Castelló project P1-1B2010-49. Lisette García-Moya has been supported by the PhD Fellowship Program of the Universitat Jaume I (PREDOC/2009/12).

Chapter 4

Storing and analysing voice of the market data in the corporate data warehouse

Lisette García-Moya, Shahad Kudama, María José Aramburu, and Rafael Berlanga.
Information Systems Frontiers 15, no. 3 (2013): 331-349.

4.1 Abstract

Web opinion feeds have become one of the most popular information sources users consult before buying products or contracting services. Negative opinions about a product can have a high impact in its sales figures. As a consequence, companies are more and more concerned about how to integrate opinion data in their business intelligence models so that they can predict sales figures or define new strategic goals. After analysing the requirements of this new application, this paper proposes a multidimensional data model to integrate sentiment data extracted from opinion posts in a traditional corporate data warehouse. Then, a new sentiment data extraction method that applies semantic annotation as a means to facilitate the integration of both types of data is presented. In this method, Wikipedia is used as the main knowledge resource, together with some well-known lexicons of opinion words and other corporate data and metadata stores describing the company products like, for example, technical specifications and user manuals. The resulting information system allows users to perform new analysis tasks by using the traditional OLAP-based data warehouse operators. We have developed a case study over a set of real opinions about digital devices which are offered by a wholesale dealer. Over this case study, the quality of the extracted sentiment data is evaluated, and some query examples that illustrate the potential uses of the integrated model are provided.

4.2 Introduction

Current data warehouse (Inmon, 2005) and OLAP (Codd, 1993) technologies are applied to analyse the structured data that companies store in databases. The context that helps to understand this data over time (e.g., the explanation of a sales fall) is usually described separately in text-rich documents. Some of these documents are self-produced internal company documents (e.g., technology reports), whereas others are available on the Web (e.g., a market-research article in a business journal). Although these documents include highly valuable information that should also be exploited by companies, they cannot be analysed by current OLAP technologies, because they are unstructured and text-rich data (Pérez et al., 2008b).

Documents specially useful in business intelligence (BI) are those expressing some sentiment about a product, service or transaction related to the organization. As pointed out in (Liu et al., 2005), to gather and study the customers' opinion of the products and services of a company is a key task in product benchmarking. This is usually a time-costly and expensive process, done manually by the marketing department of the company. Depending on the reference documents, typical tasks can be classified into two study groups. The studies made on any combination of documents internal to the company (e.g., email, warranty reports, call center logs, and so on), which are usually called *Voice of the Customer* (VoC), and the studies made on documents external to the company (e.g., blogs, forums, tweets, and so on), which are usually called *Voice of the Market* (VoM). Listening to the external market allows setting the strategic direction of a business based on in-depth customer insights, whereas listening to the internal market helps to identify better ways of targeting and retaining customers. As pointed by (Johne, 1994; Reidenbach, 2009), both perspectives are important if the aim is to build long-term competitive advantage. VoC focuses individual customers (and perhaps classes of customers), while VoM is about collective rather than individual voices.

VoM usually involves analysing the opinions published on web sites like blogs and forums which contain opinion posts about some business object (e.g. product, service and so on). The techniques that are employed to analyse opinion text elements are known as *sentiment analysis* (Pang and Lee, 2008), whose main aim consists of extracting the opined features of the business objects along with a score that measures the sentiment degree of the opinion (e.g., satisfaction, orientation, and so on). Sentiment analysis is playing an increasingly relevant role in BI analysis, and some preliminary work is showing a real impact in prediction tasks (Archak et al., 2007). However, as far as we know, there is no approach that fully integrates the extracted sentiment data into the corporate data warehouses where strategic data models are implemented for decision-making.

To build a data warehouse with opinions, we first must analyse the nature of the information sources to integrate, namely: the corporate warehouse and opinion posts. On the one hand, corporate warehouses are well-structured, homogeneous and subject-oriented. They rely on multidimensional data models whose central elements are facts, which consist of dimension and measure values. Broadly speaking, dimensions give the context of certain observations over the corporate warehouse (i.e., measures). This information is efficiently managed and queried by means of OLAP-based techniques. On the other hand, opinion posts are brief text fragments expressing some sentiment about a business object, pointing out the pros and cons, and regarding to different features of the product.

Current state-of-the-art approaches that integrate documents into data warehouses (e.g. EROCS (Bhide et al., 2008) and R-Cubes (Pérez et al., 2007)) are mainly aimed at identifying corporate facts within external documents (e.g., concrete sales, deliveries, and so on). However, opinion documents are not intended to refer to such corporate facts. Instead, opinion posts are intended to evaluate product aspects and features. In this context, the only dimensions shared with the corporate warehouse are those explicitly expressed in the contents and metadata of the posts, namely: the product, location and time dimensions. Moreover, current integration approaches do not regard the extraction of measures usually associated to sentiment data such as ratings, satisfaction and assessment scores. Summarizing, these approaches are more suited to deal with VoC document streams than VoM ones.

As illustrated in Figure 4.1, in this paper we propose a true integration of sentiment data into the corporate data warehouse so that new analysis tasks involving both the company strategic data and the VoM documents can be performed.

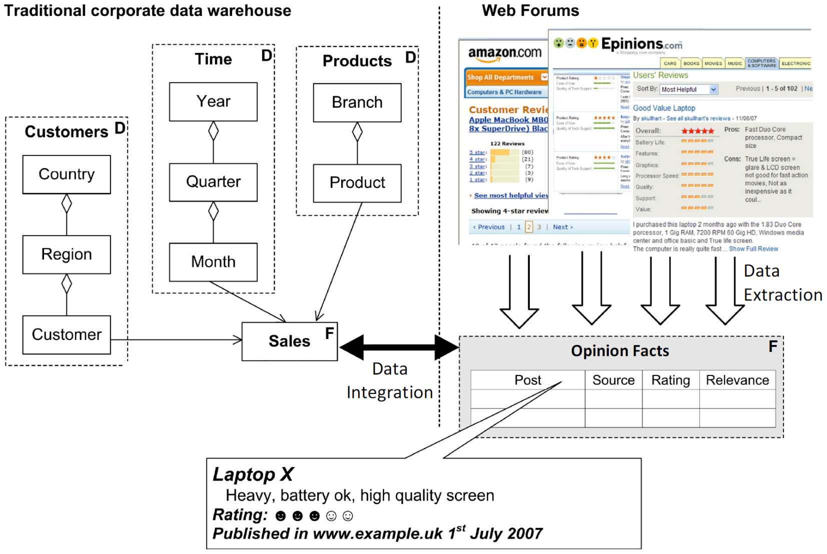


Fig. 4.1 Generic architecture of the proposed integration of corporate and opinion data.

The rest of the paper is organized as follows. Section 4.3 describes the main requirements of the intended information system from the point of view of the opinion posts, their integration into the corporate data warehouse, the analysis of the combined data and the quality of the obtained information. Section 4.4 presents the proposed multidimensional data model which consists of two main parts: the traditional corporate data warehouse and the tables with the sentiment data extracted from the opinion posts. Section 4.5 addresses the problem of sentiment data extraction from opinion posts and presents a new method that applies semantic annotation to facilitate the integration of sentiment and corporate data. Section 4.6 is dedicated to evaluate the results, related work is summarized in Section 4.7, and finally, we give some conclusions and future work in Section 4.8.

4.3 Analysis of requirements

The main aim of our work is to include VoM documents in the set of data warehouse and decision making resources of an organization. We assume the existence of a corporate data warehouse where relevant sales facts are stored and queried through OLAP-based tools. Regarding opinions, they consist of brief text documents that give an evaluation of some product or service. In this section, we analyse the requirements of this application from the point of view of the opinion posts, their integration into the corporate data warehouse, the analysis of the combined data and the quality of the obtained information.

4.3.1 Opinion posts

Opinion posts can be accesed on-line from specialized web forums, usually through web feeds associated to each product being assessed. These data are quite unstructured, mainly consisting of a free text section and a few fields with data about the reviewers and some ratings. More specifically, each opinion typically provides two global scores and a descriptive part with the pros and cons of the evaluated item. The first global score is inserted by the reviewer that has written the opinion and gives a general sentiment about the product (e.g., 5-stars rating). We will denote this score as *product assessment*. The polarity of an opinion (positive, negative or neutral) is directly related to this rating. Our example post in Figure 4.1 has a slightly positive polarity (3 out of 5 stars). The second global score is a summary of the ratings made by the readers of the opinion to indicate how useful it has been for them. Here, we will denote it *opinion assessment*. Notice that this score can be also considered as a global indicator of the quality of the provided information. In the descriptive part of each post, users provide their assessments of the particular features of a product. For example, the post of Figure 4.1 criticizes the weight of the laptop that constitutes a feature opinion with negative polarity, but at the same time praises its battery and the screen, which can be considered as two different product features with positive polarity.

For analysis tasks it is necessary to identify in the opinion texts both *product features* and *product feature assessments*, being a product feature a property that has been assessed by a reviewer when writing an opinion about the product. Product features can be different for each product type and even for each specific product. For example, most digital devices have common features subject to assessment such as weight and battery-life. However, for some particular camera model, reviewers will rate also other features specially relevant for that camera, like its resolution, its capacity, and so on.

It is important to notice that in the set of assessed product features not only participate the product attributes, but also product parts and functionalities. *Product attributes* are well known in advance as they are internally defined by the company as, for example, in the technical specification of a digital device. See Table 4.1 for an example excerpt of the technical specification of a camera consisting in the definition of several groups of product attributes. However, reviewers can assess about other features they consider relevant but that were not regarded in the internal company documents. For example, in an opinion

General	
Body type	Compact
Weight (inc. battery)	481g (17.0oz)
Dimensions (inc. grip)	121 x 74 x 70mm (4.8 x 2.9 x 2.8 in)
Sensor	
Max resolution	2272 x 1704
Image ratio w:h	4:3
Effective pixels	3.9 megapixels
Sensor photo detectors	4.1 megapixels
Sensor size	1/1.8" (7.144 x 5.358 mm)
Sensor type	CCD
Image	
ISO	Auto, 50, 100, 200, 400
White balance presets	6
Custom white balance	Yes
Image stabilization	No
Uncompressed format	RAW
JPEG quality levels	Super-Fine Fine, Normal
Optics and Focus	
Focal length (equiv.)	35 - 140 mm
Optical zoom	4x
Autofocus	Contrast Detect (sensor), Single, Live View
Digital zoom	Yes (x3.6)
Manual focus	Yes
Normal focus range	50 cm (19.69")
Macro focus range	5 cm (1.97")

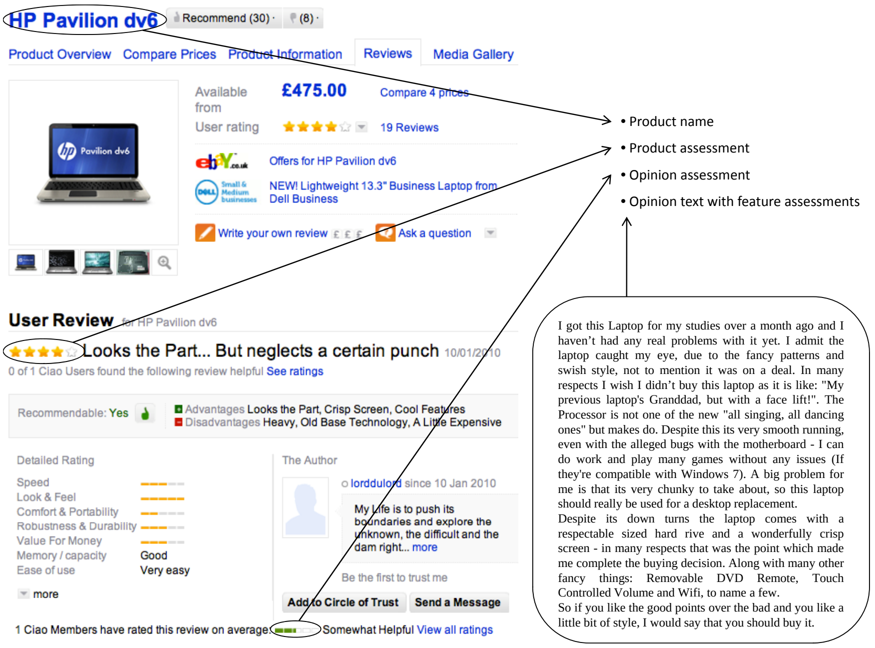
Table 4.1 Excerpt of the product specifications of the Canon PowerShot G3 camera taken from the CNET site.

about a cellular phone users may say that the charger becomes lose and falls out, being the charging system a product attribute that is rarely included in the technical description of a cellular phone. Identifying opined product features different from the internally defined product attributes is an interesting way of discovering unexpected information about the company products. Moreover, it makes possible to analyse the assessments of a product from the point of view of both its predefined attributes and the rest of features assessed by the reviewers.

About the range of product features to be taken into account by a particular system, it is important to note that it changes as new products appear and disappear in the market, and that their relevance also varies depending on both the reviewers and the analysts concerns. For example, ten years ago the battery-life of a cellular phone was not considered a feature so important as today.

4.3.2 Integration issues

Despite the different nature of these information sources, there are some interesting aspects that can enable their integration. Notice that like data warehouses, opinion forums are subject-oriented, and the posts are usually organized into sections where each final section is dedicated to the evaluation of a given product (e.g., a particular laptop). Thus, considering that each opinion post is about a single product clearly indicated by the post section, it will be possible to recognise in each opinion the corresponding product of the corporate files. Furthermore, taking into account the structured part of each opinion post, we can find some interesting measures as the global ratings that can be directly used for estimating its polarity (*product assessment*) and its quality (*opinion assessment*). Figure 4.2 illustrates all these concepts with an example opinion post taken from Ciao![1]. The metadata of the opinion posts also provides useful information to be integrated into the corporate data warehouse. For example, we can use the date and the country from which the post was published as dimension values.


HP Pavilion dv6 Recommend (30) · (8) ·

Available from £475.00 Compare 4 prices

User rating ★★★★☆ 19 Reviews

Offers for HP Pavilion dv6

NEW! Lightweight 13.3" Business Laptop from Dell Business

User Review for HP Pavilion dv6

★★★★☆ Looks the Part... But neglects a certain punch 10/01/2010

0 of 1 Ciao Users found the following review helpful See ratings

Recommendable: Yes

Advantages Looks the Part, Crisp Screen, Cool Features

Disadvantages Heavy, Old Base Technology, A Little Expensive

Detailed Rating

Memory / capacity Good

Ease of use Very easy

The Author

lorddulord since 10 Jan 2010

My Life is to push its boundaries and explore the unknown, the difficult and the dam right... more

Be the first to trust me

1 Ciao Members have rated this review on average: Somewhat Helpful View all ratings

I got this Laptop for my studies over a month ago and I haven’t had any real problems with it yet. I admit the laptop caught my eye, due to the fancy patterns and swish style, not to mention it was on a deal. In many respects I wish I didn’t buy this laptop as it is like: "My previous laptop's Granddad, but with a face lift!". The Processor is not one of the new "all singing, all dancing ones" but makes do. Despite this its very smooth running, even with the alleged bugs with the motherboard - I can do work and play many games without any issues (If they're compatible with Windows 7). A big problem for me is that its very chunky to take about, so this laptop should really be used for a desktop replacement.

Despite its down turns the laptop comes with a respectable sized hard rive and a wonderfully crisp screen - in many respects that was the point which made me complete the buying decision. Along with many other fancy things: Removable DVD Remote, Touch Controlled Volume and Wifi, to name a few.

So if you like the good points over the bad and you like a little bit of style, I would say that you should buy it.

Fig. 4.2 Integration of opinion data from an example post.

To integrate the textual contents of opinion posts into the data warehouse is a difficult task because they are unstructured data. Our purpose is that, after processed, the opinions given by the users in their posts should be in a format able to be analysed by OLAP operators. This can be achieved by extracting structured data from the unstructured texts. The data to be extracted are dimension and fact table values, and the way of integrating them is by identifying in the posts both the dimension values that already appear in the dimension tables of the corporate data warehouse (mainly product, location and time dimension values), and the values of the new dimension and fact tables specifically created to analyse opinions. Undoubtedly, product and opinion assessments, product features, and product feature assessments, as previously described, are values that once extracted from the post texts could be used to integrate sentiment data into the corporate data warehouse.

The extraction of sentiment data from opinion post texts and its integration

[1] `http://www.ciao.co.uk/`

into the corporate data warehouse is a complex task that needs to be automatically made without any assistance. Today, companies produce a large amount of information about their products as manuals, technical specifications, product brochures and catalogues. Some companies have databases to manage the descriptions and attributes of their products. The metadata store of the corporate data warehouse can also be a good place where to find information about the company products. In our opinion, the knowledge provided by all these internal sources can be applied to recognise the product attributes in the opinion texts and to extract the sentiment data to be integrated in the corporate data warehouse. In fact, these sources should not be limited to internal information, as there may be many external sources with new information useful to recognise further opined product features different from the internally defined product attributes.

In this paper we will propose a process that automatically recognises and extracts product feature assessments from opinion posts to be stored in the corporate data warehouse. This process uses a large knowledge base with information coming from several internal and external sources to semantically annotate the opinion posts and, in this way, to extract the data required by the sentiment part of the data warehouse. The group of internal and external knowledge resources needed to implement a particular data warehouse should be defined for each case. In this way, the data warehouse can be customized to analyse the sentiment data about the products of each kind of business. The knowledge included into these resources will determine the set of product features recognised by providing the system with an application specific vocabulary to guide the process of extracting sentiment data from the opinion texts.

4.3.3 Analysis tasks

Our main aim includes the execution of complex queries over the integrated data warehouse to satisfy the information requirements of VoM applications as previously defined. Below, there is a list of example queries that can be involved in a VoM study:

- The list of the best/worst sold products in the company in a determined period of time.
- The list of the best/worst opined features of each product sold.
- The comparative study of the sales values and the users' opinions of each product. This study can show us how important are the opinions and how they affect the sales of each product in the company.

In general terms, it can be said that the users of a VoM application need to execute complex queries to analyse the market opinion of both the company products and the individual features of its products, as well as to study the impact of these opinions on the sales figures of the company.

For analysis purposes, it is necessary to include in the data warehouse an interesting group of categories of analysis useful to study the market opinion of products and product features at several levels of detail and from different points of view. Although most companies have available an internal classification of their products and product attributes, these taxonomies do not always take part of their corporate data warehouses. In order to analyse the market opinion, they should be supported by the system. In our approach, internal information can be applied to obtain a classification of the company products attributes. For example, Table 4.1 presents the technical specification of a camera where the product attributes are grouped into several sections. With this information, it is possible to automatically build a taxonomy to analyse the opinions of the camera attributes at different levels of detail. As pointed out before, recognising the elements of these taxonomies in the opinion texts also helps to extract sentiment data from the opinion posts and to integrate them in the sentiment part of the data warehouse.

Notice that external sources of information can be also used to generate further categories of analysis, different from those defined by the company, but that can also be useful to analyse the market opinion from other points of view. For example, the knowledge included in an electronic commerce application can be used to find out categories of the product features different from those internally defined and that can provide an interesting group of categories of analysis. These taxonomies are external to the company and facilitate a new range of analysis queries.

Depending on the specific knowledge sources used to build a particular data warehouse, the set of available analysis categories will vary. In this way, the system can be adapted to the analysis requirements of a particular application. For example, some companies may only need to analyse the opinions of their products from the point of view of their internal attributes whereas other companies may use a large variety of external knowledge sources and build several taxonomies in order to analyse their products from different points of view.

4.3.4 Quality of data

In the pool of opinions coming into the system, it is possible to find some spam and noisy posts that must be identified and discarded because inserting them into the data warehouse would reduce its global credibility. Spam posts are useless because they do not provide any valuable information. Noisy posts can be defined as those whose assessments are out the average range of opinions about that product. Other kinds of useless posts to be discarded are those that come from some undesirable sources or that refer to a discontinued product. For the rest of posts inserted in the data warehouse, two measures of quality can be defined.

On the one hand, each post comes with an indicator of its usability that is provided by the readers of the web site opinions. Similarly, the product feature assessments of the opinions should have assigned a level of usefulness. In this way, it could be possible to associate a usefulness measure to each product feature assessment fact inserted in the data warehouse and then use it to calculate a global measure of the usefulness for the summarised information provided to the analyst. However, in the available opinion forums, the readers of an opinion are not able to rate the usefulness of each specific assessment made by its reviewer, only the quality of the whole opinion can be scored. To obtain this information, it will be necessary to analyse the distribution of the assessed product features among the posts of each product. The assessments of product features that only appear in the most useful posts can be considered more useful and, conversely, the assessments of product features that only appear in the less useful posts can be considered less useful. For the rest of product features that are frequently assessed in many posts, the quality of each assessment can be estimated as a function of the global quality of the post. It is worth mentioning that some recent work has been proposed to automatically obtain these distributions by applying regression models (Wang et al., 2010a), but this issue goes beyond the scope of this paper.

On the other hand, as previously explained, the relevance that the reviewers of a product and the analysts of a company give to each product feature is different and can change over time. In some cases, the relevance of each product feature for its reviewers can be automatically estimated by the sentiment analysis processes. In other cases, this score could be defined by the analyst within the data warehouse to adapt the answers of the system to the requirements of a specific VoM application.

From our point of view, these two measures of data quality can be applied to select the most valuable opinion posts or to specify some quality conditions to the information to be considered when executing an analysis query.

