# i-SLOD: Towards an Infrastructure for Enabling the Dissemination and Analysis of Sentiment Data

Rafael Berlanga, Dolores Mª Llidó, Lisette García, Victoria Nebot,
María José Aramburu and Ismael Sanz
*Universitat Jaume I, Avda. Vicent Sos Baynat s/n, Castellón de la Plana, Spain*

Keywords: Opinion Analysis, Open Linked Data, Business Intelligence.

Abstract: This paper proposes a new data infrastructure for massive opinion analysis, called i-SLOD, from a Business Intelligence (BI) perspective. This infrastructure aims to allow analysts to re-use the existing review data about products and services publicly available in the Web. It should also take advantage from the external relationships of i-SLOD data in order to perform new exploratory analyses now unfeasible with traditional BI tools. We consider the adoption of Linked Open Data (LOD) technology to build this infrastructure. In this way, i-SLOD data will be published as distributed linked open data by using the RDF and OWL formats. Moreover, we propose to apply automatic semantic annotation to perform the basic tasks in i-SLOD, mainly the extraction of opinion facts from raw text, and linking opinion data to the i-SLOD and other related LOD datasets.

## 1 INTRODUCTION

The massive publication of opinions about product and services has produced a burst of methods for sentiment analysis (Liu, 2012). Most of these approaches directly deal with the review texts to identify global assessments (reputation) of certain products and services. They are mainly focused on detecting the subject of the opinion (e.g., some product or some aspect of it) as well as the orientation of the opinion (i.e., polarity). Massive mining of opinions allow obtaining good indicators about the Voice of the Market (García-Moya et al., 2013a). Due to the high interest of this kind of data, a good number of commercial tools have recently appeared in the market, for example Swotti, Radian6 Insight, Media Miser, Scout Labs, Wise Window and Sinthesio, to mention a few. Unfortunately, most of these tools just provide web reports targeted to end-users, and the sentiment data is not publicly available for third party applications.

Apart from the sentiment analysis approaches, there is also a great interest on publishing strategic data for Business Intelligence (BI) tasks within the Linked Open Data (LOD) cloud (Heath and Bizer, 2011). Initatives like *Schema.org* are allowing the massive publication of product offers as microdata, as well as specific vocabularies for e-commerce applications. Unfortunately, both worlds, sentiment data and LOD technology, have kept unconnected to each other until recently. Some preliminary projects such as MARL (Westerski and Iglesias, 2011) attempt to provide standarized schemas for expressing opinion data as linked data. However, 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 propose i-SLOD, a new data infrastructure for sentiment data aimed at satisfying the necessity of generating and analysing opinion data from a BI perspective in the context of the LOD initative.

## 2 i-SLOD ROAD MAP

Traditional BI assumes the existence of a controlled set of data sources, from which summarized data is obtained for decision making tasks. BI architectures usually rely on a data warehouse defined under a multidimensional model (i.e., just consisting of measures and dimensions), which is fed with data extracted from existing data sources by applying the

Berlanga R., Mª Llidó D., García L., Nebot V., José Aramburu M. and Sanz I..
i-SLOD: Towards an Infrastructure for Enabling the Dissemination and Analysis of Sentiment Data.
DOI: 10.5220/0004627902140219
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing* (KDIR-2013), pages 214-219
ISBN: 978-989-8565-75-4

so-called Extraction, Transform and Load (ETL) processes. Finally, data is summarized by applying efficient BI tools such as OLAP.

From a BI point of view, opinion data can be also multidimensionally modelled and analysed. For example, the reputation of a product, the most outstanding features of some product brand, or the opined aspects can be efficiently computed with OLAP-like operations (García-Moya et al., 2013a).

The main BI e-commerce patterns we consider in this project are summarized in Figure 1. Facts such as sales, offers and opinions account for spatio-temporal observations of some measure (e.g., units sold, units offered, number of positive reviews, and so on), whereas dimensions (labelled with 'D') account for the contexts of such observations. Dimensions can provide different detail levels (labelled with 'L'). In this paper we will mainly focus on the specification and generation of both review and opinion facts. Notice that every review produces two kind of sentiment facts: the global review assessment about the item (review fact), and the specific criticisms to the item features/aspects (opinion facts).

