Inf Technol Tourism
DOI 10.1007/s40558-015-0047-7

ORIGINAL RESEARCH

Automatic analysis of textual hotel reviews

Aitor García-Pablos[1] · Montse Cuadros[2] ·
Maria Teresa Linaza[1]

Received: 22 April 2015 / Accepted: 10 December 2015

Abstract Social Media and consumer-generated content continue to grow and impact the hospitality domain. Consumers write online reviews to indicate their level of satisfaction with a hotel and inform other consumers on the Internet of their hotel stay experience. A number of websites specialized in tourism and hospitality have flourished on the Web (e.g. Tripadvisor). The tremendous growth of these data-generating sources demands new tools to deal with them. To cope with big amounts of customer-generated reviews and comments, Natural Language Processing (NLP) tools have become necessary to automatically process and manage textual customer reviews (e.g. to perform Sentiment Analysis). This work describes OpeNER, a NLP platform applied to the hospitality domain to automatically process customer-generated textual content and obtain valuable information from it. The presented platform consists of a set of Open Source and free NLP tools to analyse text based on a modular architecture to ease its modification and extension. The training and evaluation has been performed using a set of manually annotated hotel reviews gathered from websites like Zoover and HolidayCheck.

Keywords Customer-generated reviews · Text analysis · Sentiment analysis

✉ Aitor García-Pablos
agarciap@vicomtech.org

Montse Cuadros
mcuadros@vicomtech.org

Maria Teresa Linaza
mtlinaza@vicomtech.org

[1] Department of eTourism and Cultural Heritage, Vicomtech-IK4, San Sebastián, Spain

[2] Department of Human Speech and Language Technologies, Vicomtech-IK4, San Sebastián, Spain

Published online: 22 December 2015

1 Introduction

Social Media and consumer-generated content, like hotel reviews, continue to grow and impact the hospitality domain (Browning et al. 2013). To reduce uncertainty and perceived risks, consumers often search for Word-of-Mouth (WOM) when making purchase decisions. Previous research has revealed extensive evidence showing the importance of WOM in purchase decision and choice behaviour. In the Internet era, the effect of WOM has been further enhanced in the form of electronic Word of Mouth (eWOM) (Litvin, Goldsmith and Pan 2008). Consumers can make their opinions easily accessible to other Internet users via message boards, Twitter, product review websites or online communities. Meanwhile, consumers are willing to search for the opinions and experiences of peer consumers before purchasing a product.

Consumers write online reviews to indicate their level of satisfaction with the hotel (Liu et al. 2013) and inform other consumers on the Internet of their hotel stay experience (Park and Allen 2013). Online reviews have become one of the most important information sources in consumers' accommodation decision making (Ye et al. 2011) and are used considerably to inform consumers about the quality of the services (Filieri and McLeay 2014). It cannot be ignored that consumers tend not to book a hotel without seeking online reviews (Kimet al. 2011).

A number of websites specialised in hospitality related offers have appeared on the web (e.g. Tripadvisor, Hotels.com, Expedia, Yelp.com, Citysearch, Orbitz, Booking.com, HolidayCheck). Many of them enable users to exchange information, ratings, opinions or recommendations concerning certain destinations, hotels, and other tourist services (O'Connor 2008; Ye et al. 2011; Liu and Park 2015). Besides the overall ratings, attribute ratings on hotel specific attributes such as service, location, price, room and cleanliness are available to customers on social media platforms, and are commonly taken into account when customers evaluate a hotel (Ramanathan and Ramanathan 2011; Zhang et al. 2013).

These online platforms provide excellent tools for tourists to document and relive their travel experience such as expressing their satisfaction level with the hotel stay experience (Filieri and McLeay 2014). Furthermore, as consistency in service quality is difficult to achieve, service failure is almost unavoidable from time to time. Online complainers can rapidly become the travel opinion leaders of the electronic age. Such dissatisfying critics negatively influence future attitudes towards hotels.