4.4 Multidimensional data model

Figure 4.3 shows the multidimensional data model that has been designed to solve all the requirements of the application outlined in the previous section. In the figure we distinguish between fact (F), dimension (D) and dimension category (L) tables. There are two main parts in the model: the upper part (cube 1) represents the corporate information, and the lower part (cube 2) represents the sentiment data extracted from opinion posts. The two parts are described in detail as follows:

- The *corporate part* of the data warehouse obtains the data from the internal databases of the company by applying traditional extract, transform and load (ETL) processes. Corporate facts usually involves typical BI measures such as sales, profits, etc.

- The *sentiment part* of the data warehouse has information about the different product reviews from opinion forums. It has two levels of detail: the global opinion about the product (*Opinion* table) and the sentiments about the specific product features mentioned in the opinion post (*OpinionFact* table). With the former one it is possible to obtain summarized information such as the average rating or the best/worst opined products, whereas the latter one provides finer summaries by taking into account the identified product features and their scores ("Feature Assessment"). Unlike the corporate part, this part requires sentiment analysis techniques for being populated. The fact table *Opinion* stores two important measures: "Product Assessment" and "Opinion Assessment", which were described in Section 4.3.1.

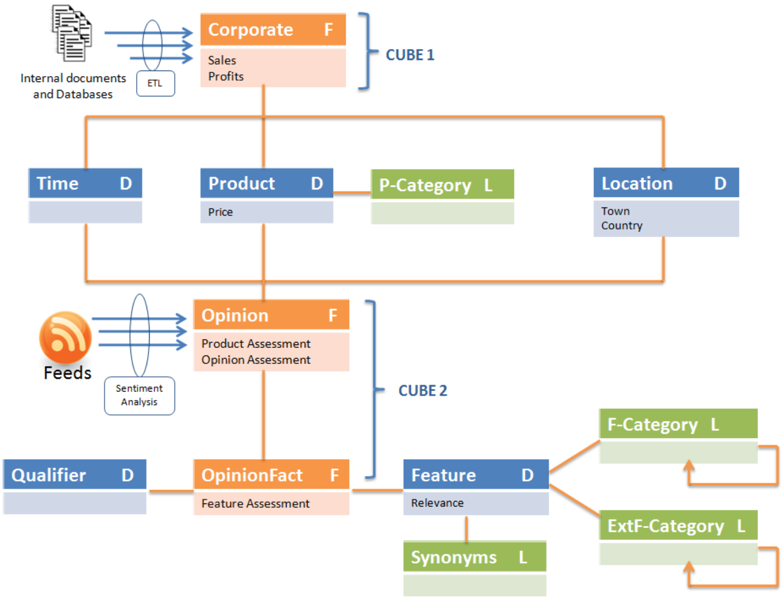


Fig. 4.3 Multidimensional model for a data warehouse integrated with sentiment data.

Notice that the integration between these two parts is achieved through a subset of shared dimensions, in this case: *Time*, *Product*, and *Location*. This makes possible the navigation from one of the cubes to the other one, giving us the opportunity of, for example, studying how the sales and profits of the company are affected by the users' opinions.

In the corporate part of the model, the dimension table *Product* can have associated categories, denoted as *P-Category*, which group the products according to the internal company taxonomies. The analyst can define as many product categories as internal perspectives are required. For example, an analyst can study the sales, profits and users' opinions grouped by products (Canon G3, Nikon, Nokia 6600 and so on), by product families (mobile phones, digital cameras, an so on), and by any other category she considers relevant. For the sake of simplicity, we have just depicted one product category in the schema of Figure 4.3.

In the sentiment part of the model, it can be seen that each opinion fact has associated one qualifier. These are the words that have been used in the text together with the product feature in order to assess it (e.g., **big** camera, **dark** screen, etc.). Storing them in the dimension category table *Qualifier* is interesting to study the range of opinions that have been given about some specific product or product feature. In the following sections, these qualifiers will be denoted *feature indicators* given that they are applied by the sentiment analysis process to identify product feature assessments in the opinion texts.

The dimension table *Feature* is the most challenging dimension of the sentiment part. This table must account for product features subject to assessment. The *Relevance* attribute of this dimension represents the global importance of each product feature according to the reviewers. This attribute is automatically calculated by the sentiment analysis process. As mentioned in Section 4.3.1, product features can stem from either internally defined product attributes or features automatically identified from opinion posts.

In different posts the same feature can be expressed in a great variety of forms, which must be unified and homogeneously represented within the data warehouse. For this reason, we define the dimension category table *Synonyms*, which accounts for all the variants of product features used during the sentiment analysis process.

The *Feature* dimension table has two possible types of analysis categories: *F-Category* and *ExtF-Category*. The first one allows grouping opined product features by the categories extracted from internal knowledge resources as technical specifications and catalogues. With this type of categories the user will be able to analyse the market opinion from the point of view of the internally defined product attributes. The second type of analysis categories, *ExtF-Category*, allows grouping opined product features by the categories extracted from the adopted external knowledge resources (e.g., Wikipedia, e-commerce thesauri, and so on). With these taxonomies, the user can analyse the market opinion from different points of view, external to the company, and without being limited to the categories of product attributes defined by the company itself. Notice that for each particular company, the system must be provided with the knowledge resources needed to build the categories that satisfy the analysis requirements of its specific users and type of business. As a consequence, the group of available categories of analysis will vary from one system to another. For the sake of simplicity, we have just depicted one external category and one internal category for the feature dimension, but the analyst can define an arbitrary number of these categories.

It is worth mentioning that when a new product is launched to the market and it must be regarded in the data warehouse, a new pool of opinions is built from the corresponding feeds. The aim of this pool is to gather enough opinion posts to generate the product features set. Usually this pool is maintained during a few months after the end of the product promotion period, when a burst of opinions is expected, and then the data warehouse will be updated. From that moment, the data warehouse will be periodically refreshed by following the schema defined by the system manager.

Next section discusses how the different tables of the sentiment part of the data warehouse are populated with the data extracted from opinion posts.

4.5 Populating the sentiment part of the data warehouse

After presenting the multidimensional data model of the proposed data warehouse, in this section we describe a novel method to populate the sentiment data tables. This method is aimed at identifying in the opinion streams the most relevant opined product features as well as their associated sentiment scores.

Unlike other approaches to product feature extraction (Pang and Lee, 2008; Zhang et al., 2010), our method applies *semantic annotation* as a means to provide more precise and complete sentiment data to the data warehouse, as well as to integrate the internal and external product-related information. As formerly stated in (Kiryakov et al., 2004), semantic annotation (SA) can be defined as the processing of text elements (e.g. data description fields, free texts chunks, and so on) with the purpose of assigning semantic descriptions from a knowledge resource (KR) to the mentioned entities and, in this way, to reduce the ambiguity present in most natural language expressions. In our case, SA is applied to give a common semantics to the extracted product features and to validate and classify them w.r.t. the KR taxonomies.

Another advantage of SA is the integration of coporate product data into the sentiment data extraction process. Textual descriptions attached to BI objects, such as technical specifications, are semantically annotated with the same knowledge resources than opinion posts. In this way, a direct mapping between BI objects and opinions can be established. In our scenario, these relations are expressed at the product feature dimension.

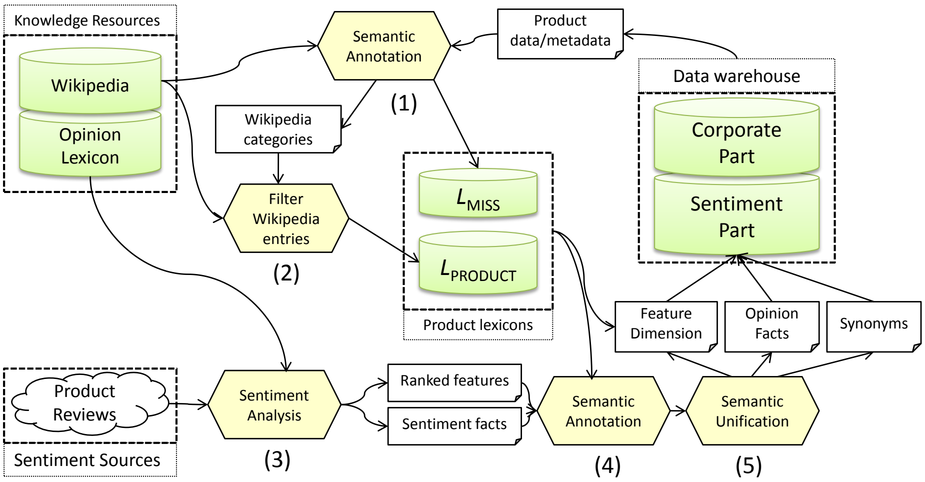


Fig. 4.4 Proposed method for populating a sentiment-aware data warehouse.

The overall method is described in Figure 4.4. It starts by semantically annotating the data and metadata that describe the company products and that are available in the corporate data files (Step 1 in Figure 4.4). Although the proposed method can deal with any KR, in this work we use Wikipedia as the main reference KR. Wikipedia is nowadays the most comprehensive publicly available KR, which includes most technical concepts involved in consumer goods and services. In Step 2, the main categories affecting the company products are manually selected to project the Wikipedia to the corporate domain. The resulting lexicon is called $\mathscr{L}_{PRODUCT}$. A second lexicon, denoted as $\mathscr{L}_{MISS}$, will store the product attributes described in the corporate data files that are not included in Wikipedia, as well as those specific words that clearly denote certain features. For example, the term '4x' is frequently referred in camera reviews denoting the optical zoom feature (see Table 4.1).

In Step 3 of Figure 4.4, opinion posts are collected from the opinion feeds and processed through sentiment analysis to obtain both a ranking of potential product features ordered by relevance, and the set of associated opinion fact table elements (i.e., feature assessments). Afterwards, the information generated by this sentiment analysis phase is semantically annotated with the $\mathscr{L}_{PRODUCT}$ and $\mathscr{L}_{MISS}$ lexicons (Step 4), and then it is processed to unify semantically similar features (Step 5). As a final result, the dimension and fact values required by the sentiment data tables of the data warehouse are obtained. Notice that the output of Step 5 consists of the data that populate the sentiment part of the data warehouse. Notice also how internal and external product information is integrated through semantic annotation processes.

4.5.1 Semantic annotation process

SA can be seen as the process of linking the *entities* mentioned in a text to their *semantic descriptions*, which are usually stored in KRs such as thesauri and domain ontologies (Kiryakov et al., 2004). Former approaches to SA were mainly guided by users (e.g., (Kahan and Koivunen, 2001)) through seed documents, manually tagged examples or ad-hoc extraction patterns. However, in our scenario, we require that the SA process is fully automatic and unsupervised. This is because the volume of data to be processed is huge, and the opined features of a product are unknown a priori. There are few approaches performing fully unsupervised SA, and they are mainly based on dictionary look-up methods or ad-hoc extraction patterns (see (Uren et al., 2006) for a review of SA concepts and approaches).

Most SA approaches assume that the entities are *named entities* (e.g., people, locations, organizations, and so on); hence named entity recognition (NER) is the basic pillar of these approaches. However, the proliferation of comprehensive KRs has extended the notion of entity to any kind of object or concept (Dánger and Berlanga, 2009).

Our SA process consists of three main steps. In the first step, the KR is processed to generate a lexicon, which should contain the lexical variants with which each concept is expressed in the written texts. We denote the set of variants of a concept C as $lex(C)$. For example, $lex(AutoFocus)$ includes the strings "automatic focus", "auto-focus", and "autofocus". The second step consists of applying some *mapping function* between the text chunks likely to contain an entity and the KR's lexicon. Finally, in the third step, the concepts whose lexical forms best fit to each text chunk are selected to generate the corresponding semantic annotation. Figure 4.5 shows an example text that has been semantically annotated with Wikipedia. References to Wikipedia pages are represented with XML tags indicating the page unique identifiers and their main categories (semantic types). For the sake of readability, in this example we have adopted the IeXML notation[2] instead other W3C standard formats like RDFa.

Quality issues

SA has three main issues that affect to the quality of data, namely: entity boundaries, ambiguity, and synonymy. The former issue refers to the fragment of text

[2] http://www.ebi.ac.uk/Rebholz-srv/IeXML/

> *Just a little overview,* <e id="W1469262:Canon_PowerShot_cameras"> powershot g3 </e> *is the flagship of canon's* <e id="W5647263:Canon_PowerShot_cameras"> powershot </e> *series and its an* <e id="W29648:Cameras_by_type"> slr-like camera </e>, *its 4* <e id="W23665:Display_technology"> megapixel < /e> *and (almost) full* <e id="F000008"> manual control </e> *gives the* <e id="W25080:Photography"> pictures </e> *a touch of brilliance.*

Fig. 4.5 Example of semantically annotated text with IeXML. Concept identifiers stemming from Wikipedia start with 'W', and those stemming from the corporate lexicon with 'F'.

that must be designated as being an entity in order to identify its proper meaning. For example, if only *software* is tagged in the chunk "software engineer", the assigned semantics is wrong as the chunk is denoting a person not a feature. Errors due to entity boundaries can affect to the data quality because some sentiments could be associated to wrong features. It is worth mentioning that the issue of entity boundaries is one of the main current challenges of SA (Jimeno-Yepes et al., 2008).

About the second issue, ambiguities occur when the same lexicon entry is associated to more than one concept of the KR. For example, the word *flash* can refer to both the "flash unit" and the "card flash" of a digital camera. Clearly, ambiguous annotations also degrade the quality of data. Disambiguation is one of the most difficult tasks in natural language processing, and few approaches regard it during the SA process (e.g., (Bryl et al., 2010; Mihalcea and Csomai, 2007)).

Finally, the third issue is related to the set of strings that potentially can take part of a concept lexicon. Usually, the KR can provide alternative labels for a given concept. However, written texts like product reviews present a great variety of surface forms like abbreviations, acronyms and short forms, which usually are not regarded in the KR. For example, the '4x' term to refer to a camera zoom. Taking into account such variants is very important for achieving good recall scores. With the purpose of storing and classifying them, the proposed multidimensional model includes the table denoted *Synonyms*.

In our approach, we have addressed these three issues in the following way. For the entity boundary issue, we have applied part-of-speech tagging (POS-tagging) during the sentiment analysis in order to find potential features (see Section 4.5.2). Ambiguity has been greatly reduced by filtering the lexicon to those concepts that are related to the product domain. For example, to annotate reviews of digital cameras we have selected the entries of Wikipedia whose categories are in the photographic domain (i.e., $\mathscr{L}_{PRODUCT}$ in Figure 4.4). Finally, synonymous surface forms are extracted from the Wikipedia redirects. Moreover, the proposed SA system applies a flexible mapping function that allows soft matching between surface forms, which is explained next.

Mapping function

We will assume that the available KRs conform a knowledge base denoted as *KB*, and the text chunks and the strings of the lexicon are represented as bags of words. For each concept C in *KB*, we denote with $C.S_i$ to the i-th string of the concept C within its lexicon $lex(C)$. In order to measure the information overlap between a text chunk T and a concept string $C.S_i$, we apply the following information-theoretic function:

$$sim(T,C.S_i) = \frac{IDF(C.S_i \cap T) - missing(C.S_i,T)}{IDF(C.S_i)} \tag{4.1}$$

Here, the function $missing(S,T)$ accounts for the part of the concept string S that is not covered by T:

$$missing(S,T) = (IDF(S) - IDF(S \cap T)) \tag{4.2}$$

The information amount of a concept string S is measured with the sum of the inverse document frequencies (IDF) of its words, which is estimated over all the Wikipedia articles as follows:

$$IDF(S) = \sum_{w \in S} IDF(w) = -\sum_{w \in S} log(P(w|Wikipedia)) \tag{4.3}$$

We use this formula instead of its common form in order to account for the information lost produced when matching two text chunks. Thus, those words with little IDF produce little information lost when performing partial matchings. Information lost is then the basis of the proposed similarity measure.

Concept ranking

Once we have retrieved all the concepts that potentially fit with the text chunk T, we have to select those that best can participate in the semantic annotation. First, in order to ensure the good quality of annotations, we have to discard those concepts with very low IDF because they are error-prone (e.g., the *line* entry for the *Electronics_stubs* category). Thus, the set of candidate concepts for a text chunk T is:

$$CS(T) = \{C.S | C \in KB, S \in lex(C), S \cap T \neq \emptyset \wedge IDF(S) > \delta\} \tag{4.4}$$

The string concepts in $CS(T)$ can be ranked by several criteria. As we are interested in high quality annotations, our ranking function takes into account the ambiguity and the number of matched words, that is: ambiguous concepts are penalized, and longer matches are rewarded. This function is formally stated as follows:

$$score(C.S,T) = sim(T,C.S) \cdot \frac{|C.S \cap T|}{|\{C' | C' \in CS(T), lex(C') \cap lex(C) \neq \emptyset\}|} \tag{4.5}$$

As an example, Table 4.2 shows the scores returned by this function for different text chunks and concept strings. As the results show, this function is

Text chunk	$C_1.S$	$C_2.S$	$C_3.S$	$C_4.S$
Canon	-0.46	0.61	0.47	-1.0
Canon camera	-0.46	0.05	2.0	0.06
Canon G3	0.40	0.05	-0.04	-1.0
Canon PowerShot G3	3.0	0.05	-0.04	-1.0
G3	0.092	-1.0	-1.0	-1.0
PowerShot G3	0.69	-1.0	-1.0	-1.0
digital camera G3	0.090	-1.0	-0.04	2.0

Table 4.2 Example of resulted scores for $C_1.S$ ='Canon Powershot G3', $C_2.S$ ='Canon Co.', $C_3.S$ ='Canon camera', and $C_4.S$='digital camera' for several text chunks.

able to identify the entity represented by concept C_1 for all the text chunks that clearly refer to the camera "Canon Powershot G3" (see the last five rows of Table 4.2).

From the concept ranking provided by the previous function, we have to select those concepts that best cover the text chunk. More specifically, top concepts whose matched words best cover the text T are selected. For example, from the ranking shown in Table 4.2, the text chunk "digital camera G3" is annotated with the concepts C_4 (for digital and for camera) and C_1 (for G3).

In the next section we discuss how the text chunks are extracted from reviews by applying sentiment analysis. These text chunks will be semantically annotated in order to select those features with precise semantics w.r.t. the selected KR.

4.5.2 Sentiment analysis

The proposed sentiment analysis methodology mainly consists of (i) identifying potential product features in reviews texts, (ii) ranking these features according to their relevance, and (iii) composing opinion facts through product feature assessments. These opinion facts will be used to populate the sentiment part of the data warehouse.

Identifying potential product features and their indicators

The extraction of opinion facts starts by identifying potential features and their indicators in the opinion post texts. Potential product features are identified by means of the extraction patterns shown in Table 4.3. Each pattern is defined in terms of an extended regular expressions over the POS-tagging labels: AJ (adjective), NN (common noun), NNP (proper noun), VBG (gerund verb), VBN (past participle verb), and DT (general determiner). These definitions allow the extraction of both simple and compound noun phrases as potential features.

We are interested on those product features on which customers have expressed their opinions. A *feature indicator* is an opinion word that occurs in the text close to a product feature and that assesses it. Feature indicators will be used to determine product feature assessments.

Name	Pattern	Examples
$NP1$	$(AJ\|NN\|NNP)+$	battery life lcd screen
$NP2$	$NP1\ (VBG\|VBN)\ NP1$	battery charging system
$PF1$	$(NP1\|NP2)$	
$PF2$	$PF1\ (of\|from\|in)\ (DT)?\ PF1$	quality of photos

Table 4.3 Extraction patterns for identifying potential features.

To build a lexicon of opinion words useful for our application, we have used the list of positive and negative opinion and sentiment words for English (around 6800 words) compiled by Hu and Liu (2004b). In addition, we have obtained a list of *polarity shifters terms*, also known as *valence shifters*, from the negative category of the General Inquirer (Stone et al., 1966a). When these terms are used before any opinion word, they change its semantic orientation, turning a negative term into a positive one (e.g., "With the automatic settings, I really **haven't** taken a *bad* picture yet"). Examples of valence shifters are: not, never, none and nobody. It is important to mention that, in a sentence, the polarity of a word is forced into the opposite class if a valence shifter is up to three words before it.

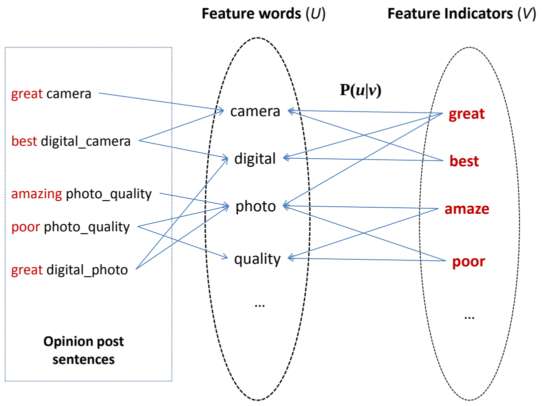


Fig. 4.6 Example bipartite graph of product features and feature indicators.