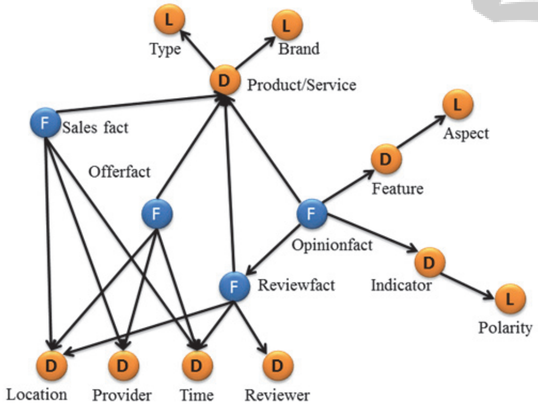

Figure 1: Main BI patterns over e-commerce facts.

In order to cover these patterns, the main components of i-SLOD data infrastructure are linked to each other as well as to other external related LOD datasets. Figure 2 shows the proposed architecture, where i-SLOD components are placed within the inner ring. The outer ring contains other LOD datasets and vocabularies (dotted boxes) that can be linked to the proposed infrastructure in order to enrich or perform exploratory BI.

Every i-SLOD component consists of a series of RDF-triples datasets regarding some of the perspectives we consider relevant for BI over sentiment data. As proposed in LOD, links between datasets are expressed with "owl:sameAs" statements. Links to external datasets like DBpedia play a very relevant role in this infrastructure since they can enormously facilitate the migration of existing review and opinion data. For example, reviews already containing microdata referring to some product in DBpedia will be automatically assigned to the product URI of the corresponding i-SLOD product dataset.

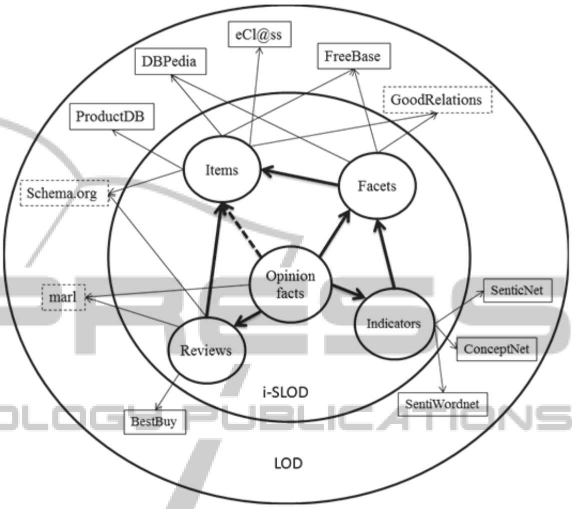

Figure 2: Main components of I-SLOD, and their relation to existing LOD vocabulary and data sets.

Regarding the nature of the data to be published in this infrastructure, we have identified some basic requirements in order to make published data useful in a real BI scenario:

- Support classification of sentiment data through taxonomical relationships.
- Support massive generation of opinion data from reviews texts.
- Support high distribution of data, providing optimal partitions w.r.t. to data usage.
- Provide fresh data by migrating as quickly as possible published reviews.
- Adapt as much as possible existing vocabularies in e-commerce in order to facilitate the load of data from different sources.
- Ensure quality and homogeneity of the i-SLOD datasets, dealing with the multi-lingual issues of this BI scenario.

## 3 i-SLOD DATASETS

In this section, we briefly describe the main datasets that will constitute the i-SLOD data infrastructure (inner ring of Figure 2). The main criteria we have followed to define these datasets are the following:

- Take profit from existing vocabularies and schemas as much as possible.
- Distribute linked data according to both the identified BI demands and the fact extraction from raw texts.
- Keep the inner datasets coherent.

The rest of the section shows the most relevant aspects of the datasets included in each component.

## 3.1 Items Component

This component contains the datasets describing concrete products and services as well as their manufacturers (e.g., product brand). These datasets must be kept as simple as possible just providing the attributes useful for BI tasks. Other attributes and relationships can be accessed through the links to externals datasets such as eCl@ss, DBpedia, ProductDB, FreeBase, etc. For the sake of simplicity, this component just regards two root classes: Item and Manufacturer. The schema for the former is shown in Table 1.