The tremendous growth of these data-generating sources demands new tools to deal with them and has inspired the development of new approaches to understand this phenomenon in a variety of disciplines. In order to cope with the big amount of customer-generated reviews and comments, Natural Language Processing (NLP) techniques are necessary (Cambria and White 2014). One of the most remarkable NLP sub-fields used to process customer-generated content is the so-called Sentiment Analysis. Sentiment Analysis techniques and tools allow computers to provide a valuable insight of what the customers perceive as positive or negative (Montejo-Ráez et al. 2014). In the hospitality field, there is an increasing interest in using customer reviews to gain insights about problems that have not been well understood by conventional methods. Indeed, Sentiment Analysis and other related NLP areas open the door to multiple opportunities to develop new knowledge to reshape the understanding in the field and to support decision making in customer relationship management in the hospitality industry.

This work presents OpeNER, a NLP platform applied to the hospitality domain in order to automatically process customer-generated text content and obtain valuable information from it. The introduced platform consists of a set of free Open Source NLP tools to analyse text based on a modular architecture to ease its modification and extension. The goal of these tools is to enable end-users (e.g. hoteliers, SMEs offering reputation analysis services and other actors in the tourism sector) to easily overcome some of the challenges of setting up NLP tools to deal with customer-generated text comments, so they can focus on building added value services upon them. The provided tools work for six languages so far: English, Spanish, French, Italian, German and Dutch. The tools achieve their interoperability using a particular result format which stores all the information, named Knowledge Annotation Format (KAF). This enables an easy extension or inclusion of new tools by anyone, for example to add new languages or functionalities, as long as KAF format is respected.

The remaining of this work is structured as follows. First, Sect. 2 deals with a brief common NLP techniques revision in the context of the tourism sector. Section 3 provides a technical description of the proposed NLP platform, accompanied by an example of how the different tools analyse textual content. Section 4 shows the evaluation of some of the implemented tools and utilities in the context of the hospitality reviews analysis. Finally, Sect. 5 contains the conclusions and future work.

2 State of the art

A large amount of information about companies and products can be gathered from the Web and organized and visualized through various text and Web mining techniques. Web intelligence, web analytics, and user-generated content collected through Web 2.0-based social and crowd-sourcing systems (Doan et al. 2011; O'Reilly 2005) have brought a new and exciting era of business intelligence research in the 2000s, centred on text and web analytics for unstructured Web contents.

Many Web 2.0 applications developed after 2004 have also created large amounts of user-generated content from various online social media such as forums, online groups, web blogs, social net-working sites, social multimedia sites (for photos and videos), and even virtual worlds and social games (O'Reilly 2005). In addition to capturing celebrity chatter, references to everyday events, and socio-political sentiments expressed in these media, Web 2.0 applications can efficiently gather a large volume of timely feedback and opinions from a diverse customer population for different types of businesses. In order to leverage the content of the customer feedback provided in text comments, specific tools and technologies are required; in particular Natural Language Processing tools.

Natural Language Processing (NLP) is a field of Computer Science that studies the use of automatic ways to process natural language. As it has been mentioned before, automatic processing of text is becoming increasingly important in the tourism sector due to the large amount of content generated by users every minute. NLP is a wide research field, with many subfields addressing specific information extraction tasks of varying complexity. Different domains and types of texts have different information extraction requirements and thus require different NLP tasks and tools (Kiyavitskaya et al. 2009). In these paragraphs we do not describe the full NLP state of the art, which would require a book to deal with its many areas and subfields. Instead we focus on some aspects relevant to analysing hotel reviews, and in particular the ones covered by the tools introduced in this paper.

2.1 Processing text

In order to process a text, it is first necessary to determine its language. There are currently many Open Source language identification tools that implement state-of-the-art algorithms, achieving a precision over 99 % for tens of languages. The most popular approaches are based on statistical distributions and probabilities of character level *n*-grams (Řehůřek and Kolkus 2009), which are sequences of *n* characters. It is proven that every language has its own particular distribution of such *n*-grams.