Features ranking

Once the set of potential product features has been identified from opinion posts, they are ranked according to their relevance (i.e., the attribute *relevance* of the *Feature* dimension in Figure 4.3). This relevance will depend on the number of sentiments each feature is receiving in the opinion posts. Therefore, the ranking should be based on the influence of the indicators over the features. With this purpose, we build a bipartite directed graph $G = (U, V, E)$ representing the relationships between feature words (U) and feature indicators (V). There is an edge $e = (u, v)$ between $u \in U$ and $v \in V$ if they co-occur in some sentence of the analysed opinion posts. Edges $(u,v) \in E$ are weighted by the conditional probability $P(u|v)$, which is estimated as follows:

$$P(u|v) = \frac{P(u,v)}{P(v)} \tag{4.6}$$

where

$$P(u,v) \propto \sum_{W \in \mathscr{W}} p(u|W) \cdot p(v|W) \cdot p(W), \tag{4.7}$$

$$P(v) = \sum_{u \in U} P(v,u), \tag{4.8}$$

and $\mathscr{W}$ is the set of all possible word windows of size k that can be formed in each sentence from the customer reviews. In the experiment carried out in this paper the best performance is achieved using $k = 5$. These probabilities are estimated using $p(u|W) = |W|_u / |W|$ and $p(w) = |\mathscr{W}|^{-1}$, where $|W|_u$ is the number of times u occurs in window W, $|W|$ is the length of W, and $|\mathscr{W}|$ is the cardinality of the set $\mathscr{W}$.

Figure 4.6 shows an example of a bipartite directed graph constructed from some opinion sentences. Following the idea proposed by Zhang et al. (2010), we have applied the HITS algorithm over this bipartite graph for obtaining the feature relevances. Basically, HITS assigns two scores for each vertex: its authority, which estimates the value of the content of the vertex, and its hub value, which estimates the value of its links to other vertexes. We apply the HITS algorithm in our scenario with the hypothesis that a highly-relevant feature word should have high authority score.

Consequently, the relevance score of a potential feature $f = w_1 \ldots w_n$ is calculated from the HITS authority scores of its words w_i as follows:

$$relevance(f) = \sqrt[n]{\prod_{i=1}^{n} authority(w_i)} \tag{4.9}$$

This expression is inspired in the language models proposed for information retrieval based on word translations (Berger and Lafferty, 1999).

Composing opinion facts

An opinion fact consists of a pair (f,q), where f is a potential feature, and q is the list of feature indicators that co-occur with that feature in an opinion post (denoted qualifiers in the multidimensional data model). Given an opinion fact $OF = (f,q)$, we denote f with the function $feature(OF)$ and q with $indicators(OF)$. A feature indicator FI is represented as a pair (o,p), where o is an opinion word, denoted with $opinion(FI)$, and p is its polarity ($p \in \{+1,-1\}$, denoted with $polarity(FI)$). This polarity value takes into account the possible polarity shifters. We assume that opinion words have always the same polarity, that is, they are not context-dependent.

Finally, given an opinion fact $OF = (f,q)$, its assessment is calculated as fol-

Features	Indicators	Relev.	Assess.	Sem. Annot.
camera	best(+), fantastic(+), fun(+), **like(-)**...	0.101	1.185	W52648 (*Camera*)
digital camera	well(+), drawback(-), first-class(+)...	0.029	0.750	W3616597 (*Digital_cameras*)
photo	beatiful(+), great(+), spectacular(+)...	0.006	0.484	W25080 (*Photography*)
use	comfortable(+), great(+), ease(+), easy(+)...	0.003	0.002	F00004 (*General*)
photo quality	amaze(+), ideal(+)	0.019	0.341	W12448204 (*Computer_graphics*)
flash unit	sophisticated(+), lose(-), delicate(+)...	0.011	0.376	W195752 (*Flash_photography*)
design	flaw(-), **correct(-)**, beatiful(+)	0.010	-0.983	F00001 (*General*)

Table 4.4 Examples of the potential features extracted with our method. Semantic annotations generated from the Wikipedia starts with 'W', and those generated from $\mathscr{L}_{MISS}$ starts with 'F'. Indicators in bold face have had their polarities shifted.

lows:

$$assessment(OF) = \sum_{o \in indicators(OF)} p(feature(OF)|opinion(o)) \cdot polarity(o) \quad (4.10)$$

In other words, the assessment of an opinion fact is the weighted sum of the sentiment polarities expressed over the feature. The weight of each indicator w.r.t the feature is its conditional probability. Notice that this makes the approach less sensitive to uncommon sentiments since the conditional probability $p(feature(OF)|opinion(o))$ is inversely proportional to the occurrence frequency of the sentiment $opinion(o)$.

Table 4.4 shows an example set of extracted opinion facts from the reviews of a digital camera (two first columns). It also shows the semantic annotations associated to the features (last column) and the feature relevance and assessment scores.

4.5.3 Feature unification

Before concluding the population phase of the sentiment part of the data warehouse, we must unify those extracted features that are semantically similar. Semantic annotations provide us with a straightforward way to perform this step, as it only requires to group the features according to their associated concepts.

The main issues to be considered when performing feature unification through semantic annotation are the treatment of ambiguous, vague and incomplete annotations. In our approach, ambiguous annotations are directly rejected, as we do not perform any kind of disambiguation processing. Fortunately, thanks to the lexicon filtering previously described, there are very few ambiguous cases in the annotated features.

Vague annotations are those that refer to some abstract category, like *Light* or *Optics*, which are too general and provide little information to the analysts. The corresponding features are not stored in the data warehouse, but their sentiment scores are distributed across all the precise annotations that have them as categories. For example, the score of *Optics* is distributed among *Filter*, *Lense*, *Focus*, etc. In this way, we properly estimate the contribution of each specific annotation.

Incomplete annotations are those that do not cover some words of the text chunk. In this case, it is very important to check if the non-tagged part is relevant for the underlying entity. For this purpose we make use of the *IDF* measure to filter out incomplete annotations. Thus, all the incomplete annotations whose missing part has an *IDF* greater than a given threshold are rejected. In our experiments, there are very few cases of incomplete annotations.

Apart from grouping features by their annotations, we also distribute the sentiment scores of other untagged features whose words occur in the tagged ones. This is a simple way to perform co-reference resolution, because untagged features usually are single words that have been used in the text as feature references. For example, it is usual to find "control" (untagged) as a co-reference of "manual control" (tagged). In absence of further information, the score of an untagged feature is uniformly distributed across all the tagged features that contain it.

Finally, it is worth mentioning that, thanks to the SA process, it is possible to extract from the KR the necessary taxonomies for defining the *Feature* dimension. As an example, in Figure 4.7 we show the extracted taxonomy for the digital cameras domain taking into account the category relationships provided by Wikipedia. In our implementation, this taxonomy has been used to populate the *ExtF-Category* table of Figure 4.3.

4.6 Evaluation of results

We have developed a full prototype of the presented method. Both the semantic annotator and the sentiment analysis method have been implemented in Python and the library NetworkX used for the HITS algorithm. POS-tagging of the review texts has been performed with the Standford NLP parser[3]. As back-end, we have implemented the data warehouse in the SQLServer Business Intelligence Studio. Corporate data elements have been synthetically generated (e.g., sale figures w.r.t. locations and dates), and sentiment data elements have been obtained from well-known opinion web sites. Next sections include the evaluation of the sentiment data extraction method and some interesting complex

[3] `http://nlp.stanford.edu/software/`

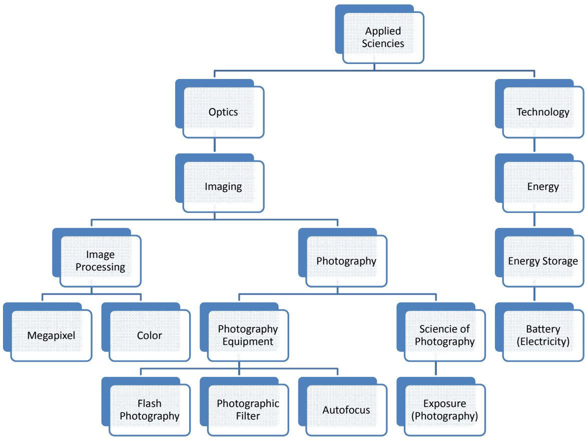


Fig. 4.7 Example of subgraph obtained from Wikipedia for the digital cameras domain.

queries over the resulting data warehouse.

4.6.1 Quality of extracted sentiment data

The quality of the extracted data has been measured by means of a golden standard (GS) consisting of a dataset with opinions collected from Amazon.net and CNET.com [4]. This GS has been widely adopted in many sentiment analysis approaches of the literature, and it consists of reviews about devices such as digital cameras, DVD players and mobile phones (see Table 4.5 for the list of products). Each review includes manual annotations of the features and polarities derived from its text.

After selecting the digital camera reviews from the GS dataset, we collected the corresponding technical specifications from the CNET site in order to annotate them with the whole Wikipedia (XML 2007 snapshot (ISLA, 2010)), which contains more than 5 million entries. Then, the resulting annotations were checked and 62 Wikipedia categories were used to build the $\mathscr{L}_{PRODUCT}$ lexicon of 4739 concepts and 13998 strings. With the untagged attributes of the technical specifications, we built the $\mathscr{L}_{MISS}$ lexicon of 22 concepts and 143 strings.

By means of the $\mathscr{L}_{PRODUCT}$ lexicon, we annotated again the technical specifications in order to measure the precision of the annotations, which was about 93%. The recall was about 77%, that is, 33% of technical attributes were not annotated with $\mathscr{L}_{PRODUCT}$. For the reviews texts, the annotator performed with a precision of 89%, and for the extracted features the precision was about 92%, and the recall about 86%. In all cases, we just kept annotations whose final score was greater than 0.4 and whose *IDF* was greater than 3 ($\delta > 3$).

In order to evaluate the quality of the feature ranking function, Table 4.5 shows the precision values of the extracted potential features w.r.t. the GS at different cut points (i.e. number of correct features present in the selected ones).

[4] This dataset is available at `http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip` and `http://www.cs.uic.edu/~liub/FBS/Reviews-9-products.rar`

Products	@10	@20	@30	@40	@50
Canon G3	100.0	95.0	93.3	92.5	92.0
Canon SD500	100.0	95.0	86.7	82.5	72.0
Canon S100	100.0	90.0	83.3	82.5	78.0
Nikon CP4300	100.0	90.0	76.7	65.0	66.0
Nokia 6610	100.0	100.0	100.0	92.5	88.0
Apex AD260	100.0	100.0	100.0	97.5	96.0
Micro MP3	90.0	85.0	66.7	65.0	64.0
Creative	80.0	75.0	80.0	82.5	80.0

Table 4.5 Precision @N for identified product features in product reviews.

Products	Extracted Features			Annotated Features		
	P	R	F1	P	R	F1
Canon G3	55.0	84.9	67.2	92.0	80.9	86.1
Canon SD500	42.6	71.9	53.5	96.0	84.0	89.0
Canon S100	53.0	67.6	59.4	95.0	87.8	91.2
Nikon CP4300	38.6	90.2	54.1	97.0	93.0	94.0

Table 4.6 Quality of the annotated and extracted features.

Notice that top ranked potential features are more likely to be true features. However, it can also be seen that for some products the precision degrades when some specific features that occur less frequently are included.

Table 4.6 summarises the quality of the extracted data with the precision (P), recall (R) and its harmonic mean (F1). In this case, we evaluated the whole set of extracted features w.r.t. the GS, and the subset of extracted features that were annotated. Notice the outstanding improvement of the measures for all the products. This fact indicates that semantic annotation is discarding noisy and incomplete data elements that were generated by the sentiment analysis method.

Finally, we have also evaluated the calculated assessments for the opinion facts w.r.t. the GS assessments. Table 4.7 reports the precision values of this comparison. Here, precision regards the number of times the obtained polarity of our method coincides with that of the GS for the same features, divided by the number of evaluated facts. Although in general, the obtained precision is acceptable, there is still room for improvement. However, checking the cases where polarities are different, we notice that in the GS many manual annotations are subjective interpretations of the sentences where the feature were mentioned, but there are no opinion words affecting them. For example, for the sentence "the lens retracts and has its own metal cover so you don't need to fuss with a lens cap", the GS assigns the assessment to the feature *lens*, whereas our method assigns it to *lens_cap*. In another review of the same camera, the metal cover is referred as "two-piece shutter-like cap", which is assigned to *lens_cap* by our method but disregarded in the GS. Such a kind of subjective details makes very difficult a precise evaluation of the assessment quality. Another important remark is that thanks to the semantic annotations we have been able to distinguish the features of the analysed product from those of other products. For

Digital Cameras	Assessment Precision
Canon G3	0.78
Canon SD500	0.87
Canon S100	0.64
Nikon CP4300	0.75

Table 4.7 Precision values of the calculated feature assessment w.r.t. the GS assessments.

example, the feature “Nikon zoom” is not a feature of a Canon camera.

4.6.2 MDX queries

The proposed data warehouse model enables complex queries that may be of particular interest to business managers, featuring the maximum profit of the data handled by the enterprise and the information found on the Internet, and helping them in the analysis tasks involved in Voice of the Market studies. We have implemented a prototype of our data model in the SQLServer Business Intelligence Studio. In order to show the usefulness of our approach, in this section we present a few examples of complex queries and the results returned by this prototype.

As an example, Table 4.8 shows the sum of scores of the opined features (rows) for the different products offered by the company (columns). With this information the user can know which are the most commented features of each product and the opinion that users have about them. Notice that although all the products are cameras, most of the listed features are different for each concrete model. Furthermore, considering the Canon G3 model, in the features ranking we can find some features that do not appear in the technical specification of the product (see Table 4.1) as for example the accessories. The results shown in Table 4.8 correspond to the MDX query:

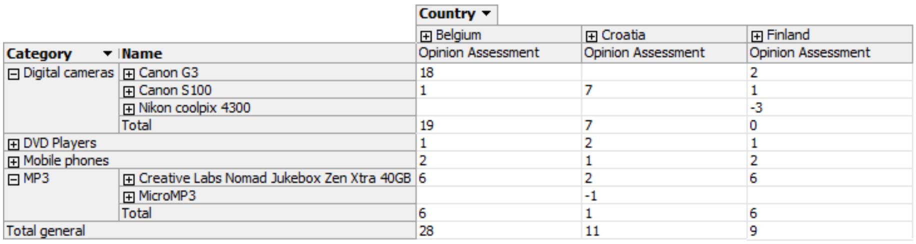
```
SELECT NON EMPTY
 {[Measures].[Feature Assessment]} ON COLUMNS,
 NON EMPTY { ( {
  [Product].[Name].[Canon G3],
  [Product].[Name].[Canon S100],
  [Product].[Name].[Nikon coolpix 4300],
  [Product].[Name].[Canon PowerShot SD500] }
  * [Feature].[Feature].ALLMEMBERS ) } ON ROWS
FROM [Project]
```

In Figure 4.8, we can see a screenshot fragment of a query result. It is completely navigable through the different values of the dimensions. We have expanded the category groups “Digital Cameras” and “MP3”. If we expand one of the products, all the features that have been mentioned in the opinions referred to that product will appear. For the columns we have selected the location dimension to analyse the opinions about the company products that are

Canon G3		Canon S100		Canon SD500		Nikon CP4300	
Camera	0.52	Camera	0.82	Camera	2.29	Camera	0.71
Photo	0.46	Camera body	-0.87	Picture	0.40	Camera lens	1.00
Price	1.00	Size	-0.5	Video-camara	1.00	Picture	0.45
Image quality	0.66	LCD screen	0.00	Image quality	0.10	Usage	0.90
Performance	0.52	Image quality	0.71	Light	-0.03	Resolution	0.86
Flash	0.58	Battery	0.66	Video quality	0.36	Accesories	1.00
Accessories	0.98	Softw. package	1.00	Size	0.06	Scene modes	0.61

Table 4.8 Ranking of the most commented features for each digital camera model. The aggregated assessment of each feature appears next to it.

being promoted at different countries. In the body part of the table, there are represented the sum of the values of the opinion assessments.



		Country ▾		
		⊞ Belgium	⊞ Croatia	⊞ Finland
Category ▾	Name	Opinion Assessment	Opinion Assessment	Opinion Assessment
⊟ Digital cameras	⊞ Canon G3	18		2
	⊞ Canon S100	1	7	1
	⊞ Nikon coolpix 4300			-3
	Total	19	7	0
⊞ DVD Players		1	2	1
⊞ Mobile phones		2	1	2
⊟ MP3	⊞ Creative Labs Nomad Jukebox Zen Xtra 40GB	6	2	6
	⊞ MicroMP3		-1	
	Total	6	1	6
Total general		28	11	9

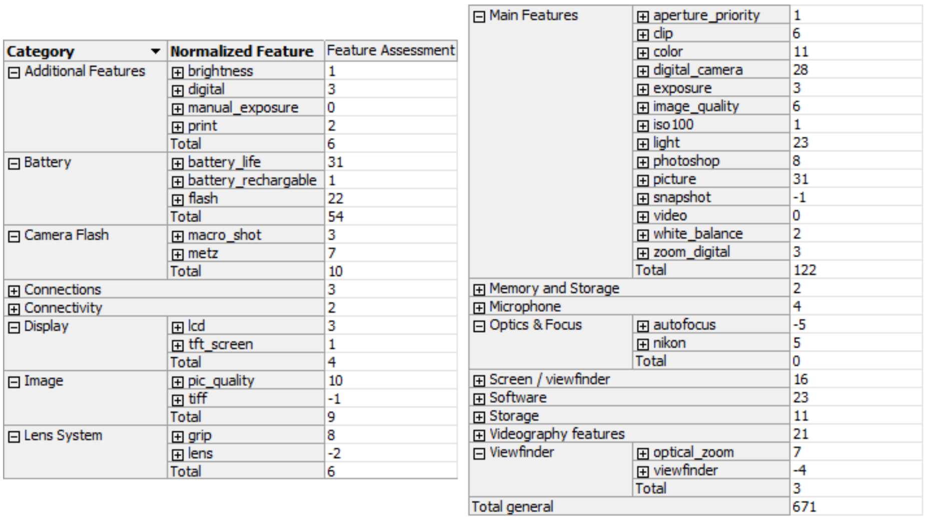
Fig. 4.8 Example of navigable cube with products and locations as dimensions and the aggregated assessment as measure.

In Figure 4.9 we can see a screenshot of a query that presents the normalised form of the opined product features grouped by the categories defined in the technical specifications of the product. For each feature, the assessment value is shown. Most of the extracted features (70%) were not included in the product technical specifications, therefore they have been manually assigned to the corporate internal categories. This example shows the usefulness of the implemented prototype to discover sentiment data on product properties that are not present in the corporate files.

Finally, in Figure 4.10, a query example that combines corporate and sentiment data is presented. In the rows, we can see the products organized in categories and in the columns we can see the time dimension. If we expand a year, all its months will be shown. The query result presents two measures: the total sales and the sum of the product assessments. With this information, the user can analyse how the opinions have influenced on the sales of each product. This is a good example to see how corporate and sentiment data can be shown together in order to give decision support to the companies. Other useful information like the product promotions, which are usually regarded at the corporate part of the data warehouse, could be also easily combined with the sentiment data in order to analyse their impact in both sales and opinions.

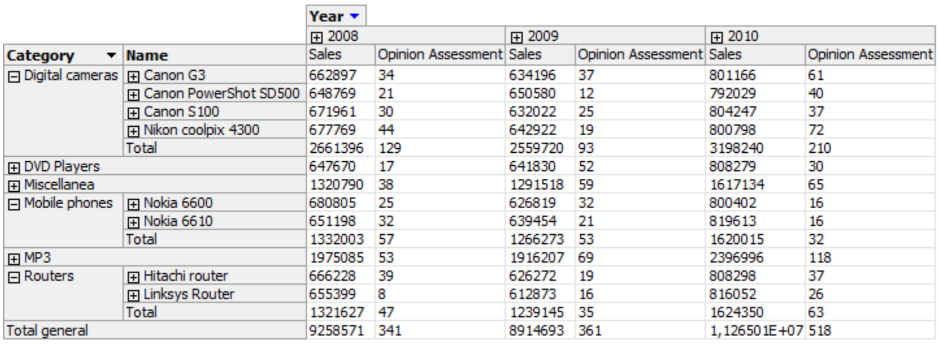
4.7 Related Work

The problem of how to exploit web opinions to extract data that could be useful for BI applications and how to integrate the extracted opinion data into the existing corporate data warehouse is still an open issue. In (Pérez et al., 2008a), we



Category	Normalized Feature	Feature Assessment
⊟ Additional Features	⊞ brightness	1
	⊞ digital	3
	⊞ manual_exposure	0
	⊞ print	2
	Total	6
⊟ Battery	⊞ battery_life	31
	⊞ battery_rechargable	1
	⊞ flash	22
	Total	54
⊟ Camera Flash	⊞ macro_shot	3
	⊞ metz	7
	Total	10
⊞ Connections		3
⊞ Connectivity		2
⊟ Display	⊞ lcd	3
	⊞ tft_screen	1
	Total	4
⊟ Image	⊞ pic_quality	10
	⊞ tiff	-1
	Total	9
⊟ Lens System	⊞ grip	8
	⊞ lens	-2
	Total	6
⊟ Main Features	⊞ aperture_priority	1
	⊞ clip	6
	⊞ color	11
	⊞ digital_camera	28
	⊞ exposure	3
	⊞ image_quality	6
	⊞ iso100	1
	⊞ light	23
	⊞ photoshop	8
	⊞ picture	31
	⊞ snapshot	-1
	⊞ video	0
	⊞ white_balance	2
	⊞ zoom_digital	3
	Total	122
⊞ Memory and Storage		2
⊞ Microphone		4
⊟ Optics & Focus	⊞ autofocus	-5
	⊞ nikon	5
	Total	0
⊞ Screen / viewfinder		16
⊞ Software		23
⊞ Storage		11
⊞ Videography features		21
⊟ Viewfinder	⊞ optical_zoom	7
	⊞ viewfinder	-4
	Total	3
Total general		671

Fig. 4.9 Example of a navigable query where features are grouped by corporate specifications. The aggregated assessment of each feature is shown as measure of the query.