Table 1: Item i-SLOD schema.

| Property | Description |
|---|---|
| s:itemID | Unique identifier of the item. |
| gr:hasManufacturer | URI of the manufacturer. |
| rdf:label | Item name. |
| slod:onDomain | Item family. |
| rdf:type | Type of item (product, service). |

For this component we adopt the vocabularies of Schema.org (*s*) and GoodRelations (*gr*). It is worth mentioning that, although there are several datasets about products in the LOD cloud, they do not cover all products and services. In order to perform BI tasks, this is a serious limitation since an analytical query requires all data be expressed under the same schema. This limitation is present in MARL approach (Westerski and Iglesias, 2011), as opinion products are arbitrarily linked to either external datasets or literals. In our case, we propose a homogeneous schema, which can be further linked to external datasets.

## 3.2 Facets Component

This component comprises all the elements subject to evaluation in the opinions. In this work, the concept *feature* is used for denoting concrete physical parts of an item (e.g., *zoom*, *room*, etc.), whereas the concept *aspect* is used for abstract concepts (e.g., *design*, *price*, etc.).

Table 2: Facets i-SLOD Schema.

| Property | Description |
|---|---|
| slod:facetID | Unique identifier of a facet. |
| rdf:label | Facet labels. |
| slod:onDomain | Item family to which it is defined. |
| rdf:type | Facet type (feature, aspect, etc.). |

There are few LOD datasets including facets subject to opinions. We can find technical specifications about products like in eCl@ss, but they do not cover well the features customers usually opine (García-Moya et al., 2013b). As a consequence, sentiment analysis approaches aim at extracting these features directly from text reviews by applying machine learning techniques (Liu, 2012).

Indeed, one of the i-SLOD goals is to conceptualize and make public facets that can be automatically extracted from reviews. For this purpose, we propose a simple schema (see Table 2) to which item facets must map to. The main issues for performing these mappings are: to group together expressions denoting the same facet, to distinguish between features and aspects, and to classify features w.r.t. aspects. Our starting point for addressing these issues is the statistical approach presented at (García-Moya et al., 2013b).

## 3.3 Indicators Component

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

### 3.3.1 Opinion Words

Opinion words, also known as sentiment 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*, *cost an arm and a leg* or *cost the earth*, in this case all are referring to the indicator concept *expensive*.

Opinion 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 features it is modifying (e.g., *unexpected* for movies (+) and electronic devices (–)). Even within the same domain, the polarity of an indicator may be different depending on the feature. 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 feature. For example, the indicator “*too expensive*” refers to the aspect “*price*”.

For the Indicators component we propose two classes: slod:Indicator and slod:Polarity. Table 3 shows the main properties for the indicator class according to the previous comments.

Table 3: Properties for opinion indicators.

| Property | Description |
|---|---|
| slod:indicatorID | Unique identifier of a sentiment. |
| rdf:label | Sentiment words and sentences. |
| rdf:type | Type of indicator. |
| slod:onFacet | Associated facet (implicit/context). |
| slod:hasPolarity | Polarity associated to the indicator. |

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 only applicable to English-written reviews, and they do not take into account context-based indicators (Lu et al., 2011). It is worth mentioning that there exist also some web services like SentiStrength (Thelwall et al., 2010) that compute polarities from free-texts. This kind of services could be applied over this dataset to infer the values of slod:hasPolarity.

### 3.3.2 Opinion Shifters

Opinion indicators may not be sufficient to determine the true or contextual polarity of the feature. The valence of a polar term may be modified by one or more words, called *contextual valence shifters*. These shifters can be categorized into several types, some of them are: negations (*not*, *never*, *none*, etc.), intensifiers (*deeply*, *very*, *little*, *rather*, etc.), modal shifters (*might*, *possibly*, etc.), and presuppositions (e.g., *lack*, *neglect*, *fail*, etc.) There are other kinds of shifters (Polanyi and Zaenen, 2006), but they are less useful for BI analysis. Table 4 shows the main properties of the shifter class.

Table 4: Properties for opinion shifters.

| Property | Description |
|---|---|
| slod:shifterID | Unique identifier for the shifter. |
| slod:change | Change applied to the indicator. |
| rdf:label | Expresions associated to the shifter. |
| rdf:type | Type of shifter. |

## 3.4 Reviews Component

Currently, we can find many proposals for representing review metadata in LOD. One of the main references is Schema.org, which has been adopted by Google for rich snippets over reviews. This vocabulary covers all aspects we need for the Reviews component, and therefore we have adopted it without extensions. Table 5 shows some properties associated to the review class.