Once the language has been identified, tokenization is commonly the following step of any text processing pipeline (Webster and Kit 1992). It is the process of breaking a text into its fundamental pieces, called tokens, which are likely to be a word, a number, a punctuation mark, or a particular combination of them.

Part-of-Speech tagging (PoS-tagging) is the next step that assigns grammatical categories to words in a text. Basically, it states that a word in a particular context is a noun, a verb, an adjective, an adverb, etc. It can also provide more information, like the gender and number of a word, or the person in case of verbs. PoS-taggers are usually based on stochastic methods like Hidden Markov Models or Maximum Entropy, trained on sets of pre-annotated data (Brants 2000; Collins 2002). The accuracy achieved by state-of-the-art taggers varies from one language to another and relies heavily on available training datasets (Giesbrecht and Evert 2009). PoS-tagging is sometimes considered as a solved problem, because the performance of state-of-the-art systems for the languages that have received major attention in the literature (e.g. English) is above 95 %. The current trends for PoS-tagging systems are about creating or adapting systems or methods to new emergent languages (Sun et al. 2014), or creating PoS-taggers that deal with jargon and specific types of non-conventional texts like Twitter messages (Derczynski et al. 2013).

Another relevant analysis that can be performed on a text is the Named Entity Recognition and Classification (also known as NERC). NERC locates and classifies rigid entity designators appearing in texts such as proper names (Nadeau and Sekine 2007). The concept of "entity" that a particular system tries to detect depends on what the system is intended for and the requirements of its target domain. In the tourism sector, the main entities are names of people, organizations (hotels, restaurants) and location names (countries, cities, or any other kind of geographical location). In other contexts, also dates, numeric expressions or currencies are detected. Common approaches described in the literature use supervised machine learning classifiers or sequence labelling. As with the PoS-tagging, for regular texts (e.g. well-written news articles) and for major languages, NERC is considered an almost solved task, with existing systems obtaining about 90 % of precision when tested on well-known test sets. Notwithstanding this fact, classical approaches and systems do not perform so well when they are applied to non-general domains (e.g. medicine, computer games or restaurants), in new languages, or in different writing styles like online reviews or social network comments (Marrero et al. 2012).

As another step in text processing, entities detected with a NERC system can be disambiguated in order to distinguish the entities referred from a set of potential candidates using Named Entity Disambiguation and Linking techniques. When possible, detected named entities are linked to well-known ontologies or knowledge-bases (Sil et al. 2012) like the Wikipedia's page of that entity. This allows uniquely identifying that entity according to a certain namespace or vocabulary (Rao et al. 2013), and aggregating or manipulating more precisely all the mentions to the same entity in order to avoid confusion with other entities with similar names, e.g. Washington as a city or as a state.

On the other hand, two different mentions in a text may refer to the same real-world entity. For example, in the following comment, "I stayed in NH in Brussels and Zurich and I really liked *them* because of *their* modern and stylish design and big rooms", the word *them* refers to "NH in Brussels and Zurich", and so does the word *their*. Detecting which mentions co-refer to the same entity is known as co-reference resolution (Bagga and Baldwin 1999). To solve co-referent expressions, both linguistic and domain knowledge are required. One of the best performing systems is a multi-pass sieve co-reference resolution system (Lee et al. 2011a, b), which combines different analysis sieves for entity mention detection and co-reference resolution in an incremental way.

Finally, Sentiment Analysis and Opinion Mining are closely related fields which refer to the application of NLP techniques to extract subjective information about how someone expresses a feeling (negative, positive or neutral) about something (Pang and Lee 2008). These tasks are increasingly important for determining the opinion about products and services, and brand reputation on the Internet. Usually, this information is the sentiment of the so-called "opinion holder" towards a particular "opinion target" (a topic, an entity or some part or feature of it) (Liu 2010). Ideally, this task is about retrieving "who" is opining "what" about "which entity" in each given piece of text. The time can be also important, especially when the opinions and sentiments change very quickly.