		Year					
		⊞ 2008		⊞ 2009		⊞ 2010	
Category	Name	Sales	Opinion Assessment	Sales	Opinion Assessment	Sales	Opinion Assessment
⊟ Digital cameras	⊞ Canon G3	662897	34	634196	37	801166	61
	⊞ Canon PowerShot SD500	648769	21	650580	12	792029	40
	⊞ Canon S100	671961	30	632022	25	804247	37
	⊞ Nikon coolpix 4300	677769	44	642922	19	800798	72
	Total	2661396	129	2559720	93	3198240	210
⊞ DVD Players		647670	17	641830	52	808279	30
⊞ Miscellanea		1320790	38	1291518	59	1617134	65
⊟ Mobile phones	⊞ Nokia 6600	680805	25	626819	32	800402	16
	⊞ Nokia 6610	651198	32	639454	21	819613	16
	Total	1332003	57	1266273	53	1620015	32
⊞ MP3		1975085	53	1916207	69	2396996	118
⊟ Routers	⊞ Hitachi router	666228	39	626272	19	808298	37
	⊞ Linksys Router	655399	8	612873	16	816052	26
	Total	1321627	47	1239145	35	1624350	63
Total general		9258571	341	8914693	361	1,126501E+07	518

Fig. 4.10 Example of navigable cube with products and dates as dimensions, and the aggregated product assessment and the total sales as measures.

proposed a first approach to this problem by contextualizing a sales data warehouse with on-line customer reviews about the company products/services. A contextualized warehouse (Pérez et al., 2008b) is a new kind of decision support system that allows users to obtain strategic information by combining all their sources of structured data and documents. The analysis cubes of a contextualized warehouse, denoted R-cubes (Pérez et al., 2007), are special since each fact is linked to an ordered list of documents. These documents provide information related to the fact (i.e., they describe the context of the fact). The position of each document in the ranking depicts the relevance of the document for the corresponding fact.

Similarly, the EROCS (Bhide et al., 2008) system basically constructs a link table between data warehouse facts and external documents. NER techniques are used to identify fact dimension values within document texts, and then valid combinations likely to represent facts are extracted to define the fact-document links. Authors propose to classify the linked documents according to some sentiment classes (e.g., satisfaction degree) and then calculate for each fact a global score. However, no concrete implementation of this proposal is given in (Bhide et al., 2008), nor how sentiment data have to be represented and aggregated within the data warehouse.

As mentioned in the introduction, these systems are more appropriate for the Voice of Customer studies, where documents are likely to refer to corporate facts (e.g., concrete sales, deliveries, and so on). Instead, opinion posts basically involve product features and aspects, which are required to be analysed in Voice of the Market studies but that cannot be regarded a priori in the data warehouse because they emerge from anonymous reviewers.

Other related work was presented in (Funk et al., 2008), which proposes to manually assign to each opinion text a qualitative category from an ontology specially developed for BI applications. The resulting annotated corpora would be translated into RDF statements to be inserted into a shared knowledge base and used by applications to track the evolution of business entities. In this paper, the authors only deal with the first steps of the process, in which they rank each opinion text into a five-way classification.

A different approach to the problem of analysing customers' opinions for business decisions is to include the extracted data into predictive econometric models. For example, the work in (Archak et al., 2007) shows that text opinions can be applied to improve product sales prediction compared to a baseline technique that simply relies on numeric data. By processing customer opinions and assigning an importance to the extracted features, the authors of this work propose an econometric model that allows the identification of the weights of qualitative features in determining the overall price of a product. Similarly, the work in (Liu et al., 2007) applied the sentiment-data extracted from the opinions to build an autoregressive model for predicting product sales performance. Although these econometric models can take part of the BI suite of a company, their development is complex and they can not be used to give rapid answers to the full range of ad-hoc queries that different users of a BI application require.

In the approach here presented, the integration of unstructured post data into the corporate data warehouse requires to extract useful information from the posts and turn it into structured data that can be stored and analysed along with the rest of structured data. Moreover, the extracted data have to be related to the corporate data for which entity resolution and classification techniques need to be applied (Berry and Castellanos, 2007; Elmagarmid et al., 2007). In our approach, we have relied on the semantic annotation of the opinion texts by using a very large lexicon extracted from Wikipedia. In order to minimize the ambiguity of the generated annotations, the Wikipedia is projected to the specific domain of the corporate warehouse following the method recently proposed in (Kudama et al., 2011a).

An important issue in this process is the extraction of sentiment data from web opinions. The extensive use of Web 2.0 technologies to produce online opinion data has motivated many research projects in this direction, being the extraction of opined product features from customer reviews an active research area (Pang and Lee, 2008; Zhang et al., 2010). A crucial part of this task is the construction of a sentiment lexicon that can be used to identify product features with their associated polarity. Although in our work, the sentiment lexicon is manually built, there are some proposals to automate this process. A complete review can be found in (Lu et al., 2011), where an optimization approach to the automatic construction of a sentiment lexicon from an unlabeled collection of reviews is proposed. The main novelty of this work is that the resulting lexicon is not only domain specific but also dependent on the aspect in context. In this case, product aspects are known a priori. In (García-Moya et al., 2011) an approach to this problem is presented which consists in a methodology for obtaining a probabilistic ranking of product features that relies on an entailment model between opinion and feature words groups being both groups known in advance.

4.8 Conclusions

Nowadays, the Web has become the greatest source of information ever known. Companies and organizations can find now information about their business environments in the Internet. Specially, customer reviews about products and services available on-line in blogs and web forums constitute a highly valuable source of information for marketing, BI and product benchmarking. This opens a novel and interesting range of possibilities for the combination of data warehouse, OLAP and opinion retrieval technologies.

In this paper, we have presented a proposal to include sentiment analysis data into the corporate BI suite. As a result, we have presented a multidimensional data model that integrates sentiment data extracted from customer opinion forums into the corporate data warehouse. The process of extracting sentiment data considers the assessments made by customers on both products and product features, and produces a semantically rich data set that enables complex queries. As a consequence, another important result of this work consists of showing the usefulness of MDX queries over the integrated data warehouse to satisfy the requirements of Voice of the Market studies (Johne, 1994; Reidenbach, 2009).

This work has successfully applied semantic annotation to both corporate and sentiment data, so that the quality of the extracted data can be checked. Semantic annotation opens new issues such as the consistency of the extracted features with respect to the sentiment words and the known data of the product. For example, if opinion words were also categorized into semantic types, we could detect wrong associations between features and opinion words, like 'heavy' → 'quality'. Future work will be focused on exploiting the taxonomies supporting the semantic annotations for both the generation of good analysis dimensions, and for disambiguation and consistency issues.

Semantic annotations could also allow us to find relationship between products, mainly comparisons, which can be extracted by using open-IE techniques (e.g., (Etzioni et al., 2008)) over the reviews stream. For example, frequently, reviewers compare some feature for different product brands (e.g., “the lens of my Nikon are better than the new Canon's lens”). To recognize this kind of comparisons and to be able to properly distinguish which opinion belong to which product brand are two challenging issues to be addressed in the future.

Regarding the sentiment analysis part, there are some open issues in our approach. Firstly, we have assumed that opinion words always have the same base polarity, which can only be changed through valence shifters. However, it is well-known that some opinion words have different polarities for different products (e.g. “large” for “mobile phone” and “screen”). As future work, we are planning to regard context-dependent indicators similarly to (Lu et al., 2011). Alternatively, we could use the HITS graph as a regularization framework (Deng et al., 2009) to induce the polarities of context-dependent indicators. Another relevant issue is the enrichment of the opinion words lexicon. In (García-Moya et al., 2011) we show how new feature indicators can be learnt from a translation-based probability model similar to that presented in Section 4.5.2. However, some quality filter should be defined in order to avoid noisy data produced by these indicators. Finally, it would be also interesting to analyze the top untagged ranked features in order to enrich both the product lexicon and corporate product data.

Acknowledgments

This work has been partially funded by the “Ministerio de Economía y Competitividad" with contract number TIN2011-24147, and the Fundació Caixa Castelló project P1-1B2010-49.

Chapter 5

SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence

Rafael Berlanga, Lisette García-Moya, Victoria Nebot, María José Aramburu, Ismael Sanz and Dolores María Llidó. International Journal of Data Warehousing and Mining (IJDWM) 11, no. 4 (2015): 1-28.

5.1 Abstract

The tremendous popularity of web-based social media is attracting the attention of the industry to take profit from the massive availability of sentiment data, which is considered of a high value for Business Intelligence (BI). So far, BI has been mainly concerned with corporate data with little or null attention to the external world. However, for BI analysts, taking into account the Voice of the Customer (VoC) and the Voice of the Market (VoM) is crucial to put in context the results of their analyses. Recent advances in Sentiment Analysis have made possible to effectively extract and summarize sentiment data from these massive social media. As a consequence, VoC and VoM can be now listened from web-based social media (e.g., blogs, reviews forums, social networks, and so on). However, new challenges arise when attempting to integrate traditional corporate data and external sentiment data. This paper deals with these issues and proposes a novel semantic data infrastructure for BI aimed at providing new opportunities for integrating traditional and social BI. This infrastructure follows the principles of the Linked Open Data initiative.

5.2 Introduction

The massive adoption of web-based social media for the daily activity of e-commerce users, from customers to marketing departments, is attracting more and more the attention of Business Intelligence (BI) companies. So far BI has been confined to corporate data, with little attention to external data. Capturing external data for contextualizing data analysis operations is a time-consuming and complex task that, however, would bring large benefits to current BI environments (Pérez et al., 2008a). The main external contexts for e-commerce applications are the Voice of the Customer (VoC) and the Voice of the Market (VoM) forums. The former regards the customer opinions about the products and services offered by a company, and the latter comprises all the information related to the target market that can affect the company business. Listening to the VoM allows setting the strategic direction of a business based on in depth consumer insights, whereas listening to the VoC helps to identify better ways of targeting and retaining customers. As pointed out by Reidenbach (2009), both perspectives are important to build long-term competitive advantage.

The traditional scenario for performing BI tasks has dramatically changed with the consolidation of the Web 2.0, and the proliferation of opinion feeds, blogs, and social networks. Nowadays, we are able to listen to the VoM and VoC directly from these new social spaces thanks to the burst of automatic methods for performing sentiment analysis over them (Liu, 2012). These methods directly deal with the posted texts to identify global assessments (i.e., reputation) over target items, to detect the subject of the opinion (i.e., aspects) and its orientation (i.e., polarity). From now on, we will consider as social data the collective information produced by customers and consumers as they actively participate in online social activities, and we will refer to all the data elements extracted from social data by means of sentiment analysis tools as sentiment data.

A good number of commercial tools have recently appeared in the market for listening and analyzing social media and product review forums, for example Salesforce Radian6 (`http://www.salesforce.com/marketing-cloud`), Media Miser (`http://www.mediamiser.com`), and Sinthesio (`http://synthesio.com`), to mention just a few. Unfortunately, these commercial tools aim to provide customized reports for end-users, and sentiment data on which these reports rely on are not publicly available (indeed this is the key of their business). Consequently, critical aspects such as the quality and reliability of the delivered data cannot be contrasted nor validated by the analysts. This fact contrasts with the high quality that BI requires for corporate data in order to make reliable decisions.

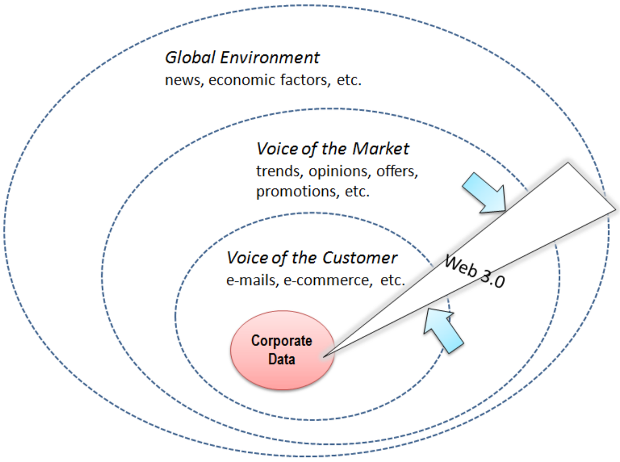

Global Environment
news, economic factors, etc.
Voice of the Market
trends, opinions, offers, promotions, etc.
Voice of the Customer
e-mails, e-commerce, etc.
Corporate Data
Web 3.0

Fig. 5.1 BI contexts and their relation to the Web 3.0 data infrastructure.

Apart from the sentiment analysis approaches, there is also a great interest on publishing strategic data for BI tasks within the Linked Open Data (LOD) cloud (Heath and Bizer, 2011). The Web 3.0 and LOD are about publishing data identified and linked to each other through a Unique Resource Identifier (URI), and providing data with well-defined semantics to allow users and machines to rightly interpret them. Projects like *Schema.org* are allowing the massive publication of product offers as micro-data, as well as specific vocabularies for e-commerce applications. Unfortunately, nowadays there is no open data infrastructure that allows users and applications to directly perform analysis tasks over huge amounts of published opinions in the Web.

In this paper we discuss the opportunities and advantages of defining new data infrastructures for performing social BI. As Figure 5.1 shows, in this social BI infrastructure, VoC and VoM sentiment data must be integrated together with all the external factors that may potentially affect a business (e.g., new legislations, financial news, etc.). We claim that such a data infrastructure must follow the principles of the LOD initiative. As a result, if web-based social data is migrated to the Web 3.0 as linked data in order to be shared, validated and eventually integrated with corporate data, a new global BI scenario for e-commerce applications is enabled. Furthermore, most data and vocabularies used by researchers and companies for performing sentiment analysis could be better exploited if they are shared, contrasted and validated by the community.

The main contributions of this paper are summarised as follows:

- We propose BI analytical patterns to combine corporate with social data.
- We propose a novel semantic data infrastructure to publish both social data and automatically extracted sentiment data. This data infrastructure follows the LOD principles, and therefore it is aimed at linking the social data with other related datasets in the LOD cloud.
- We propose a novel method for data provisioning, called ETLink, which covers the requirements identified in this scenario.

The rest of the paper is organized as follows. Next section is dedicated to describe the background of the proposal. In Sections 5.4 and 5.5 respectively, the proposed social BI infrastructure is presented and its main component datasets described. Afterwards, Section 5.6 discusses how the main components of the SLOD-BI data infrastructure are populated from the social resources. Section 5.7 presents the evaluation. An illustrative application of this infrastructure and some example analysis operations are depicted in Section 5.8 and the overall conclusions are summarized in Section 5.9.

5.3 Background

BI refers to the methodologies, architectures and technologies that transform raw data into meaningful and useful information to enable more effective decision-making. BI technologies provide historical, current and predictive views of business operations. Common functions of BI are reporting, online analytical processing (OLAP), data mining, complex event processing and text mining among others. Often BI applications use data gathered from a data warehouse (DW) or a data mart. In fact, one of the most successful approaches to BI has been the combination of DW and OLAP (Codd et al., 1993).

Traditional BI follows a three-layered architecture consisting of the data sources layer, where all the potential data of any nature is gathered, the integration layer, which transforms and cleanses the data from the sources and stores them in a DW, and the analysis layer, where different tools exploit the integrated data to extract useful knowledge that is presented to the analyst as charts, reports, cubes, etc. For the integration layer, the multidimensional model (MD) is used, where factual data gathered from the data sources layer must be expressed in terms of numerical measures and categorical dimensions. The semantics of this model consists in representing any interesting observation of the domain (measures) at its context (dimensions). The typical processes in charge of translating data from the data sources layer to the integration layer are called ETL processes (extract, transform, and load).

Even though this traditional architecture has proved useful to analyze corporate data, it presents several limitations that make it unsuitable to meet the analytical requirements of social BI. First of all, the previous architecture only works well in a closed-world scenario, where both the data sources and the user requirements are static and known in advance. Moreover, the ETL processes are meant to periodically load well-structured data in batch mode, as they usually apply heavy cleansing transformations. The massive availability of web-based social media related to business processes has become a valuable asset for the BI community. The integration of these external and heterogeneous data sources with corporate data would enable more insightful analysis and would bring new marketing opportunities so far unexplored. The need of incorporating external data to the traditional analysis processes is not new. The majority of approaches try to incorporate external data to the already existing MD structures by establishing mappings. Thus, the integration is only circumstantial and problems such as the lack of dynamicity and freshness remain. These problems require a shift in the traditional BI architecture towards a more dynamic, open and flexible infrastructure.

In recent years, opinion mining and sentiment analysis have been an important research area that combines techniques from Machine Learning (ML) and Natural Language Processing (NLP). One of the most relevant applications of sentiment analysis is the aspect-based summarization (Liu, 2012). Given a stream of opinion posts, aspect-based summarization is aimed at giving the most relevant opined aspects, also called features or facets, along their sentiment orientation, usually given by a score and a polarity. For example, given a stream of opinions about digital cameras, some relevant aspects can be the battery life, the quality of the lenses, etc. Sentences like "the battery life is too short" will contribute to the negative orientation of the battery life aspect, whereas others like "we took very good pictures" will contribute to the positive orientation of the picture quality aspect. Aspect-based summarization has been usually divided into three main tasks, namely: sentiment classification, subjectivity classification and aspect identification. The first one is focused on detecting the sentiment orientation of a sentence, the second one consists of detecting if a sentence is subjective (i.e., if it contains a sentiment), and the latter one consists of detecting the most relevant aspects of an opinion stream. ML supervised approaches have been widely adopted to solve these problems, as they can be easily modelled as traditional classification problems. Unfortunately, it is unfeasible to get training examples for all the items and potential aspects regarded in opinion streams. Thus, supervised approaches have been restricted to obtain sentiment lexicons and to detect sentence subjectivity with them (Liu, 2012). As a consequence, sentiment analysis in open scenarios should rely on unsupervised or semi-supervised methods (Garcia-Moya et al., 2013). Moreover, sentiment analysis must be blended with social network analysis, which basically aims to predict the diffusion and popularity of opinions spread across social networks (Guille et al., 2013).

The problem of how to exploit social data to extract sentiment data that could be useful for BI applications and how to integrate the extracted opinion data into the existing corporate DW is still an open issue. Pérez et al. (2008a) proposed a first approach to this problem by contextualizing a sales DW with on-line customer reviews about the company products/services. A contextualized warehouse allows users to obtain strategic information by combining all their sources of structured data and documents (Pérez et al., 2008b). The analysis cubes of a contextualized warehouse, denoted R-cubes, are special since each fact is linked to an ordered list of documents. These documents provide information related to the fact (i.e., they describe the context of the fact). Similarly, the EROCS (Bhide et al., 2008) system basically constructs a link table between DW facts and external documents. Named entity recognition (NER) techniques are used to identify fact dimension values within document texts, and then valid combinations likely to represent facts are extracted to define the fact-document links. However, all these approaches only regard the explicit rates of the opinion posts (e.g., 5-star ratings) as sentiment measures, which are clearly not enough to perform social BI analysis. Firstly, many opinion sources do not provide explicit ratings. Furthermore, most interesting BI analysis involves facet-sentiment pairs, which must be extracted from post texts.

A recent approach to integrate BI with sentiment data was proposed by García-Moya et al. (2013) where a corporate DW is enriched with sentiment data from opinion posts. In this approach, sentiment data are extracted from opinion posts and then stored into the corporate DW. As a result, sentiment and corporate data can be jointly analyzed by means of OLAP tools. The main limitation of this approach is that sentiment data must be fitted to a predefined MD schema, which reduces the range of analytical operations that can be performed over the extracted sentiment data. In contrast to the closed and rigid scenario of DW/OLAP, in this paper we propose an open and dynamic framework based on LOD, where data can be linked to external sources on demand, without being attached to rigid data structures or schemas.

As BI mainly involves the integration of disparate and heterogeneous information sources, semantic issues are highly required for effectively discovering and merging data. Most work proposed in this direction can be classified either in those focusing on Web data (Pérez et al., 2008), where the presence of Semantic Web (SW) technologies is granted, and those using SW technologies to tackle integration in any scenario. A pioneer approach was presented in (Mena et al., 2000), where multiple data sources are expressed and integrated via description logics. The main idea behind this model is to achieve a loose coupling between the integrated data sources through semi-automatic ontology mapping tools. It is worth mentioning that this is the main leitmotiv behind the LOD initiative (Bizer et al., 2009).

The LOD initiative aims at creating a global web-scale infrastructure for data. Relying on the existing web protocols, this initiative proposes to publish data under the same principles that web documents, that is, they must be identified through a Unique Resource Identifier (URI), with which any user or machine can access to their contents. Similarly to web documents, these data can also be linked to each other through their URIs. In order to manage the resulting data network, data must be provided with well-defined semantics to allow users and machines to rightly interpret them. For this purpose, the W3C consortium has proposed several standards to publish and semantically describe data, mainly the Resource Description Framework (RDF) and the Ontology Web Language (OWL). In this paper we refer as semantic data infrastructures to the data networks resulted from publishing and linking data with the standard formats RDF and OWL.