Table 5: Properties for review objects.

| Property | Description |
|---|---|
| s:reviewrating | Overall assessment (s:rating). |
| s:itemreviewed | Item reviewed. |
| s:reviewer | Author of the review. |
| s:dtreviewed | Publication date of the review. |

## 3.5 Opinion Facts Component

Opinion facts express the associations between features/aspects to opinion indicators that appear at the review texts.

Table 6: Opinion facts properties.

| Property | Description |
|---|---|
| slod:opinionId | Unique identifier of an opinion fact. |
| slod:onFacet | Opined facet. |
| slod:fromReview | Review reference. |
| slod:onTargetItem | In comparisons, the compared item. |
| slod:compOperator | In comparisons, the operator being applied (e.g., *better*, *worst*, *faster*, etc.) |

In our approach, an opinion fact is always linked to the review object from which it was identified. Consequently, each opinion fact takes the time and place dimensions from its linked review. Thus, the schema of an opinion fact can be just expressed with the feature/aspect and indicator/shifters involved in the fact. Table 6 summarizes the properties associated to the opinion fact class.

Another kind of opinion facts regarded in (Liu, 2012) is that of product comparisons. To represent comparisons, two properties to the opinion fact class are added: slod:onTargetItem and slod:comOperator. 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*”).

The most similar approach for expressing opinions in LOD is that of MARL (Westerski and Iglesias, 2011). The main differences of our approach w.r.t MARL are the following ones. In our approach, opinion facts must be always linked to datasets within i-SLOD. In this way, we can ensure coherence and homogeneity of data for BI analysis. Moreover, our proposal uncouples the opinion fact from its polarity, which should be inferred from indicators and shifters. Finally, we do not allow opinion aggregations, as they will be performed by the analytical tools (see Section 4.3).

# 4 i-SLOD POPULATION

This section discusses how to populate the main components of i-SLOD data infrastructure.

## 4.1 ELT Processes

Similarly to traditional data warehouses (DW), we propose to populate the i-SLOD infrastructure by means of Extraction, Load and Transform (ETL) processes. These processes will be in charge of continuously processing published reviews to update i-SLOD datasets. In this context, each component presents a different dynamicity degree. For example, review and opinion facts will grow very quickly, whereas products, features and indicators will change more slowly.

Table 7: Proposed i-SLOD ETLs.

| Component | Operators | Dynamicity |
|---|---|---|
| Product/Service | LOD Linking | Low |
| Feature/Aspects | Sentiment analysis<br>LOD Linking | Low |
| Opinion indicators | Lexica extraction<br>Sentiment analysis | Low |
| Review | Microdata xtraction<br>LOD Linking | High |
| Opinion fact | Semantic Annotation | High |

Unlike traditional ETLing, i-SLOD processes deal with RDF and web data. Table 7 shows the main ETL operators involved in the i-SLOD components. As it can be noticed, one critical operator consists of linking all the loaded data to internal and external datasets (see Figure 2). Another critical operator consists of applying sentiment analysis to extract and rank relevant feature/aspects and indicators to be included in the corresponding datasets.

## 4.2 Semantic Annotation

We propose to apply automatic semantic annotation for extracting opinion facts from raw texts, and linking data. Semantic annotation consists in identifying concept mentions in the free-texts in order to link them to existing knowledge resources. This technique is gaining popularity within the LOD community as it allows linking unstructured data to reference knowledge resources (Mendes et al., 2011). Unfortunately, current tools are all targeted to Wikipedia.

In our context, semantic annotation should be performed with any lexicon that can be extracted from the i-SLOD datasets (rdf:label statements). Particularly, we are interested on identifying features, indicators and shifters in the review text to extract opinion facts. An example of opinion fact extraction is shown in Table 8.

Table 8: Example of opinion facts.

| review1: “I don’t like the image and sound of this camera” | |
|---|---|
| (slod:oatom1, slod:fromReview, slod:review1) | |
| (slod:oatom1, slod:onFacet, slod:feature123) | image |
| (slod:oatom1, slod:withIndicator, slod:indct2) | like |
| (slod:oatom1, slod:hasShifter, slod:shifter10) | don’t |
| (slod:oatom2, slod:fromReview, review1) | |
| (slod:oatom2, slod:onFacet, feature231) | sound |
| (slod:oatom2, slod:withIndicator, slod:indct2) | like |
| (slod:oatom2, slod:hasShifter, slod:shifter10) | don’t |

The work in (García-Moya et al., 2013a) will serve us as basis to define the tailored semantic annotators necessary to extract opinion facts.