There are plenty of different approaches to perform Sentiment Analysis and Opinion Mining. Not all the available systems and techniques aim to extract the same type of information or with the same granularity. Some are oriented to just finding the overall polarity of a full sentence, paragraph or document, while others aim at finding the polarity of a product or service feature basis (e.g. distinguishing whether a particular opinion is about the rooms of a hotel or about the breakfast).

Furthermore, most of them involve some kind of machine learning techniques combined with specific language resources (Cambria et al. 2013) in order to classify and group existing content, extract or infer information or predict trends. Usually, those tools are language and domain dependent. This means that most of the techniques and tools work better for a target language and domain and thus require some adaptions to work for other languages or domains.

Tourism sector related domains, like hospitality and restaurants, are remarkable examples of domains for which NLP tools and approaches are actively developed. All the described techniques are used or might be used to obtain valuable insights on tourism sector related text content generated by online customers.

2.2 Application to the tourism sector

The increasing growth and popularity of user-generated contents on the Web, such as customer comments on social networks and specialized websites, has led to a new area of research in the application of text mining techniques. Applications of Sentiment Analysis and Opinion Mining based on text reviews have grown very quickly during the last decade in the tourism sector.

The earliest approaches focused on Sentiment Analysis of product reviews, which were clustered as positive or negative on the basis of specific sentiment structures (Hu and Liu 2004; Lau et al. 2005; Popescu and Etzioni 2005). Four steps were defined for online text mining: definition of mining context and concepts; data collection; dictionary construction; and data analysis. Several analyses have been done related to the profile of a hotel or the price of a room.

More recently, sentiment classification of consumer reviews is addressing bigger challenges, since the Opinion Mining systems try to deal with more complex tasks and results, as customers may provide a mixed review, combining positive and negative aspects of the same product or service. Ghose et al. (2009) used a 4-grams Dynamic Language Model classifier to acquire a subjectivity confidence score for each sentence in a hotel review and derive the mean and standard deviation of this score. The analysis of the content focused on polarity classification, sentiment classification of customer reviews, or automated extraction of product attributes. They have further used text-mining techniques to include textual information from hotel reviews in demand estimation models on the basis of the user-generated hotel reviews from Travelocity and TripAdvisor.

Ye et al. (2009) presented a study to analyse the existing approaches to perform automatic classifications based on Sentiment Analysis of online reviews related to travel destinations. Furthermore, the study analyses different supervised machine learning algorithms, in particular Naïve Bayes, Support Vector Machines and character based n-gram models for sentiment classification. The authors evaluate for each approach the effect on the different amount of training corpus to various performance measurements in terms of accuracy, precision, and recall in the sentiment classification of online reviews about tourist destinations. The algorithms evaluate the reviews about seven popular travel destinations in Europe and North America.

Moreover, Lee et al. (2011a, b) used text mining techniques to extract keywords from descriptive comments from hotel customers in order to identify areas of service failures and recovery actions. CATPAC software was used to classify and identify main topics based on the frequency of key terms. Furthermore, Kasper and Vela (2011) have implemented a service for hotel managers that collects customer reviews from various sites on the web; analyzes and classifies the textual content of the review; and presents the results in a systematic way. The customer reviews sentiment polarity classification is achieved through a supervised statistical classifier based on character n-grams and trained using a corpus of positive and negative reviews. Its main disadvantage is that it is available only in German.

Gräbner et al. (2012) have proposed a system that classifies customer reviews of hotels on the basis of lexicon based Sentiment Analysis techniques. The study includes building a lexicon with a semantic orientation of the relevant words of the given corpus; the application of Sentiment Analysis based on the generated lexicon to generate a classification of customer reviews; and the evaluation of the results with quantitative ratings.