Semantic data infrastructures provide a series of standards and tools for editing, publishing and querying their data. The basic component of this infrastructure is the *dataset*, which consists of a set of RDF triples that can be linked to other LOD datasets. These datasets usually provide a SPARQL endpoint, with which data can be accessed via declarative queries. Additionally, SPARQL also enables distributed queries over linked datasets. These data infrastructures are opening new opportunities to both data providers and consumers to develop new applications, which goes beyond the corporate boundaries. More specifically, LOD has opened new ways to perform e-commerce activities such as retailing, promotion, and so on. Proposals like *schema.org* and *GoodRelations* (`http://www.heppnetz.de/projects/goodrelations`) are allowing the massive publication of product offers as micro-data, as well as specific vocabularies for e-commerce applications. Additionally, commercial search engines like Google and Yandex are adopting these formats to improve the search of these data. As far as we know, there is no open data infrastructure that allows users and applications to directly perform analysis tasks over huge amounts of published opinions in the Web. Some preliminary work such as MARL (Westerski et al., 2011) attempts to provide proper schemas for expressing opinion data as linked data. However, MARL has not been devised for performing large-scale BI analyses, and consequently it disregards the BI patterns with which data should be aggregated, as well as data provisioning methods to populate the intended data infrastructure.

Although these three worlds, BI, sentiment data and LOD technology, have kept unconnected to each other until recently, in this paper we advocate for a BI paradigm shift towards a LOD infrastructure of sentiment data extracted from social media. With this infrastructure companies are able to execute complex analysis operations that dynamically integrate corporate data with relevant social data. With this infrastructure, it is possible to study the response of consumers to the company strategic decisions, to identify the sentiments that company products and services produce among consumers or to analyse social data with the purpose of predicting new demands of the market.

5.4 A Social BI Data Infrastructure

From a BI point of view, social data can be regarded as a multidimensional model that can be blended with company data for helping decision-making. For example, the reputation of a product, the most outstanding features of some brand, or the opined aspects of an item can be represented as multidimensional data, and efficiently computed through OLAP tools (García-Moya et al., 2013). In this section we first present a new set of analytical patterns that combine corporate and social data. Then, in Section 5.4.2, the global requirements of our social BI infrastructure are established. Finally, a structural view and a functional architecture for implementing the infrastructure are introduced in Section 5.4.3.

5.4.1 Analytical patterns for social BI

The main BI patterns to analyze and combine corporate and social data are summarized in Figure 5.2. The analysis patterns at the corporate data side of the figure correspond to the traditional MD model of a typical DW (Codd et al., 1993). Patterns at the social data side constitute the main contribution of our proposal, and they are explained in the next paragraphs.

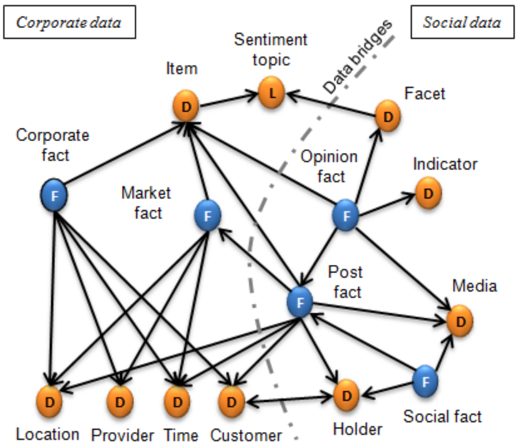


Fig. 5.2 Main BI patterns in a social analysis context scenario. Notice that some facts also act as dimensions of other facts (e.g., *Post fact* and *Market fact*).

In the figure, facts (labelled with 'F') represent spatio-temporal observations of some measure (e.g., units sold, units offered, number of positive reviews, and so on), whereas dimensions (labelled with 'D') represent the contexts of such observations. In some cases, facts can have a dual nature, behaving as either facts or dimensions according to the analyses at hand. For example, in Figure 5.2, a post can be considered either as a fact or as a dimension of analysis for opinion and social facts. Dimensions can further provide different detail levels (labelled with 'L'). For example, the dimension *Item* is provided with the level Sentiment topic. In Figure 5.2 we have distinguished two kinds of corporate data that can be combined with social data, namely: *Corporate fact*, which concerns business transactions (e.g., sales, contracts, etc.), and *Market fact*, which the promotions and offers of the company products and services.

Measure	Example values	Fact type
Polarity	(-1,0,+1)	Opinion
Rating	(★,★★,★★★,...)	Post
Like	(👍, 👎)	Post
Popularity	(-10,...,+10)	Social
Credebility	(0,...10)	Social

Table 5.1 Examples of measures for social BI facts.

The main facts concerning social data are *opinion facts*, *post facts*, and *social facts*. *Opinion facts* are observations about sentiments expressed by opinion holders concerning concrete facets about an item, along with their sentiment indicators. For example, the sentence "I don't like the camera zoom" expresses an opinion fact where the facet is "zoom", and the sentiment indicator is "don't like" (negative polarity). *Post facts* are observations of published information about some target item, which can include a series of opinion facts. Examples of post facts can be reviews, tweets, and comments published in a social network. Notice that opinion facts are usually expressed as free texts in the posts, and therefore it is necessary to process these texts to extract the facts (Liu, 2012). Finally, *social facts* are observations about the opinion holders that interchange sentiments about some topic. These facts are usually extracted from social networks by analyzing the structure emerged when the opinion holders discuss about some topic (Pak and Paroubek, 2010). Notice that topic-based communities can be very dynamic as they rise and fall according to time-dependant topics (e.g., news, events, and so on).

As for the measures associated to these facts, Table 5.1 shows some examples of typical measures used in the literature for sentiment and social analysis.

It is important to notice that in Figure 5.2 the corporate and social BI patterns are separated by a dotted line. As the intended data infrastructure is aimed at facilitating the integration of information, we define *data bridges* between corporate and social data elements (see the arrows that cross the dotted line in Figure 5.2). *Data bridges* are the patterns that can be used to execute analysis operations that combine corporate and social data. Each data bridge consists of an internal data element and an external data element (i.e., dimensions or facts) that are related in the analysis scenario. For example, the analytical pattern between market facts and post facts can be applied to study the features of marketing campaigns from the point of view of its acceptance by consumers, and in the other way, to analyse consumer opinions in the context of each campaign. Different applications and different scenarios can make use of different data bridges to integrate data.

Data bridges support the communication channels between the internal and external data sources and it is very important for companies to enable all the means necessary to implement them. Of special interest are the data bridges that relate *Sentiment Topic* to *Facet*, and *Customer* to *Holder* dimensions. In the first case, the company can specify the most important topics in its items (products or services) that require some sentiment analysis and that use to coincide with some of the facets that appear in the opinions of a post. In order to facilitate the implementation of this data bridge, companies and social media users could apply the same *hashtags* to mark up these topics. In the second case, it is important to note that when the holder of an opinion is a known customer, both entities must be identified as the same. With respect to these data bridges companies must ensure that the corporate data and metadata files include key information to enable the recognition of corporate entities in social data by means of sentiment analysis tools.

Some examples of interesting sentiment and social analysis operations that can be done with the previous BI patters are the following ones:

- To identify troublesome parts of some product or service (i.e., item).
- To measure the popularity of product promotions from rival companies.
- To find the best nodes in social networks to advertise a product.
- To predict the popularity of a topic (i.e., sentiment topic) in the different communities.
- To analyze the evolution of an item sentiment within a topic-based community.

5.4.2 Requirements of a social BI data infrastructure

Regarding the nature of the data to be published in the data infrastructure, we have identified a set of global requirements that are not covered yet by current proposals, namely:

1. The infrastructure must give support for *massive generation of sentiment data from posts* (e.g., reviews, tweets, etc.) so that high volumes of crawled data can be quickly processed and expressed as linked data. As in DWs, a series of ETL processes are needed to periodically feed the data infrastructure. These ETL processes are quite unconventional, as they deal with semi-structured web data, perform some kind of sentiment analysis, and output RDF triples.

2. Sentiment data published in the infrastructure must be semantically represented under *well-controlled vocabularies and useful taxonomical relationships*. Currently, sentiment data is automatically extracted from texts with either statistical (Garcia-Moya et al., 2013) or NLP methods (Liu, 2012), but they do not bring well-defined semantics for enabling BI analyses. For example, most automatic methods capture facets and sentiment indicators from text reviews but these data are not organized into semantic groups (e.g., optics, storage and image quality for cameras) to properly calculate the partial scores for each semantic group. In this context, we have to say that the success of traditional BI partially stems from the capacity of OLAP tools for exploring data through hierarchical dimensions. Another relevant aspect to take into account is the context-dependent nature of these data (Lu et al., 2011), which may also require inference capabilities.

3. Analysing social data can imply the massive generation of opinion facts from social media sources (i.e., Big Data). Consequently, the infrastructure must support *massive processing and distribution of data*, providing optimal partitions with respect to data usage. Since BI analysis is subject-oriented, data distribution should take profit from the topics around which opinions are generated. For example, opinion facts should be organized into item families (e.g., electronic products, tourist services, etc.) and allocated into separate distributed datasets.

4. The infrastructure must provide *fresh data* by migrating as quickly as possible published posts. In this respect, depending on the scenario and other features, social data elements have a different lifespan during which they can be considered fresh for real time applications.

5. The infrastructure must ensure the *quality and homogeneity of the datasets*, dealing with the potential multi-lingual issues of a BI scenario. At this respect, it is essential to focus only in the posts with opinions that are relevant, discarding all those social data elements without a clear and valid meaning. As e-commerce acts in a global market, sentiment data extracted from different countries will be expressed in different languages. Datasets must support multi-lingual expressions as well as organize them around well-understood semantic concepts (see requirement 2). Additionally, links between datasets of the intended infrastructure must be as coherent as possible, using the appropriate classes and data types offered by our infrastructure. Some current approaches like MARL (Westerski et al., 2011) allow users to express opinion facts with any kind of resource (e.g., a string, a URI to an external entity, etc.) Despite the fact that this makes the schema much more flexible to accommodate any opinion fact, it makes unfeasible to perform a BI analysis over these data.

6. The infrastructure must support *complex analysis operations* that integrate data with two different purposes. In many cases, companies will exploit the social datasets to execute analysis operations on internal corporate data but contextualized with external sentiment data. For this purpose, social data should be easily structured, loaded and integrated with corporate data in order to analyse it with the available BI applications. In the other cases, advanced applications working on the cloud will analyse relevant social data in the context of some company events such as marketing campaigns or special offers. Although these applications will mainly use social data, they will also need relevant data coming from the corporate databases.

In this paper, we mainly focus on the points 1, 2, 5 and 6. Points 3 and 4 will be left to the future work since they depend on the growth rate of the infrastructure: number of followed opinion streams, variety of domains to be regarded, and so on.

5.4.3 SLOD-BI overview

Regarding the previous requirements, Figure 5.3 proposes the architecture for the intended social BI data infrastructure. First, we divide the involved datasets into two layers. Thus, the inner ring of Figure 5.3 regards the main vocabularies and datasets of the proposed infrastructure, whereas the outer ring comprises the external linked open vocabularies (LOV), and the datasets that are directly related to the infrastructure (e.g., DBpedia and productDB). Every SLOD-BI component consists of a series of RDF-triple datasets regarding some of the perspectives we consider relevant for BI over sentiment data. For example, in the Item Component each dataset holds the products associated to a particular domain (e.g., cars, domestic devices, etc.) These datasets are elaborated and updated independently of each other, and can be allocated in different servers. All the datasets of a component share exactly the same schema (i.e., set of properties), reflecting the BI patterns defined in Section 5.4.1.

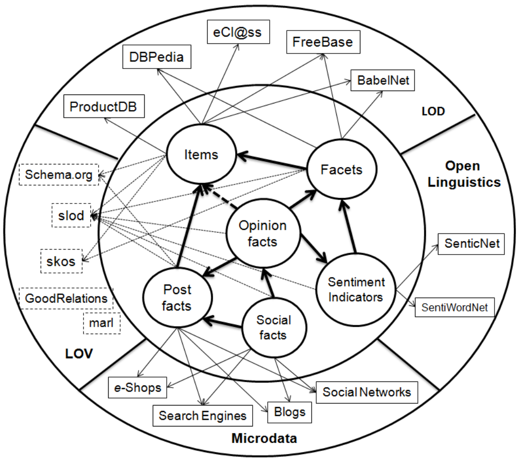


Fig. 5.3 Structural view of SLOD-BI. LOD stands for Linked Open Data (`http://linkeddata.org`), and LOV for Linked Open Vocabularies (`http://lov.okfn.org`).

In Figure 5.3, links between components are considered hard links, in the sense that they must be semantically coherent, and they are frequently used when performing analysis tasks. Consequently, the infrastructure should facilitate join operations between triples of these datasets. On the other hand, links between infrastructure components and external datasets are considered soft links, as they just establish possible connections between entities of the infrastructure and external datasets. These external datasets are useful when performing exploratory analyses, that is, when new dimensions of analysis could be identified in these external datasets. Links to external datasets like DBpedia play a very relevant role in this infrastructure since they can facilitate the migration of existing review and opinion data. For example, reviews already containing micro-data referring to some product in DBpedia will be automatically assigned to the item URI of the corresponding SLOD-BI dataset.

Figure 5.4 summarizes the functional view for the proposed data infrastructure. At the bottom layer, the external Web data sources are selected and continuously monitored to extract, transform and link (ETLink) their contents according to the SLOD-BI infrastructure. As earlier stated, social BI facts are regarded

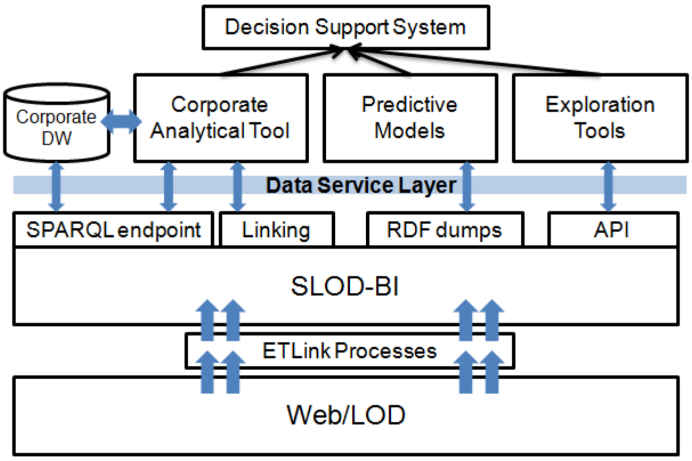


Fig. 5.4 Proposed functional view for SLOD-BI infrastructure.

as spatio-temporal observations of user sentiments in social media. Therefore, both spatial and temporal attributes must be captured and explicitly reflected in the ETLink processes.

The SLOD-BI infrastructure is exploited by means of the *data service layer*, which is in charge of hosting all the services consuming sentiment data to produce the required data for the analytical tools. These services are implemented on top of a series of basic services provided by the infrastructure, namely: a SPARQL endpoint to directly perform queries over sentiment data, a Linking service to map corporate data to the infrastructure data (e.g., product names, locations, etc.), an RDF dumper to provide parts of the SLOD-BI to batch-processing services, an API for performing specific operations over the infrastructure (e.g., registering, implementing access restrictions over parts of the infrastructure, etc.), and visual tools for data exploration.

Notice that in the proposed functional view, sentiment data is integrated with corporate data at the corporate analytical tool, by making use of some intermediate data service. In this case, corporate and sentiment data is aggregated separately and joined inside the analytical tools through a cross-join. This process is similar to Pentaho blending processes to integrate external and internal data (`http://www.pentaho.com/big-data-blend-of-the-week`). The predictive models and exploration tools will allow the execution of complex processes over the sentiment data in the infrastructure. In both cases, the data service layer will facilitate the retrieval of the relevant corporate data as necessary. An important advantage of using the data service layer to query the corporate DW is that it helps to maintain the appropriate level of data governance and security necessary for accurate and reliable analysis (Carey et al., 2012).

5.5 SLOD-BI Datasets

In this section, we describe the main datasets that will constitute the SLOD-BI data infrastructure as represented in the inner ring of Figure 5.3. In addition to complying with the W3C recommendations about publishing linked data, these datasets have been defined according to the following criteria:

Property	Range	Description
s:brand	rdfs:Resource	Link to the manufacturer or brand of the item.
slod:onDomain	skos:Concept	Item family (taxonomy).
skos:related	slod:Item	Related items.

Table 5.2 Basic properties for describing items.

- Take profit from existing vocabularies and schemas as much as possible, mainly from *schema.org*, which is the de facto standard in e-commerce.
- Distribute data according to the identified BI demands (e.g., subject and topics), in order to achieve high scalability.
- Keep the inner datasets as coherent as possible, so that they can be easily queried for analytical tasks.
- Provide soft schemas in order to accommodate incomplete data.
- Provide data provenance metadata: all sentiment data captured from the web should be attached to their location (URL) and time, and all calculated measures should be attached to the service (URL) used for such calculations.

The rest of the section shows the most relevant aspects of the datasets included in each component. In the specific schemas, we do not include the standard properties that are common to all datasets, namely: rdfs:label to specify possible variants and synonyms of the described entity, owl:sameAs to specify mappings between infrastructure elements and external datasets, and rdf:type to classify instances into classes. Moreover, to represent and organize topics within the infrastructure we use the Simple Knowledge Organization System (`http://www.w3.org/TR/skos-reference/`), and for data provenance the Dublin Core vocabulary (dc). We also adopt whenever is possible the vocabulary of *schema.org* (name space s), since it is the standard de facto for e-commerce microdata.

5.5.1 Items component

This component contains the datasets describing concrete products and services as well as their manufacturers (e.g., product brand, providers, facilitators, etc.) These datasets must be kept as simple as possible, just providing useful attributes for BI tasks. Furthermore, additional attributes and relationships can be accessed through the links to externals datasets such as eCl@ss, DBpedia, ProductDB, FreeBase, etc. The main class of this component is slod:Item, whose basic properties are summarized in Table 5.2.

5.5.2 Facets component

This component comprises all the elements subject to evaluation in the user's opinions, which are called *facets* (slod:Facet). According to the analytical pat-

Property	Range	Description
slod:onDomain	skos:Concept	Item family where the facet is defined.
slod:onTopic	skos:Concept	Sentiment topic to which the facet belongs.
skos:related	slod:Facet	Related facets.

Table 5.3 Basic properties for describing facets.

terns of Section 5.4.1, we consider two detail levels about judged elements, namely: *sentiment topic* and *facet*. A *sentiment topic* describes some BI perspective of an item family, like “design”, “safety” and “comfortability” for cars. Sentiment topics group facets, which can be any abstract or concrete aspect opined by the users (e.g., “engine”, “diesel engine”, etc.). In order to account for the semantic relationships between facets (e.g. “diesel engine” is-an “engine”), we make use of the SKOS vocabulary. However, as this kind of relationships is not required for BI analysis they can be omitted.

Currently there are scarce LOD datasets including facets subject to opinions (e.g., some small *GoodRelations* ontologies). We may also consider technical specifications about products like in eCl@ss, but they do not properly cover the features customers usually opine on (Garcia-Moya et al., 2013). As a consequence, facets should be extracted directly from text reviews by applying sentiment analysis methods (Liu, 2012). Indeed, one of the SLOD-BI goals is to conceptualize and make public facets automatically extracted from reviews (see Section 5.6.2). For this purpose, we propose a simple schema (Table 5.3) to which item facets must map to. The main issues for performing these mappings are: to group together expressions denoting a same facet (they should appear as different labels of the same instance) and to classify facets into sentiment topics. For the former issue, we make use of external datasets such as BabelNet by using an automatic linking process (see Section 5.6.3), whereas the latter issue is addressed by manually defining the require mappings according to corporate criteria.

5.5.3 Sentiment indicators component

Sentiment analysis relies on the existence of a set of words and expressions that indicate some opinions about a subject. The Sentiment Indicators Component is mainly based on linguistic resources that allow identifying facets from review texts as well as sentiments associated to them.

Sentiment words, also known as opinion words, are the most important indicators of sentiments about a subject. These are words commonly used to express positive or negative opinions. For example “excellent”, “amazing”, “good” are positive words whereas “bad”, “terrible”, “awful” are negative ones. Additionally, there also exist sentences used for expressing opinions, for example, “cost a pretty penny” and “cost an arm and a leg” all are referring to the indicator *expensive*.

Sentiment indicators could be defined as context-independent or context-dependent (Lu et al., 2011). An opinion indicator is context-dependent when its polarity depends on the domain and/or the facets it is modifying (e.g., “un-

Property	Range	Description
slod:onFacet	slod:Facet	(Optional) Associated facet (implicit/context).
slod:hasPolarity	xsd:integer	Default polarity associated to the indicator.
dc:subject	skos:Concept	Item family to which the indicator can be applied.

Table 5.4 Properties for describing sentiment indicators.

Property	Range	Description
dc:publisher	rdfs:Resource	Link to the media in which the post has been published.
s:rating	s:Rating	Overall assessment.
s:itemreviewed	s:Item	Link to the reviewed item.
s:reviewer	rdfs:Resource	Author of the review (Holder).
s:dtreviewed	s:Date	Publication date of the review.

Table 5.5 Properties for describing post facts.

expected" for movies (+) and electronic devices (-)). Even within the same domain, the polarity of an indicator may be different depending on the facet it applies to. For example, the word "long" in digital cameras: "long delay between shots" (-) and "long battery life" (+). Another interesting kind of opinion indicators consists of expressions that implicitly bring the facet. For example, the indicator "too expensive" refers to the aspect "price". The main class of the Sentiment Indicator Component is slod:Indicator. Table 5.4 shows its main properties.