## 4.3 BI Analysis in i-SLOD

The i-SLOD infrastructure is meant to hold large datasets of semi-structured data. Linked data is used as an integrating tool and provides a new architectural pattern for mapping and interconnecting data from a variety of sources. Such infrastructure should provide the analyst with the means for executing analytic queries.

BI tools provide a summarized view by aggregating the data over numerical measures according to contexts (i.e., dimensions). However, traditional BI is not suitable for linked data.

Complex queries over the i-SLOD infrastructure require a data processing model for a cloud architecture that integrates advanced information extraction and advanced analysis operations (i.e., OLAP operators). Fur such purpose, the datasets in the inner ring of i-SLOD can be partitioned and distributed according to the BI demands. For example, datasets can be partitioned with respect to domains and time slices. Moreover, functional map-reduce implementations (Dean and Ghemawat, 2004) can process such distributed partitions and parallelize complex analysis operators such as filter, join and aggregate (Sridhar et al., 2009).

In order to speed-up costly operations within the inner i-SLOD datasets, additional indexing mechanisms can be applied. For example, instead of performing a join between the opinion atom and the indicator datasets every time a user asks a query involving such datasets, we can build an index that associates each opinion atom with its indicators. More challenging is however, to efficiently perform BI operations involving external datasets, as we do not have control over the external sources.

On the other hand, the semantics introduced by the linked data flavour of the i-SLOD also require new scalable, distributed reasoning techniques able to efficiently compute new inferences so that they can be used in the analysis process.

## 5 CONCLUSIONS

We have presented i-SLOD, a proposal for a data infrastructure of open linked sentiment data. Its purpose is to facilitate the massive analysis of sentiment data by exploiting the ever-increasing amount of publicly available open linked data.

The i-SLOD components are designed to describe all necessary information for opinion analysis (products/services, features/aspects, and opinion indicators, reviews and facts), and also to incorporate the functionality required to perform massive opinion analysis: the extraction of opinion facts from text reviews, and the linkage of opinion data to other datasets, using semantic annotation as a key enabling technology.

This allows the exploitation of opinion-related dimensions of analysis that are out of reach for traditional BI applications, thus allowing the incorporation of crucial strategic information.

## ACKNOWLEDGEMENTS

This work has been partially funded by the “Ministerio de Economía y Competitividad” with contract number TIN2011-24147.

## REFERENCES

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

Dean, J., Ghemawat, S. (2004). Mapreduce: Simplified data processing on large clusters. OSDI ’04, pages 137–150.

Esuli, A., Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proc. LREC’06*, 417-422.

García-Moya, L., Kudama, S., Aramburu, M.J., Berlanga, R. (2013a). Storing and analysing voice of the market data in the corporate data warehouse. *Information Systems Frontiers*, 1-19, DOI: 10.1007/s10796-012-9400-y.

García-Moya., Anaya-Sánchez, H., Berlanga, R. (2013b). A Language Model Approach for Retrieving Product Features and Opinions from Customer Reviews, *IEEE Intelligent Systems*, DOI: 10.1109/MIS.2013.37.

Heath, T., Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, San Rafael, CA, 1st Edition.

Liu, B., (2012). *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers.

Lu, Y., Castellanos, M., Dayal, U., Zhai, C. X. (2011). Automatic construction of a context-aware sentiment lexicon: an optimization approach. *WWW 2011*, 347-356.

Mendes, P., Jakob, M., García-Silva, A., Bizer, C., 2011. DBpedia spotlight: shedding light on the web of documents. In *Proc. of I-Semantics* '11.

Polanyi. L, Zaenen, A. 2006. Contextual valence shifters. Computing Attitude and Affect in Text: Theory and Applications: The Information Retrieval Series Volume 20, 1-10, doi: 10.1007/1-4020-4102-0_1.

Sridhar, R., Ravindra, P., Anyanwu, K. (2009). RAPID: Enabling Scalable Ad-Hoc Analytics on the Semantic Web. In *Proc.s of ISWC* '09, 715-730.

Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., Kappas, A. (2010). Sentiment strength detection in short informal text. *JASIST*, 61(12), 2544–2558.

Westerski, A., Iglesias, C. A. (2011). Exploiting Structured Linked Data in Enterprise Knowledge Management Systems. An Idea Management Case Study. *In Proc. EDOCW*, 395-403.