Finally, Xiang et al. (2015) have published a study about using Big Data and text analytics to assess hotel guest experience and satisfaction. According to their conclusions the analysis of customer-generated content in form of text reviews can provide new and relevant insights to understand customers' preferences, likes and dislikes. In this scenario text analytics, and thus text analysis and natural language processing tools play an important role to help dealing with large amounts of customer-generated comments.

3 A framework for text analysis

The implemented framework provides a set of Open Source and ready-to-use tools to perform NLP analysis in six languages (English, Spanish, Italian, Dutch, German and French). This framework has been designed and developed to be public and freely available[1] and enables different agents from the tourism sector to automatically extract textual feedback on the basis of NLP technologies focusing on Sentiment Analysis and Opinion Mining. Several text processing modules have been implemented including functionalities related to language detection; sentence splitting and tokenisation; Part-of-Speech tagging; Named Entity Detection and Classification; Named Entity Linking; co-reference resolution; and Sentiment Analysis and opinion detection (Agerri et al. 2013). The platform also provides some tools to perform domain adaptation of the existing resources, for example to adapt sentiment lexicons to a new domain, to train new models for opinion detection, etc. Some of the provided tools are based on already available third-party tools, like Apache OpenNLP library[2] or DBpedia Spotlight[3] that have been adapted

[1] The complete information can be found at http://www.opener-project.eu.

[2] https://opennlp.apache.org/index.html.

[3] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki.

and conveniently wrapped to achieve the scalability and modularity desired for the modules of the platform.

3.1 General architecture of the platform

The platform is designed and implemented on an individual module basis. Each module receives a single input; performs a single text processing task; and returns a single output. Both the input and the output are documents in KAF format (see Sect. 3.2) except for the language identifier and the tokeniser, which are the first modules of the analysis process and receive plain text as input (i.e. the text that is going to be analysed). This allows a very easy integration among modules to build a full analysis pipeline. Implemented NLP modules use Java, Ruby or Python (depending on the requirements and the pre-existing resources for each NLP task). There are no integration problems among modules implemented in different programming languages as long as each module processes KAF correctly.

Figure 1 shows a possible way of composing proposed NLP modules to perform different types of analysis. The depicted architecture shows that the interaction among modules is mostly based on the sharing of KAF documents. Thus, it is possible to fully replace or customise any module, just ensuring that KAF is properly read and written. It is also possible to add brand new modules, improving