Nowadays there exist many sentiment lexicons, some of them available in LOD. The most popular ones are SentiWordNet (Esuli and Sebastiani, 2006) and SenticNet (Cambria et al., 2013), which provide sentiment-based characterizations for common words in English. Unfortunately, these lexicons are of limited use because they are of general purpose and do not take into account context-based indicators (Lu et al., 2011). Additionally, there is a proliferation of web services for computing polarities from free-texts (Thelwall et al., 2010). This kind of services could be applied to obtain the values of the property slod:hasPolarity. In order to account for both context-based indicators and sentiment indicators implying a facet, we include the property slod:onFacet. For example, the following sentiment indicators also imply a facet: *expensive* → *cost*, *delicious* → *taste*, *spacious* → *comfort*.

5.5.4 Post and opinion facts components

Currently, we can find several proposals for representing metadata of reviews and social data in LOD. One of the main references is *schema.org*, which has been adopted by Google for rich snippets over posts. This vocabulary covers all aspects we need for the Post Component, and therefore we have adopted it without any extensions. Table 5.5 shows the main properties associated to the post fact class.

Opinion facts express the associations between features/aspects to opinion

Property	Range	Description
dc:publisher	rdfs:Resource	Link to the media in which the post has been published.
slod:onFacet	slod:Facet	Link to each facet concept involved in the fact.
slod:fromPost	slod:PostFact	Link to the post from which the fact was extracted.
slod:onTargetItem	slod:Item	In comparisons, link to the compared item.
slod:opinionExpression	xsd:string	Opinion linguistic expression.
slod:facetExpression	xsd:string	Facet linguistic expression.
slod:withIndicator	slod:Indicator	Link to each opinion concept involved in the fact.
slod:hasPolarity	xsd:integer	Estimated polarity of the fact.
slod:validTime	s:Date	Time point at which the fact occurs.

Table 5.6 Properties for describing opinion facts.

indicators that appear at the post texts. In our approach, an opinion fact is always linked to the post object from which it was identified. Consequently, each opinion fact takes the time and place dimensions from its linked post. Thus, the schema of an opinion fact can be just expressed with the feature/aspect and indicator/shifters involved in the fact. Table 5.6 summarizes the properties associated to the opinion fact class.

Another kind of opinion fact regarded in (Liu, 2012) is that of product comparisons. To represent comparisons, the property slod:onTargetItem is added. Notice that we can combine these properties to express for example a comparison between two products w.r.t. some aspect (e.g., “it has better zoom than camera Y”). In this case, the indicator "better" represents the comparison operator, and the target item is “camera Y”.

5.5.5 Social facts component

There are a few useful sources for extracting social data. The main one is that provided by the social networks’ own APIs (e.g., the Twitter API, The Google+ API and Facebook’s Graph API). Opinions formulated in the context of these social networks usually have associated a large amount of meta-data, which is accessed through these APIs. Opinion meta-data can be used to find indicators about the *impact* of the opinion in the context of the social network (Guille et al., 2013). We refer to these indicators as *social facts*. Thus, the aim of social facts is to provide relevance indicators about holders and their opinions in the context of the community they belong to. Measures such as the number of followers and the number of times an opinion was shared, are indirect indicators of both opinions and holders relevance as perceived by the social community. As a consequence, these metrics resemble those used to assess the reach of social media campaigns.

Unfortunately, despite the great demand of social data for analysis, currently there is no standard vocabulary to express this kind of data as open linked data, although some preliminary drafts are being discussed within the Open Social Foundation (`http://opensocial.org`). As a consequence, we have devised a set of metrics which are (i) useful for analytic purposes (specifically, time-oriented); (ii) practical, in the sense that they can be obtained from data actually available

Property	Range	Description
slod:fromPost	slod:postFact	Link to the opinion post fact that defines the observation point for measures (reference).
dc:author	rdfs:Resource	Link to the post author (i.e., reviewer or holder).
dc:publisher	rdfs:Resource	Link to the social media where the observation is performed (e.g. Twitter, Facebook, etc.).
slod:communityImpactId	xsd:string	Unique identifier for the community impact information.
slod:reviewer_indegree	xsd:integer	Average number of followers (or friends, etc.) of the reviewer during the considered time.
slod:reviewer_mentions	xsd:integer	Count of the number of mentions of the reviewer in the social network during the considered time.
slod:in_response_to	slod:postFact	If the reference opinion (slod:fromPost) is a response to another opinion post, the reference of the latter.
slod:repostings	xsd:integer	Number of times the opinion has been reposted in the social network (e.g., by retweeting or sharing in Facebook) during the considered time.
slod:positive_feedback	xsd:integer	Number of times the opinion has been shared in the social network (e.g., by marking as favourite in Twitter or indicating +1 in Google+) during the considered time.
slod:shared_with	xsd:integer	Number of community members to which the opinion fact was posted during the considered time.
slod:responses	xsd:integer	Number of the direct responses to this opinion (e.g., the number of Twitter replies or Facebook comments) during the considered time.
slod:validTime	s:Date	Time point at which the fact occurs.

Table 5.7 Properties for describing social facts.

by exploiting the APIs of relevant social networks, such as Twitter, Facebook or Google+; and (iii) general, in the sense that they can be applied to different social networks. This is a strong constraint, since the data models provided by social networks can have significant differences. In most cases, simple mappings between the concepts in the different social networks can be found; for example, we will use the property *positive_feedback* to record the fact that an opinion has been favourite (in Twitter), liked (in Facebook) or +1'd (in Google+) a given number of times in a time interval. However, in a few instances, there is no direct equivalence; the most prominent case is *user mentions* (which gives an indicator of the influence of the reviewer), which can be directly measured in Twitter and Google+, but only indirectly, and incompletely, assessed in Facebook. Table 5.7 shows some properties that fulfil these criteria.

5.6 Data Provisioning

This section discusses how the main components of the SLOD-BI data infrastructure are populated from the selected Web resources (e.g., blogs, twitter streams, product reviews sites, and so on). The whole process of data provisioning is summarized as follows:

1. For each followed opinion stream (which is associated to a particular item), all metadata and micro-data are extracted and processed to generate the corresponding social and post facts.
2. From each post, textual contents are pre-processed for normalization issues (Section 5.6.1). Then, the system automatically extracts facet and opinion expressions by applying the vocabularies learnt from domain and background corpora (Section 5.6.2).
3. Automatically extracted sentiment data is then linked to the infrastructure, generating thus the opinion facts associated to posts (Section 5.6.3).

The different steps of the whole process are performed within a framework called ETLink which provide the necessary operators to generate all required data. A brief description of this framework is provided at section 5.6.4.

5.6.1 Text pre-processing

It is well known that products review web sites, forums, social networks, and so on, are written in casual language without paying attention to the spelling. As our method is fully unsupervised and results are statistical by nature (meaningless words are oriented to reach low-probability values), the presence of many spelling errors affects significantly the results. Repeatedly misspelled adverbs, prepositions, conjunctions and so on may be considered as "new" words by the method and, therefore, erroneously classified as facet or sentiment indicators. In order to alleviate this issue, the target collection should be fixed before applying the learning method. Basically, the text pre-processing phase is divided into the following steps:

1. Fix negative contractions (English): When one or more letters are missed out in a contraction, an apostrophe is inserted (e.g., "isnt" is replaced by "isn't").

2. Remove unnecessary repeated letters. Duplicating letters in texts is a growing phenomenon in social networks. People may duplicate letters in an effort to emphasize their sentiments. However, some cases can be ambiguous, like in "ohhhh, gooooooood", which can refer either to "good" or "god". In the current version, each repeated letter is left with a maximum of two repetitions (e.g., "ohhhhh" is replaced by "ohh" and "goooood" by "good").

3. Fix potential spelling errors. For this purpose, we apply the Suggester Spell Check (`http://www.softcorporation.com/products/spellcheck`), which can be used with an already pre-compiled dictionary. Due to the presence of acronyms and specific words used in particular domains, we restrict the corrections to words not found in Wordnet, with more than four letters, and a resulting score lower than a given threshold (in the experiments is set to 100).

5.6.2 Vocabularies construction

Regarding the structure of the proposed infrastructure (Figure 5.3), the first step towards its population consists in identifying the basic vocabularies that allow describing sentiments over products and services. Basically, we need to distinguish between two almost disjoint vocabularies, namely: facets and sentiment indicators. The construction of these vocabularies will be performed for each particular domain (e.g., cars, cameras, etc.) since each domain exhibits different terminologies and writing styles. It is worth mentioning that domain ontologies (if existing) are usually targeted to other purposes different from sentiment analysis, such as e-commerce (e.g., technical product aspects), and therefore they are usually incomplete for describing social sentiments. This is why some machine learning method is necessary to comprehensively capture potential facets and sentiment indicators from target collections. Once these potential concepts are identified, they can be linked to existing ontologies to get a richer view of them. For this purpose, we adopt the unsupervised statistical method proposed in (Garcia-Moya et al., 2013), which aims at assigning probabilities to words acting as either facets or indicators. This method is summarized as follows.

We consider stochastic mappings between words to estimate a unigram language model of facets from a probabilistic model of opinion words. The initial unigram language model for facets P is defined as follows:

$$P(w_i) = (T^k \cdot Q)_i \tag{5.1}$$

The matrix $T = \{p(w_i|w_j)\}_{1\leq i,j\leq n}$ represents the word-word entailment probabilities, which are estimated from local contexts of a large collection of opinion posts of the target domain. The unigram model Q is a generative model of opinion words, which assigns to each word w the likelihood of being an opinion word, denoted $Q(w)$.

In addition, we consider refining the unigram model P to avoid the assignment of high-probability values to meaningless words such as prepositions and conjunctions. The refined unigram language model P' is obtained by means of an expectation-maximization (EM) (Neal and Hinton, 1998) process which minimizes the cross entropy with respect to a background model P_{bg}:

$$- \sum_{1 \dots n} P(w_i) \cdot log(\lambda \cdot P'(w_i) + (1 - \lambda) \cdot P_{bg}(w_i)) \tag{5.2}$$

In (Garcia-Moya et al., 2013), these statistical models are used to generate the ranking of facet-sentiment pairs of either a review or the whole collection. In this work, our aim is slightly different, as we aim to build two basic vocabularies for the data infrastructure, namely: words acting as facets (model H), and words acting as sentiment indicators (model O). For this purpose, we apply the following iterative process:

Repeat until $H^{(j)}$ and $O^{(j)}$ do not change with respect to step $j - 1$:

$$\begin{aligned}
j \quad &+= \quad 1 && (5.3)\\
H^{(j)} \quad &= \quad \arg\max_{H} P' \log(\alpha \cdot H + (1-\alpha) \cdot O^{(j-1)}) && (5.4)\\
O^{(j)} \quad &= \quad \arg\max_{O} P' \log(\alpha \cdot H^{(i)} + (1-\alpha) \cdot O) && (5.5)
\end{aligned}$$

The optimal H and O models at step j are obtained by applying the EM algorithm, taking as reference the model P' defined above. The initial model $O^{(0)}$ is set to the model of opinion words Q, and α is set to 0.5. Finally, by applying some threshold over the H and O models we obtain the vocabularies to be used for facet and sentiment indicators, respectively.

5.6.3 Automatic linking of data

Opinion posts (e.g., product reviews, tweets) usually consist of free text fields where users express their opinions. Therefore, opinion facts are usually expressed in these fields as natural language expressions. In order to extract these expressions and mapping them to the infrastructure, it is necessary to define an automatic semantic annotation process. In our proposal, this process consists of the following phases:

1. Segment the text into sentences.

2. Recognize chunks corresponding to facet expressions and those corresponding to opinion expressions.

3. Find the relations between facet and opinion expressions.

4. Calculate the polarity score of the opinion expressions.

5. Link facet expressions to published facets.

6. Finally, generate the opinion fact for each pair facet-opinion expressions and the information related to them (i.e., polarity and links).

Given a sentence, it is first cleaned and represented as a plain sequence of words. To chunk the sentence, we take into account four categories: facet words, sentiment indicator words, shifters, and connector words. Facet words and sentiment indicator words are extracted from the corresponding datasets of SLOD-BI (rdfs:label statements). Additionally, we also identify words that can change the valence of the polarity assigned to sentiment indicators, like negations (“not”, “never”, “none”, etc.), intensifiers (“deeply”, “very”, “little”, “rather”, etc.), modal shifters (“might”, “possibly”, etc.), and presuppositions (e.g., “lack”, “neglect”, “fail”, etc.) These words constitute the lexicon of shifters (Polanyi and Zaenen, 2006). Finally, connector words are those that connect words to express concepts (e.g., prepositions). In this way, facet expressions are sub-sequence of consecutive words categorized as either facet or connector, whereas opinion expressions are subsequence of consecutive words categorized as either sentiment expression or shifter.

Once facet and opinion expressions are identified, each facet expressions must be associated to their opinion expressions. Accurate results can be obtained by using dependency analysis, thus assigning to each facet expression the opinion expressions whose words are syntactically related to the facet expression words. However, this operation is time consuming and dependent of the language of the posts. A simpler heuristic consists of just taking the opinion expressions adjacent to the facet expression, and checking if they entail each other by applying the statistical model described in the previous section. For the current prototype of the infrastructure, we have applied this simple strategy.

Each opinion expression is analyzed to assign a polarity. A simple algorithm to perform this analysis consists of the following steps: first assign each sentiment indicator word to its polarity score, then invert the sign of the words affected by the shifters, and finally sum all the scores. Notice that the polarity score of a word can depend on its context (i.e., the facet expression to which it is assigned).

In order to link the extracted data from posts to the resources of the corresponding datasets, we use the concept retrieval technique described in (Berlanga et al., 2010). Basically, given a text chunk T associated to either a feature or opinion expression, we score the candidate resources R, whose labels are denoted with $labels(R)$, as follows:

$$\begin{aligned} score(T,R) &= max_{L \in labels(R)}(info(T \cap L)/info(L)) && (5.6) \\ info(S) &= -\sum_{w \in S} log(P_{bg}(w)) && (5.7) \end{aligned}$$

In these functions, both text chunks and labels are expressed as sets of words. The *info* measure accounts for the relevance of matched words w.r.t. candidate resources, which is captured by its inverse frequency in a background corpus $P_{bg}(w)$. In practice, strings T and L are previously normalized by applying

Operator type	Input	Output	Description
Extractor	Web data	CSV	Extracts tabular data from Web data (HTML, XML, JSON, etc.) by scrapping the sources.
StatisticalModel	CSV	CSV	Given a CSV, it estimates either a unigram or a bigram model from the text-rich columns.
StatisticalRefinement	CSV	CSV	Given a series of input models, it generates their refined models (see section 5.6.2).
SentimentAnalyzer	CSV	CSV	Automatically extracts facet and opinion expressions from a text column, and calculates the polarity of each extracted facet expression.
Linker	CSV	CSV	Adds a column with the URIs of the entities recognised in a text column. The resulting column can be multi-valued.
Ungroup	CSV	CSV	Generates a "flat" CSV by performing the Cartesian product on the selected multi-valued columns.
RDFizer	CSV	RDF	Produces an RDF triple collection by taking one column as subject, and the rest as objects with the column name as predicate.

Table 5.8 Proposed ETLink operator types.

stopword removal, case lowering and word lemmatizing, in order to favour their match.

Finally, the top scored resources R whose scores are greater than a given threshold and that best cover the chunk T are selected to link the opinion fact to the corresponding datasets. For example, the feature expression "my 308sw" would be linked to the resource "slod:Peugeot_308SW".

5.6.4 ETLink processes

Similarly to traditional DWs, we propose to populate the SLOD-BI infrastructure by means of ETL processes. An ETL process consists of a data flow that periodically extracts data from the sources, and transforms them into elements of the DW (i.e., dimensions and facts). The processing units of an ETL are called operators, which consume and produce tabular data. Operators perform SQL-like operations (e.g., selection, join, union, group by, and so on), as well as other data transformations such as concatenation/split of columns, function application to columns, and so on.

In our scenario the nature of the extraction, transformation and load phases are completely different from traditional DWs. Firstly, extraction nodes must be

ETLink process	**Workflow**	**Frequency**
Item Process	Extractor (ReferenceExternalDataset) → I RDFizer(I) → ItemDataset	Low
OpinionFacet Process	Extractor (ReferenceCorpora)→TextCollection Extractor (ReferenceOpinionLexicon)→WordSet StatisticalModel(TextCollection)→ T StatisticalModel(WordSet)→ Q StatisticalRefinement(T, Q)→(H, O) Select(O, threshold1)→FacetWords Select(H, threshold2)→OpWords RDFizer(FacetWords)→FacetDataset RDFizer(OpWords)→OpinionDataset	Low
PostFact Process	Extractor (OpinionStream)→ CSV Linker(CSV)→ CSV′ RDFizer(CSV′)→PostFactDataset	High
OpinionFact Process	Extractor (OpinonStream)→ CSV SentimenAnalyzer(CSV)→ CSV′ Linker(CSV′)→ CSV″ Ungroup(CSV‴)→ CSV^{iv} RDFizer(CSV^{iv})→OpinionFact Dataset	High
SocialFact Process	Extractor (OpinionStream)→ CSV Linker(CSV)→ CSV′ RDFizer(CSV′)→SocialFactDataset	Medium

Table 5.9 Summary of the ETLink processes for the SLOD-BI infrastructure.

able to deal with semi-structured web formats, which usually require scrapping to obtain structured data. These nodes should also connect with web services (APIs) to query and extract data from social networks. Transformations are also different from traditional ETL as they mainly rely on text-processing operations to generate sentiment data, and require frequent look-ups to the data infrastructure to link the produced data. Finally, in the load phase all data must be expressed as RDF according to the data infrastructure schemata. For these reasons, we call these data flows ETLink processes to distinguish them from the traditional ones.

The implementation of ETLink processes follows the same spirit as *pygrametl* (Thomsen and Bach Pedersen, 2009). Broadly speaking, *pygrametl* provides a series of Python classes for performing data transformations and for populating DW structures (i.e., dimensions and facts). Data flows are then specified with Python scripts using these classes. In our approach, workflow operators consume and produce either tabular data (CSV) or RDF triples. Instead of using DW structures, we use RDF primitives (RDFLib library) to generate the intermediate data, and SPARQL to perform the required look-up operations. Moreover, we provide operators to perform both sentiment analysis and data linkage.

In the current implementation, ETLink processes are designed as workflows of web data services. Thus, each operator of a concrete ETLink process is uniquely identified with a URI with which other operators can interact as either providers or consumers. In this way, third party tools, like syntactic parsers or polarity cal-

Source	Post facts	Opinion facts	
		Positive	Negative
Car and Driver, Auto Express, WhatCar?	1.038	8.117	
		5.017	3.100
Twitter (4 months)	34.236	21.037	
		14.610	6.427

Table 5.10 Statistics of opinion posts processed.

culators, may take place in this platform as wrapped services. Table 5.8 shows the main ETL operator types involved in ETLink processes. In this table, we have not included the SQL-like operators, which are also available. Operators for specific ETLink processes are configured to properly deal with the data infrastructure, and to perform the specific task they are aimed at. Table 5.9 shows the main steps of the ETLink processes involved in the infrastructure data provisioning. The implementation of some of these operators has been already described along Section 5.6.

5.7 Evaluation

To populate the SLOD-BI infrastructure we have selected a subset of opinion posts from several social media sources of information specialized on vehicles and from Twitter. Table 5.10 summarizes the main statistics. Although there are much more opinion facts extracted from Twitter, opinion facts from specialized forums exhibit a much higher quality. In global, there are much more positive comments than negative ones, when the usual situation is that negative comments dominate social sentiment data. This seems a particularity of this domain (cars), where customers are usually satisfied with their vehicles.

According to the structural view of SLOD-BI in Figure 5.3, the inner ring must be populated with the vocabularies and datasets for the car rental domain. The construction of the facets and the sentiment indicators components has been performed as explained in Section 5.6.2. For example, given a stream of opinions about cars, some relevant aspects are *interior*, *engine*, *cost*, *consumption*, etc. From sentences like “The interior design is attractive” or “The interior is superb quality and just so comfortable” we can extract the facet “interior” and the positive sentiment indicators “attractive” and “superb quality”. In the use case developed in the following section, facets will be classified into six sentiment topics useful for analysis. Therefore, aspects such as “interior”, “style” and “dashboard” will belong to the “design” topic, whereas aspects such as “clutch”, “wheel” and “gearbox” will belong to the “mechanical” topic, and so on. The prototype of this dataset can be accessed through the SPARQL endpoint `http://krono.act.uji.es/SLOD-BI/sparql`.

The rest of the datasets of the SLOD-BI infrastructure (opinion facts, items, etc.) are populated and linked by semantically annotating and processing the post facts as explained in Section 5.6.3. As a result, Figure 5.5 shows an excerpt of an opinion fact that has been extracted from a post fact and linked in the

SLOD-BI infrastructure.

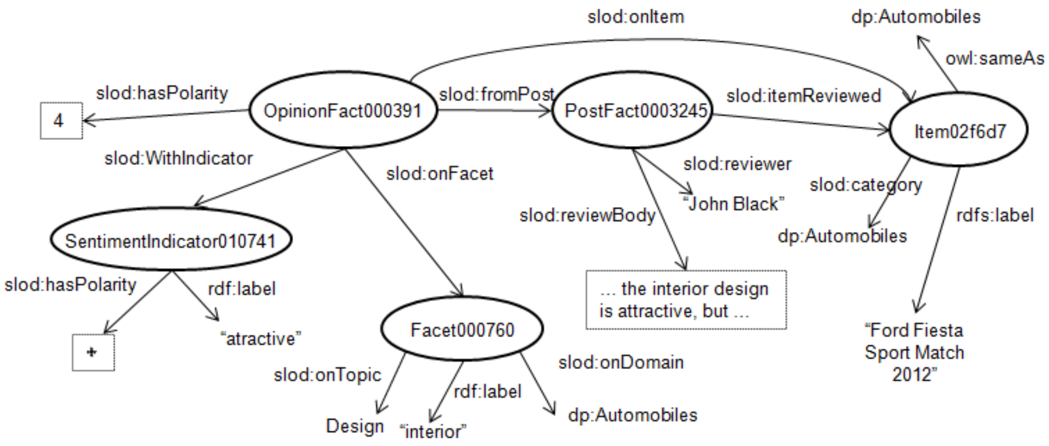


Fig. 5.5 Example of opinion fact in the SLOD-BI infrastructure.

5.7.1 Quality results

As semantic data is automatically extracted without any supervision, it is necessary to evaluate the quality of the generated data, as well as to find the best parameter settings of the learning algorithms to achieve good enough results. For this purpose, we have built two reference lexicons (unigram models), for facet and sentiment words respectively, restricted to a particular domain. To build the lexicon of sentiment words, we have downloaded and merged more than ten lists of opinion word lists that are freely available. The probability of each word is estimated from a large corpus of reviews of the specific domain. To build the lexicon of facet words, we manually chose a set of Wikipedia categories falling within the target domain, and then selected all Wikipedia entries having at least one of these categories.

With these two reference lexicons, we can evaluate the quality of the automatically extracted sentiment and facet expressions. Firstly, each word of the constructed vocabulary (Section 5.6.2) is classified as either facet or sentiment according to its probabilities in the reference lexicons. The overall precision of the method is then calculated as the total number of rightly classified words divided by the total number of classified words. Notice that some words may remain unclassified because either they do not appear at the reference lexicons or they cannot be statistically classified. Table 5.11 shows the precision results for the automatically generated vocabularies for facets (H) and sentiment words (O) in the "cars" domain. Results are shown with respect to the probability threshold applied to the models.

We can see that the quality achieved by sentiment words is quite good across the different probability thresholds. Results are not so good for the facets vocabulary, and this is because many words affected by sentiment indicators do not belong to the domain (e.g., expressions like "I had a nice day"). In order to improve the quality of this vocabulary we make use of the entailments derived from the target collection and BabelNet (Navigli and Ponzetto, 2010). Thus, for

		$p > 10^{-2}$	$p > 10^{-3}$	$p > 10^{-4}$	$p > 10^{-5}$	$p > 0$
Facet words	Precision	**0.69**	0.67	0.658	0.659	0.659
	#Words	627	2566	2936	2939	2961
	%Unclass.	0.306	0.4	0.45	0.45	0.45
Sentiment words	Precision	0.96	**0.974**	0.855	0.837	0.7
	#Words	298	1564	1955	1990	2132
	%Unclass.	0.0004	0.004	0.067	0.077	0.126

Table 5.11 Precision results of the generated vocabularies for the "cars" domain.

the final vocabulary we just consider those facet words that participate in at least an entailment of each translations model. As a result, the facet vocabulary is reduced to 638 words, achieving a precision of 0.93.

5.8 An Example of Social Analysis with SLOD-BI

To demonstrate our proposal, we have developed a prototypical SLOD-BI infrastructure for the car rental domain. At the core of each rent-a-car company lies the idea of providing their customers cost effective and quality services. This vision must be reflected on each of their business activities, which range from accepting reservations for new and existing customers to providing cars to customers, handling car upgrades when there is a shortage of cars or selecting the best promotional offer plan for each customer, among others.

To ensure business success, companies often have a series of strategic goals, such as optimum utilization of resources, customer satisfaction or controlling costs, which are materialized by more specific and measurable objectives. The objectives are set up as the result of a decision making process, which usually involves complex analytical queries over corporate data. The most established approach is to use a DW to periodically store information subject to analysis. In the case of a car rental company, the DW schema to analyze rental agreements could be similar to the one proposed in (Frias et al., 2003), where typical analysis dimensions include the rented vehicles, locations, customer features, etc. In order to make decisions, analysts often request the generation of reports involving analytical queries, e.g., number of rental agreements per location and time or preferred rented vehicles by location.

Apart from traditional analytic queries involving corporate data, there is a need to get more insight of the business internal processes in real time to be able to react more efficiently. In particular, customer satisfaction has become the greatest asset to success and there is a growing need of knowing customers' opinions about the companies' products and services. In this way, companies are able to dynamically integrate corporate data with relevant social data to analyze the answer of customers to its strategic decisions or to predict the demands of the market.

For a successful analytical experience, the company must specify the most important topics in its items (products or services) that require some sentiment analysis. In our use case, the company is interested in knowing people's opinion about the vehicles that they offer for renting, therefore, they have set up six sentiment topics that they consider of interest such as *comfortability*, *safety*, *driving perception*, *design*, *mechanical issues* and *price*. By analyzing people's opinion of their vehicles with respect to these topics, the company is able to detect vehicles implying high maintenance costs, the preferred vehicles by their design, etc.

Once the SLOD-BI is set up for the car rental domain, sentiment data can be consumed by means of the data service layer in order to produce the required data for the analytical tools. In the following, we present a series of examples of interesting analytical queries over the SLOD-BI that can be integrated with corporate data.

1. The analyst has executed a query over the corporate DW to find out the top rented cars by location. However, they would like to gain more insight by aligning the top rented cars with the people's opinion about such cars with respect to the design, to check if design is a relevant aspect for the customers behind those rentals. Figure 5.6 shows people's opinion (i.e., polarity) of different cars with respect to the sentiment topic *design*. This graph is the result of executing a query using the SPARQL endpoint service provided by SLOD-BI. The query aggregates the polarities of all the aspects classified under the topic *design*. Notice that whereas the first car has a high positive polarity, the last four cars have negative polarity, meaning that users are not happy with the design aspects of such cars.

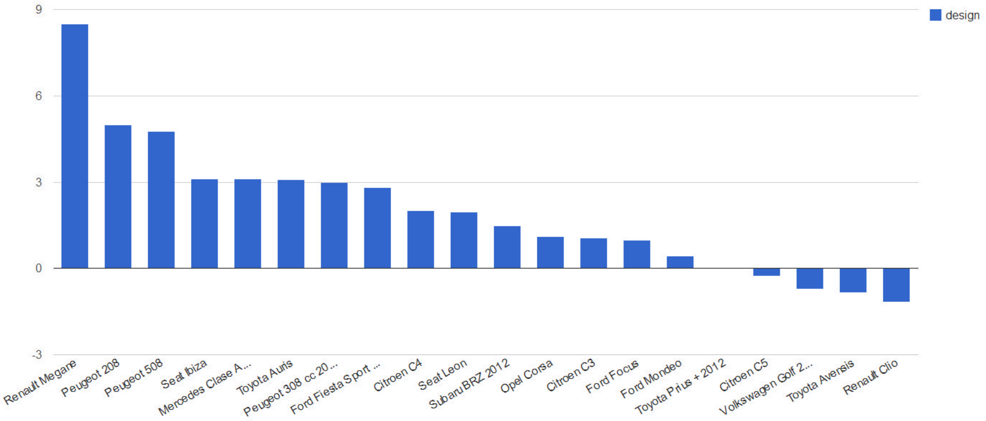


Fig. 5.6 Opinion of cars w.r.t. the sentiment topic *design*.

2. The company is interested in acquiring new fleet, but first, they would like to analyse people's opinion about cars with respect to mechanical issues, in order to avoid the acquisition of cars that usually involve more mechanical problems. Figure 5.7 shows the result of aggregating people's opinion about the topic *mechanical issues*. Notice that the last two cars show a high negative polarity, and therefore, the acquisition of these cars should be avoided.

3. The firm Peugeot has offered the rental company a special price if they acquire more than 10 units of the "Peugeot 208". However, the company

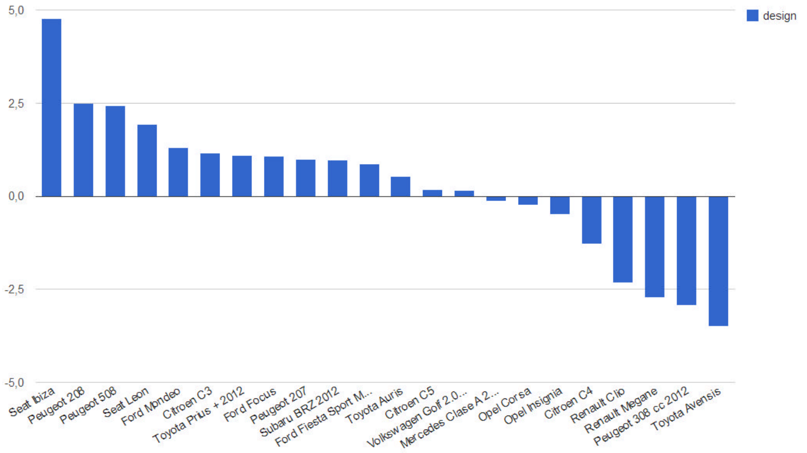


Fig. 5.7 Opinion of cars w.r.t. the sentiment topic *mechanical issues*.

would like to know people's opinion about this specific car with respect to the topics that they consider relevant. Figure 5.8 shows the results in the form of a bar chart. From the graph, we observe that *design* and *safety* are the highest rated aspects, whereas *price* is the lowest.

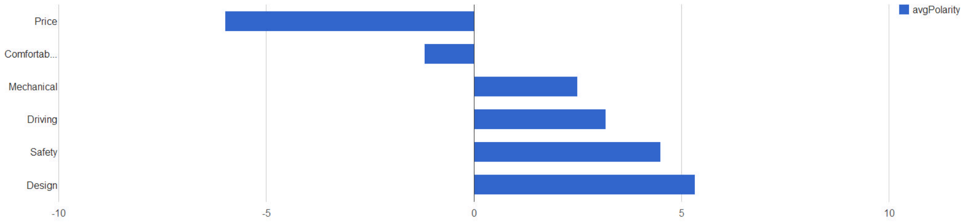


Fig. 5.8 Opinion about "Peugeot 208" w.r.t. each sentiment topic of interest.

4. Finally, the company is interested in blending corporate and social data to get some insights about how design opinions can affect to the number of contracts with respect to the company fleet. For this purpose, the popular corporate analytical tool KNIME (`http://www.knime.org/`) is used. In a few words, KNIME is an open source data analytics, reporting and integration platform with a graphical user interface that allows assembly of nodes for data pre-processing, modelling, analysis and visualization. We have implemented a node for performing SPARQL queries on SLOD-BI, and then we have used the workflow nodes of KNIME to integrate the social and corporate data. The resulting workflow is shown in Figure 5.9a. The bottom node queries corporate data to extract the number of rentals by car during 2013. The result is a table with two columns, the car and the number of rentals. The RDF *QueryAP* node executes a SPARQL query over the data service layer of the infrastructure to extract sentiment data about the topic "*design*". After some processing, the *Joiner* node merges the two tables by the car column and the resulting chart (Figure 5.9b) displays the number of car rentals (in blue) vs. the aggregation opinion on "Design" aspects (in red) by car. In general, we observe a positive correlation between the two variables, as the mostly rented cars (i.e., Renault Megane,

Peugeot 208 and 508) are the ones with highest rating of design aspects.

5.9 Conclusions

This paper has presented SLOD-BI, a new semantic data infrastructure for capturing and publishing sentiment data to enable Social BI. The infrastructure components are designed to cover the main BI patterns we have identified for analysing both corporate and social data in an integrated way. The infrastructure also provides the functionality required to perform massive opinion analysis, for example the automatic extraction of sentiment data from posts, and their linkage to the infrastructure. As a result, users will be able to incorporate opinion-related dimensions in their analysis, which is out of reach of traditional BI.

For future work, we will study the performance of complex queries over the SLOD-BI infrastructure, for example OLAP-like operations, which may require massive data processing methods. For this purpose, the datasets in the inner ring of SLOD-BI must be properly partitioned and distributed according to the BI demands. For example, datasets should be partitioned with respect to domains and time slices. Moreover, functional map-reduce implementations (Dean and Ghemawat, 2008) can process such distributed partitions and parallelize complex analysis operators such as filter, join and aggregate (Sridhar et al., 2009). Additionally, to speed-up costly operations within the inner SLOD-BI datasets, ad-hoc indexing mechanisms should be defined. More challenging is however, to efficiently perform BI operations involving external datasets, as we do not have control over them. Finally, to extend the functionality of the infrastructure we aim at linking data to multi-lingual resources such as BabelNet. We also plan to introduce services for transforming the query results to the RDF Data Cube vocabulary, so that they can be included in tools designed for this vocabulary.

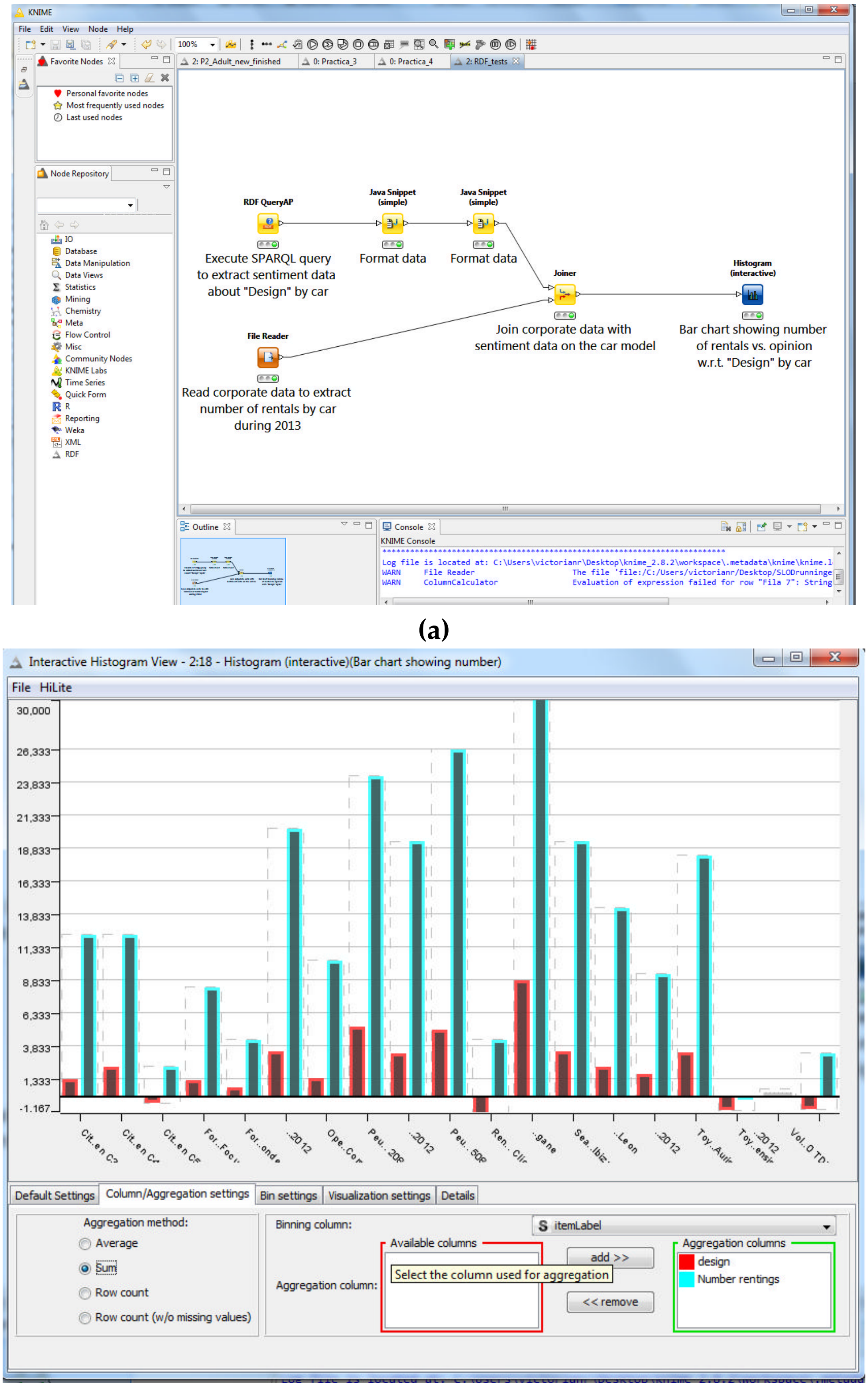
Another issue to be addressed in the future work is how the infrastructure can manage the high dynamicity of certain topics in some domains. Unfortunately, the problem of adapting sentiment analysis tools to evolving topics has been poorly treated in the literature. Moreover, the validation of a self-adapted approach for sentiment analysis requires a huge amount of data recorded during a long time in order to detect fast iteration cycles.

Finally, another open issue of the infrastructure is the mapping of automatically extracted facets to corporate sentiment topics. Currently, these mappings are manually performed, but this process has a high cost and is prone to errors. In the future work we plan to study semi-automatic methods for performing these crucial mappings of the infrastructure.

5.10 Acknowledgements

This work has been partially funded by the "Ministerio de Economía y Competitividad" with contract number TIN2011-24147. We would like to thank Avelino

Font for helping us in implementing the first prototype of the SLOD-BI infrastructure.



Execute SPARQL query to extract sentiment data about "Design" by car

Format data

Format data

Join corporate data with sentiment data on the car model

Bar chart showing number of rentals vs. opinion w.r.t. "Design" by car

Read corporate data to extract number of rentals by car during 2013

KNIME Console

```
*****************************************************************************
Log file is located at: C:\Users\victorianr\Desktop\knime_2.8.2\workspace\.metadata\knime\knime.l
WARN     File Reader                           The file 'file:/C:/Users/victorianr/Desktop/SLODrunninge
WARN     ColumnCalculator                      Evaluation of expression failed for row "Fila 7": String
```

(a)

Interactive Histogram View - 2:18 - Histogram (interactive)(Bar chart showing number)

Aggregation method:
○ Average
● Sum
○ Row count
○ Row count (w/o missing values)

Binning column: S itemLabel

Aggregation column:

Available columns: Select the column used for aggregation

Aggregation columns:
🟥 design
🟦 Number rentings

(b)

Fig. 5.9 Blending corporate and social data from SLOD-BI through KNIME.

Chapter 6

Conclusions

This thesis addresses the problem of aspect-based sentiment summarization, as well as providing methodologies to integrate sentiment data extracted from opinion posts into BI models.

The main contributions of the thesis, the obtained results and future work are summarized in the next sections.

6.1 Main contributions

The main contributions of this thesis can be summarized as follows:

1. Based on the idea that product features can be modeled by means of a statistical language model (see our second main hypothesis in Section 1.3), a preliminary methodology to model a language of product features from a set of opinion words is proposed in Chapter 2. The main novelty of the methodology is that it is based on a stochastic mapping model between words, also called self-translation model, which aims to capture the entailment between opinions and their targets in the context of opinion posts. The proposed methodology is also able to obtain a ranking of opinion words from the posts.

2. From this preliminary methodology, and taking into account our two main hypotheses, a new domain-independent methodology to model and extract product features as well as extracting their corresponding opinions is introduced.

3. As part of the methodology, a novel method to induce a lexicon model of opinions for a target product/service is introduced based on a kernel function between word distributions. In our work, the lexicon model is employed to learn a refined language model of product features from which we finally implement a method to retrieve both the features and their respective opinions.

4. Based on the methodology to extract product features and their corresponding opinions, a new methodology is introduced to integrate sentiment data from opinion posts into a BI model implemented as a traditional corporate DW.

5. To perform the integration, a new fully-automatic and unsupervised semantic annotation method is proposed to perform feature unification for populating the DW. Besides, a polarity assessment method is proposed to conform opinion facts.

6. Finally, new integration challenges are identified and addressed by means of a novel semantic data infrastructure for BI, called SLOD-BI, which is principled on the LOD initiative.

It is also worth mentioning that the methodologies proposed in this work do not need exhaustive natural language processing (except for POS-tagging/ lemmatization) to obtain a good performance in the retrieval of sentiment data. These methodologies can be applied to any language and domain given a seed set of general-domain opinion words as input.

6.2 Results

From the articles reviewed in previous chapters, we obtained the main results of this thesis, which corroborate our main hypotheses(see Section 1.3). Besides, we can response to our research questions as follows:

RQ1 *Is there a (statistical) language model capable of modeling a lexicon of opinion words for a given opinion domain or product from a seed set of general domain opinion words?*

Response: Yes, there is such a language. In Chapter 3, it is shown how a kernel-based model can be employed to successfully learn a statistical language model of domain opinions from a seed set of opinion words with an averaged precision ranging from 0.77 to 0.81 in opinion datasets in English, and about 0.85 in Spanish.

RQ2 *If so, can we model the collection of opinion targets as another language model from the lexicon model of opinions? Is it possible to define the language model of opinion targets as a translation from the lexicon model?*

Response: Yes, we can also model the opinion targets in an opinion dataset by means of a language model that can be learned using successive linear transformations of a model of opinion words.