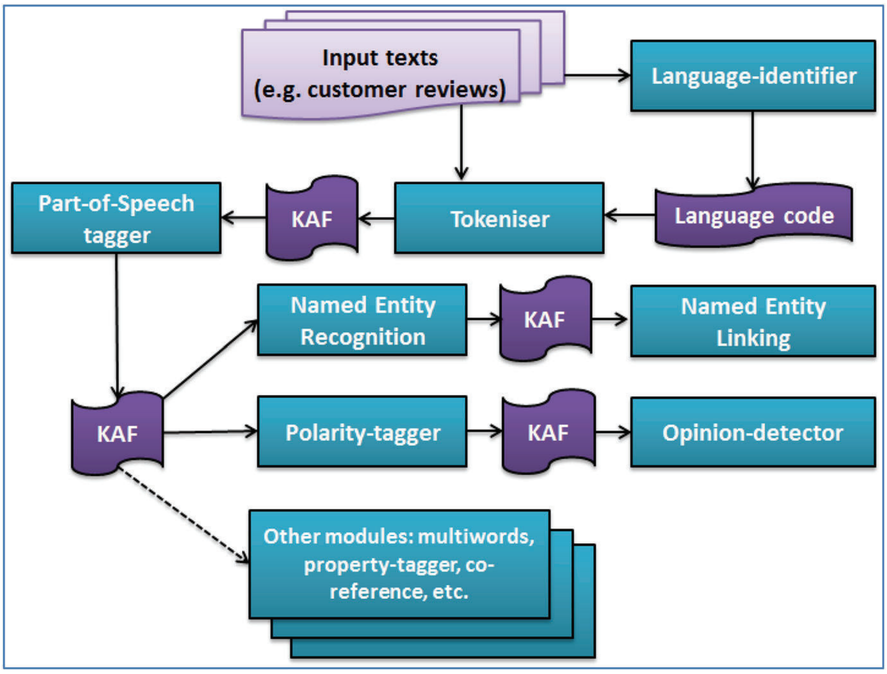


Fig. 1 Some possible text analysis pipelines

or extending the functionalities of the platform just by making sure that the input and output KAF documents are appropriately processed.

3.2 KAF: a layered format to represent the analysis results

One of the main features of the platform is the modularity of each component or the piece of software in charge of a particular NLP task. This has been achieved using a single yet expressive data representation format called KAF (Knowledge Annotation Format[4]) (Bosma, Vossen and Soroa 2009), which is the only connection among modules. Each module of the platform is responsible of providing valid KAF as output, which is then used as input for another module. The lack of tight coupling among modules provides a seamless extensibility when a new module is added or a module changes.

KAF documents are structured in several layers, each of them corresponding to different text processing tasks. Each layer is independent from the others, except for referencing an element of a previous layer. Figure 2 displays an example of a KAF document corresponding to the tokenisation, Part-of-Speech tagging and lemmatisation of the sentence "This is a sample text". This example does not include any Named Entity or opinionated words and hence the layers corresponding to these analysis processes do not appear in the example. Although KAF documents become verbose and difficult to read when the input text is long, the platform provides parsers to help handling KAF documents easily, reading and printing the information contained inside them. More examples to illustrate different KAF layers are given later in this section.

3.3 Workflow of the platform

The following text from a hotel review will be taken as an example to explain the workflow of the platform.

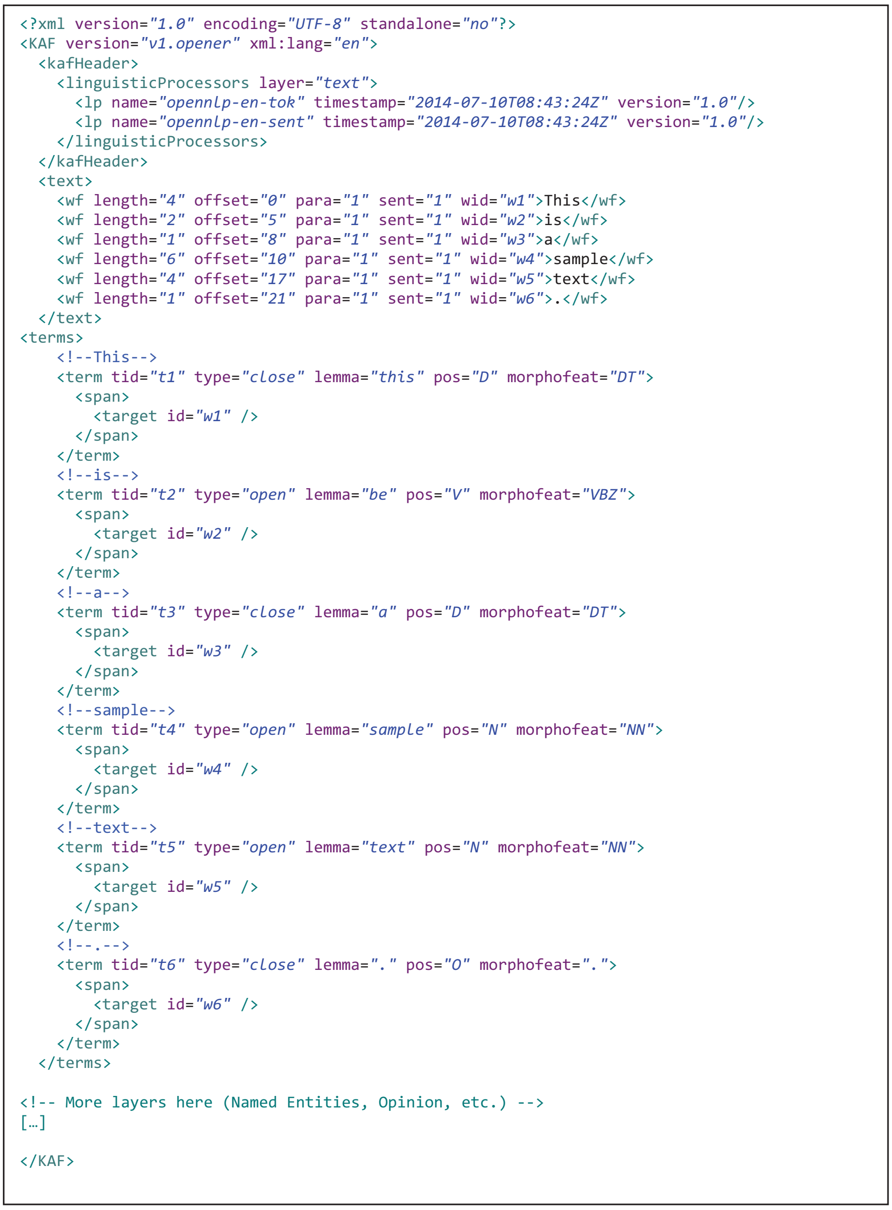
> I have been at Albergo Acquarello hotel at Lugano and I liked the beautiful decoration. The rooms were very comfortable. On the other hand, the restaurant was really expensive.

First, the text to be analysed is sent to the language identifier which returns the language code corresponding to the language detected in the text. The language identifier module internally uses an Open Source language detection library,[5] which includes trained language models to identify 47 different languages. In this case, the language code would be "en" for English.

Secondly, the tokeniser module receives the raw hotel review and the language code, and performs the tokenisation of the words outputting the result as a KAF document. This means that the tokeniser breaks the review text into individual

[4] Formerly KAF acronym stood for *Kyoto Annotation Format*, due to the name of the project in which a first version of KAF was designed. Since then KAF has evolved and the K letter changed its meaning to "Knowledge".

[5] https://github.com/shuyo/language-detection.



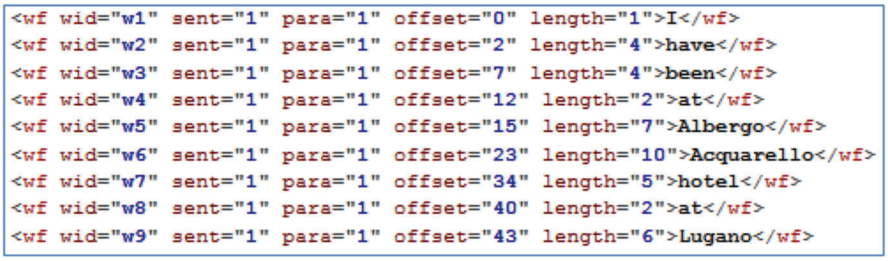
```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<KAF version="v1.opener" xml:lang="en">
  <kafHeader>
    <linguisticProcessors layer="text">
      <lp name="opennlp-en-tok" timestamp="2014-07-10T08:43:24Z" version="1.0"/>
      <lp name="opennlp-en-sent" timestamp="2014-07-10T08:43:24Z" version="1.0"/>
    </linguisticProcessors>
  </kafHeader>
  <text>
    <wf length="4" offset="0" para="1" sent="1" wid="w1">This</wf>
    <wf length="2" offset="5" para="1" sent="1" wid="w2">is</wf>
    <wf length="1" offset="8" para="1" sent="1" wid="w3">a</wf>
    <wf length="6" offset="10" para="1" sent="1" wid="w4">sample</wf>
    <wf length="4" offset="17" para="1" sent="1" wid="w5">text</wf>
    <wf length="1" offset="21" para="1" sent="1" wid="w6">.</wf>
  </text>
<terms>
    <!--This-->
    <term tid="t1" type="close" lemma="this" pos="D" morphofeat="DT">
      <span>
        <target id="w1" />
      </span>
    </term>
    <!--is-->
    <term tid="t2" type="open" lemma="be" pos="V" morphofeat="VBZ">
      <span>
        <target id="w2" />
      </span>
    </term>
    <!--a-->
    <term tid="t3" type="close" lemma="a" pos="D" morphofeat="DT">
      <span>
        <target id="w3" />
      </span>
    </term>
    <!--sample-->
    <term tid="t4" type="open" lemma="sample" pos="N" morphofeat="NN">
      <span>
        <target id="w4" />
      </span>
    </term>
    <!--text-->
    <term tid="t5" type="open" lemma="text" pos="N" morphofeat="NN">
      <span>
        <target id="w5" />
      </span>
    </term>
    <!--.-->
    <term tid="t6" type="close" lemma="." pos="O" morphofeat=".">
      <span>
        <target id="w6" />
      </span>
    </term>
  </terms>

<!-- More layers here (Named Entities, Opinion, etc.) -->
[…]

</KAF>
```