In Chapter 2, we show that it is possible to successfully learn a language model of features from a set of opinion words. Experimentally, the models learned in this chapter obtained precision values ranging from 0.80 to 0.96 at top 50, and values from 0.78 to 0.95 at top 100 words. Here, the precision values correspond to the percentages of words in the top of the ranks that are part of a feature name.

In Chapter 3, the models of features were employed to obtain a ranking of multi-word phrases to represent features. The MAP values for the obtained rankings were about 0.20, which indicates that the feature models were accurately enough learned to retrieve the product features.

RQ3 *Can we effectively retrieve the subjective data (i.e., structures with the form product feature-opinion) from the above models?*

Response: In addition to measure the performance of the retrieval of (multiword) product features in terms of MAP in Chapter 3, we also assess the performance of retrieving the opinions corresponding to the product features. In this case, the obtained MAP values were overall above 0.60. This corroborates that the overall retrieval of subjective data (i.e., the features and their associated opinions) can be carried out in an effective manner; which, again, corroborates our main hypotheses and usefulness of the learned models to retrieve the sentiment data.

RQ4 *Regarding the issue of storing and publishing sentiment data, can we integrate the extracted sentiment data into a traditional corporate data warehouse (DW) to enable BI? Which are the main challenges to achieve this integration? Is this solution suitable to dynamic scenarios where both the data sources and the user requirements may change over time?*

Response: In Chapter 4, we have introduced a new methodology to integrate sentiment data obtained by means of the methodology in Chapter 3 into a BI model implemented by means of a corporate DW. To perform such an integration, we had to face several challenges; being both the semantic annotation and the scoring of sentiment facts the most important ones. Thus, we firstly developed a semantic annotation method to perform the unification of features, and then we proposed a preliminary approach to measure the semantic orientation of the opinion facts (i.e., tuples with the form feature-opinion).

In addition, In Chapter 5 we have identified new challenges on the integration of sentiment data into BI models concerning the dynamic scenario of VoC and VoM. Thus, we have proposed SLOD-BI as an open data infrastructure based on LOD where BI data can be linked to external sources on demand, without being attached to predefined (rigid) data structures or multidimensional schema.

6.2.1 Scientific publications

The following publications are directly related to the development of this thesis:

- Rafael Berlanga, Lisette García-Moya, Victoria Nebot, María José Aramburu, Ismael Sanz, and Dolores María Llidó. SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence. *International Journal of Data Warehousing and Mining (IJDWM)*, 11(4):1–28, 2015

- Rafael Berlanga, María José Aramburu, Dolores M Llidó, and Lisette García-Moya. Towards a semantic data infrastructure for social business intelligence. In *New Trends in Databases and Information Systems*, pages 319–327. Springer International Publishing, 2014

- Lisette Garcia-Moya, Henry Anaya-Sanchez, and Rafael Berlanga-Llavori. Retrieving product features and opinions from customer reviews. *IEEE Intelligent Systems*, 28(3):19–27, 2013

- Lisette García-Moya, Shahad Kudama, María José Aramburu, and Rafael Berlanga. Storing and analysing voice of the market data in the corporate data warehouse. *Information Systems Frontiers*, 15(3):331–349, 2013
- Rafael Berlanga Llavori, Dolores María Llidó, Lisette García-Moya, Victoria Nebot, María José Aramburu, and Ismael Sanz. i-SLOD: Towards an Infrastructure for Enabling the Dissemination and Analysis of Sentiment Data. In *KDIR/KMIS*, pages 214–219, 2013
- Lisette Garcıa-Moya, Rafael Berlanga-Llavori, and Marıa José Aramburu-Cabo. Extraction and ranking of product aspects based on word dependency relations. In *CERI*, 2012
- Marıa Pérez, Lisette Garcıa-Moya, and Rafael Berlanga. A translation model for facet-based retrieval in open registries. In *CERI*, 2012
- Lisette García Moya, Rafael Berlanga Llavori, and Henry Anaya Sánchez. Learning a statistical model of product aspects for sentiment analysis. *Procesamiento del Lenguaje Natural*, 49:157–162, 2012
- Lisette García-Moya, Henry Anaya-Sánchez, and Rafael Berlanga-Llavori. Combining Probabilistic Language Models for Aspect-Based Sentiment Retrieval. In *Proceedings of the 34th European Conference on Information Retrieval*, volume 7224 of *Lecture Notes in Computer Science*, pages 561–564. Springer-Verlag, 2012
- Shahad Kudama, Rafael Berlanga Llavori, Lisette Garcia-Moya, Victoria Nebot, and Maria Jose Aramburu Cabo. Towards tailored semantic annotation systems from Wikipedia. In *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on*, pages 478–482. IEEE, 2011b
- Lisette García Moya, Shahad Kudama, María José Aramburu Cabo, and Rafael Berlanga Llavori. Integrating web feed opinions into a corporate data warehouse. In *Proceedings of the 2nd International Workshop on Business intelligencE and the WEB*, pages 20–27. ACM, 2011
- Lisette García-Moya, Henry Anaya-Sánchez, Rafel Berlanga, and María José Aramburu. Probabilistic ranking of product features from customer reviews. In *Pattern Recognition and Image Analysis*, pages 208–215. Springer, 2011

6.3 Future work

As future work, we mainly consider two research directions. Firstly, we consider to address the problem of polarity classification of the opinion facts (i.e., the tuples feature-opinions) that we extract by means of the methodology proposed in Chapter 3.

Despite a preliminary approach has been employed in Chapter 4 in order to store the opinion facts in a corporate DW, there are some open issues in this approach. Firstly, we have assumed that opinion words always have the same base polarity, which can only be changed through valence shifters. However, it is well-known that some opinion words have different polarities for different products (e.g. “large” for “mobile phone” and “screen”). Thus, we are planning to regard context-dependent indicators similarly to (Lu et al., 2011) in order to induce polarities.

We may also consider other kinds of polarity shifters such as those ones introduced by irony and figurative language in general since ironic opinions about a product can have a high (negative) impact in its sales (which is important from the BI perspective). This issue has been recently addressed in a shared task at SemEval on sentiment analysis with figurative language (Ghosh et al., 2015), as well as by several recent approaches (as for instance (Reyes and Rosso, 2014)).

In addition, a pre-processing step could be also added to the proposed methodology in order to filter out those customer reviews that are deceptive (Hernández et al., 2014; Ott et al., 2011) to consider only truthful reviews as input for the aspect-based summarization.

Secondly, we consider to address the problem of assessing the performance of the integration of sentiment data into the BI models. For this purpose, we regard to evaluate the use of sentiment data in the prediction of BI measures (e.g. sales or profits). This will favor the automatic induction of corporate sentiment topics.

Finally, another future work consists in applying the learned models and kernels as input embeddings for deep learning sentiment classifiers, which have shown very good precision scores (dos Santos and Gatti, 2014).

References

R. Agrawal and R. Srikant. Fast algorithm for mining association rules. In *Proceedings of the International Conference on Very Large Data Base*, volume 487, page 499, 1994.

Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

Nikolay Archak, Anindya Ghose, and Panagiotis G. Ipeirotis. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD*, pages 56–65, 2007.

Adam Berger and John Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*, pages 222–229, Berkeley, CA., 1999.

Rafael Berlanga, Victoria Nebot, and Ernesto Jimenez. Semantic annotation of biomedical texts through concept retrieval. *Procesamiento del Lenguaje Natural*, 45:247–250, 2010.

Rafael Berlanga, María José Aramburu, Dolores M Llidó, and Lisette García-Moya. Towards a semantic data infrastructure for social business intelligence. In *New Trends in Databases and Information Systems*, pages 319–327. Springer International Publishing, 2014.

Rafael Berlanga, Lisette García-Moya, Victoria Nebot, María José Aramburu, Ismael Sanz, and Dolores María Llidó. SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence. *International Journal of Data Warehousing and Mining (IJDWM)*, 11(4):1–28, 2015.

Michael W. Berry and Malu Castellanos. *Survey of Text Mining II: Clustering, Classification, and Retrieval*. 1 edition, 2007. ISBN 1848000456, 9781848000452.

M. Bhide, V. Chakravarthy, A. Gupta, H. Gupta, M. Mohania, K. Puniyani, P. Roy, S. Roy, and V. Sengar. Enhanced Business Intelligence using EROCS. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 1616–1619, 2008.

Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227, 2009.

P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

References

Volha Bryl, Claudio Giuliano, Luciano Serafini, and Kateryna Tymoshenko. Supporting natural language processing with background knowledge: Coreference resolution case. In *International Semantic Web Conference (1)*, pages 80–95, 2010.

E. Cambria, Y. Song, H. Wang, and N. Howard. Semantic Multi-Dimensional Scaling for Open-Domain Sentiment Analysis. *Intelligent Systems, IEEE*, 29(2): 44–51, 2013. doi: 10.1109/MIS.2012.118.

Erik Cambria, Catherine Havasi, and Amir Hussain. SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In *FLAIRS Conference*, pages 202–207, 2012.

G. Carenini, R. Ng, and A. Pauls. Multi-document summarization of evaluative text. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 305–312, 2006.

Michael J Carey, Nicola Onose, and Michalis Petropoulos. Data services. *Communications of the ACM*, 55(6):86–97, 2012.

E. F. Codd. Providing OLAP to user-analysts: An IT mandate, 1993.

Edgar F Codd, Sharon B Codd, and Clynch T Salley. Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. *Codd and Date*, 32, 1993.

Fermín L. Cruz. *Extracción de opiniones sobre características: un enfoque práctico adaptable al dominio*. PhD thesis, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Sevilla, October 2011.

F.L. Cruz, J.A. Troyano, F. Enríquez, F.J. Ortega, and C.G. Vallejo. A knowledge-rich approach to feature-based opinion extraction from product reviews. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 13–20. ACM, 2010.

Roxana Dánger and Rafael Berlanga. Generating complex ontology instances from documents. *J. Algorithms*, 64(1):16–30, 2009.

Mark. Davies. Word frequency data from the Corpus of Contemporary American English (COCA). Downloaded from http://www.wordfrequency.info on June 01, 2011., 2011.

M.C. De Marneffe, B. MacCartney, and C.D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

Hongbo Deng, Michael R. Lyu, and Irwin King. A generalized Co-HITS algorithm and its application to bipartite graphs. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–248. ACM, 2009. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557051. URL `http://dx.doi.org/10.1145/1557019.1557051`.

J. Dillon, Y. Mao, G. Lebanon, and J. Zhang. Statistical Translation, Heat Kernels, and Expected Distance. In *Proc. of the 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

Xiaowen Ding, Bing Liu, and Philip S. Yu. A Holistic Lexicon-based Approach to Opinion Mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 231–240. ACM, 2008.

Cıcero Nogueira dos Santos and Maıra Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING), Dublin, Ireland*, 2014.

Koji Eguchi and Victor Lavrenko. Sentiment retrieval using generative models. In *Proceedings of the EMNLP 2006*, pages 345–354. Association for Computational Linguistics, 2006.

Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 19:1–16, 2007. ISSN 1041-4347. doi: http://doi.ieeecomputersociety.org/10.1109/TKDE.2007.9.

Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. Open information extraction from the web. *Commun. ACM*, 51:68–74, 2008. ISSN 0001-0782.

Leonor Frias, Anna Queralt, and Antoni Olivé Ramon. EU-Rent Car Rentals Specification. Technical report, Research report of Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya. LSI-03-59-R, 2003.

Adam Funk, Yaoyong Li, Horacio Saggion, Kalina Bontcheva, and Christian Leibold. Opinion analysis for business intelligence applications. In *Proceedings of the first international workshop on Ontology-supported business intelligence*, page 3. ACM, 2008.

Lisette García-Moya, Henry Anaya-Sánchez, Rafel Berlanga, and María José Aramburu. Probabilistic ranking of product features from customer reviews. In *Pattern Recognition and Image Analysis*, pages 208–215. Springer, 2011.

Lisette García-Moya, Henry Anaya-Sánchez, and Rafael Berlanga-Llavori. Combining Probabilistic Language Models for Aspect-Based Sentiment Retrieval. In *Proceedings of the 34th European Conference on Information Retrieval*, volume 7224 of *Lecture Notes in Computer Science*, pages 561–564. Springer-Verlag, 2012.

Lisette García Moya, Rafael Berlanga Llavori, and Henry Anaya Sánchez. Learning a statistical model of product aspects for sentiment analysis. *Procesamiento del Lenguaje Natural*, 49:157–162, 2012.

Lisette Garcıa-Moya, Rafael Berlanga-Llavori, and Marıa José Aramburu-Cabo. Extraction and ranking of product aspects based on word dependency relations. In *CERI*, 2012.

Lisette Garcia-Moya, Henry Anaya-Sanchez, and Rafael Berlanga-Llavori. Retrieving product features and opinions from customer reviews. *IEEE Intelligent Systems*, 28(3):19–27, 2013.

Lisette García-Moya, Shahad Kudama, María José Aramburu, and Rafael Berlanga. Storing and analysing voice of the market data in the corporate data warehouse. *Information Systems Frontiers*, 15(3):331–349, 2013.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and John Barnden. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proc. 9th Int. Workshop on Semantic Evaluation (SemEval-2015), Co-located with NAACL, Denver, Colorado*, pages 470–478, 2015.

Oren Glickman, Ido Dagan, and Moshe Koppel. A lexical alignment model for probabilistic textual entailment. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, pages 287–298. Springer-Verlag, 2006.

Adrien Guille, Hakim Hacid, Cécile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2): 17–28, 2013.

Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1): 1–136, 2011.

Donato Hernández, Manuel Montes-y Gómez, Paolo Rosso, and Rafael Guzmán. Detecting positive and negative deceptive opinions using pu-learning. *Information Processing & Management*, 51(4):433–443, 2014.

Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760, 2004a.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004b.

W. H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, 2005.

ISLA. The WIKIXML Collection. http://ilps.science.uva.nl/WikiXML/, 2010. URL `http://ilps.science.uva.nl/WikiXML/`.

Antonio Jimeno-Yepes, Ernesto Jiménez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(S-3), 2008.

Wei Jin, Hung Hay Ho, and Rohini K. Srihari. OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1195–1204. ACM, 2009. ISBN 978-1-60558-495-9.

Yohan Jo and Alice H. Oh. Aspect and Sentiment Unification Model for Online Review Analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 815–824. ACM, 2011.

Axel Johne. Listening to the Voice of the Market. *International Marketing Review*, 11(1):47–59, 1994.

José Kahan and Marja-Ritta Koivunen. Annotea: an open rdf infrastructure for shared web annotations. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 623–632, New York, NY, USA, 2001. ACM. ISBN 1-58113-348-0. doi: http://doi.acm.org/10.1145/371920.372166. URL `http://doi.acm.org/10.1145/371920.372166`.

Maryam Karimzadehgan and ChengXiang Zhai. Axiomatic Analysis of Translation Language Model for Information Retrieval. In *Advances in Information Retrieval*, volume 7224 of *Lecture Notes in Computer Science*, pages 268–280. Springer Berlin / Heidelberg, 2012.

Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval. *J. Web Sem.*, 2(1): 49–79, 2004.

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

Shahad Kudama, Rafael Berlanga, Lisette García, Victoria Nebot, and María José Aramburu. Towards tailored semantic annotation systems from Wikipedia . In *Proceedings of the DEXA Workshop*, DEXA 2011. IEEE, 2011a.

Shahad Kudama, Rafael Berlanga Llavori, Lisette Garcia-Moya, Victoria Nebot, and Maria Jose Aramburu Cabo. Towards tailored semantic annotation systems from Wikipedia. In *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on*, pages 478–482. IEEE, 2011b.

Bing Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

Bing Liu, Wynne Hsu, and Yiming Ma. Integrating Classification and Association Rule Mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, 1998.

Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. ARSA: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 607–614, 2007.

Rafael Berlanga Llavori, Dolores María Llidó, Lisette García-Moya, Victoria Nebot, María José Aramburu, and Ismael Sanz. i-SLOD: Towards an Infrastructure for Enabling the Dissemination and Analysis of Sentiment Data. In *KDIR/KMIS*, pages 214–219, 2013.

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. Rated Aspect Summarization of Short Comments. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 131–140. ACM, 2009.

Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 347–356. ACM, 2011. ISBN 978-1-4503-0632-4.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 171–180. ACM, 2007.

Eduardo Mena, Arantza Illarramendi, Vipul Kashyap, and Amit P Sheth. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *Distributed and parallel Databases*, 8(2):223–271, 2000.

Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM, 2007. ISBN 978-1-59593-803-9. doi: 10.1145/1321440.1321475. URL `http://dx.doi.org/10.1145/1321440.1321475`.

Lisette García Moya, Shahad Kudama, María José Aramburu Cabo, and Rafael Berlanga Llavori. Integrating web feed opinions into a corporate data warehouse. In *Proceedings of the 2nd International Workshop on Business intelligencE and the WEB*, pages 20–27. ACM, 2011.

Roberto Navigli and Simone Paolo Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics, 2010.

Radford M Neal and Geoffrey E Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics, 2011.

Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREC*, volume 10, pages 1320–1326, 2010.

B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, 2008. ISSN 1554-0669. doi: 10.1561/1500000011.

Juan Manuel Pérez, Rafael Berlanga, María José Aramburu, and Torben Bach Pedersen. R-Cubes: OLAP Cubes Contextualized with Documents. In *Proceedings of the IEEE 23rd international conference on Data engineering*, pages 1477–1478, 2007.

Juan Manuel Pérez, Rafael Berlanga, María José Aramburu, and Torben Bach Pedersen. Towards a Data Warehouse Contextualized with Web Opinions. In *Proceedings of the 2008 IEEE International Conference on e-Business Engineering*, pages 697 –702, 2008a.

Juan Manuel Pérez, Rafael Berlanga, María José Aramburu, and Torben Bach Pedersen. Contextualizing data warehouses with documents. *Decision Support Systems*, 45(1):77–94, 2008b.

Juan Manuel Pérez, Rafael Berlanga, Maria Jose Aramburu, and Torben Bach Pedersen. Integrating data warehouses with web data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 20(7):940–955, 2008.

Marıa Pérez, Lisette Garcıa-Moya, and Rafael Berlanga. A translation model for facet-based retrieval in open registries. In *CERI*, 2012.

Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. Learning Sentiment Lexicons in Spanish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

Livia Polanyi and Annie Zaenen. Contextual valence shifters. In JamesG. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 1–10. Springer Netherlands, 2006. ISBN 978-1-4020-4026-9. doi: 10.1007/1-4020-4102-0_1.

Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28. Springer, 2007.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding domain sentiment lexicon through double propagation. In *Proceedings of the IJCAI-09*, pages 1199–1204, 2009.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, 37 (1):9–27, 2011.

R. Eric Reidenbach. *Listening to the Voice of the Market: How to Increase Market Share and Satisfy Current Customers*. Crc Press, 2009.

Antonio Reyes and Paolo Rosso. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, 40(3):595–614, 2014.

Gamgarn Somprasertsri and Pattarachai Lalitrojwong. Automatic product feature extraction from online product reviews using maximum entropy with lexical and syntactic features. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*, pages 250–255. IEEE Systems, Man, and Cybernetics Society, 2008.

Radhika Sridhar, Padmashree Ravindra, and Kemafor Anyanwu. RAPID: Enabling Scalable Ad-Hoc Analytics on the Semantic Web. In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, pages 715–730. Springer-Verlag, 2009. ISBN 978-3-642-04929-3.

Philip J Stone, D C Dunphy, M S Smith, and D M Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*, volume 08. MIT Press, 1966a.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966b.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

Christian Thomsen and Torben Bach Pedersen. pygrametl: A powerful programming framework for extract-transform-load programmers. In *Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP*, pages 49–56. ACM, 2009.

Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120, 2008.

A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18, 2006.

Victoria Uren, Philipp Cimiano, Jose Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta, and Fabio Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14–28, 2006. ISSN 15708268. doi: 10.1016/j.websem.2005.10.002. URL `http://dx.doi.org/10.1016/j.websem.2005.10.002`.

Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 783–792, New York, NY, USA, 2010a. ACM. doi: http://doi.acm.org/10.1145/1835804.1835903.

Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010b.

Hongning Wang, Yue Lu, and ChengXiang Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.

Chih-Ping Wei, Yen-Ming Chen, Chin-Sheng Yang, and Christopher C. Yang. Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and E-Business Management*, pages 149–167, 2009.

Chih-Ping Wei, Yen-Ming Chen, Chin-Sheng Yang, and ChristopherC. Yang. Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and e-Business Management*, 8(2):149–167, 2010. doi: 10.1007/s10257-009-0113-9.

Adam Westerski, Carlos Iglesias, et al. Exploiting structured linked data in enterprise knowledge management systems: An idea management case study. In *Enterprise Distributed Object Computing Conference Workshops (EDOCW), 2011 15th IEEE International*, pages 395–403. IEEE, 2011.

Tak-Lam Wong and Wai Lam. Hot Item Mining and Summarization from Multiple Auction Web Sites. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 797–800, Washington, DC, USA, 2005. IEEE Computer Society.

Tak-Lam Wong and Wai Lam. Learning to extract and summarize hot item features from multiple auction web sites. *Knowl. Inf. Syst.*, 14(2):143–160, 2008.

Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1541, 2009.

Christopher C. Yang, Y. C. Wong, and Chih-Ping Wei. Classifying web review opinions for consumer product analysis. In *Proceedings of the 11th International Conference on Electronic Commerce*, pages 57–63, New York, NY, USA, 2009. ACM.

Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. Aspect ranking: identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, 2011.

Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1462–1470, 2010.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 56–65. Association for Computational Linguistics, 2010.