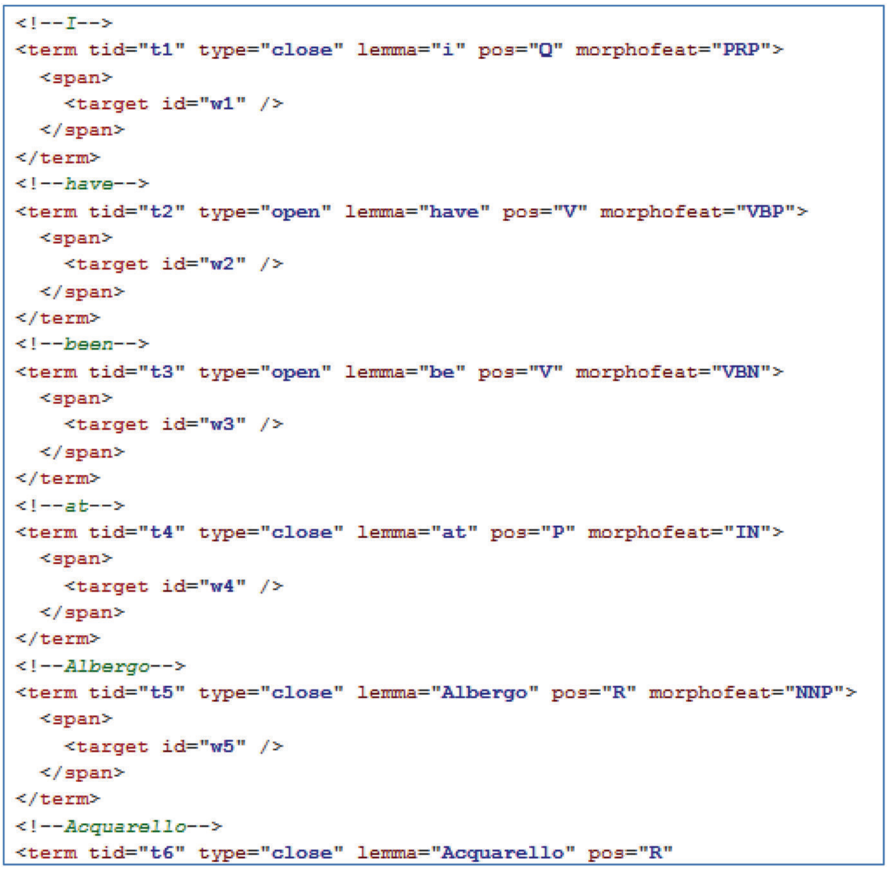
Fig. 2 Example of a KAF document to represent a simple analysed sentence

sentences and tokens (i.e. separating words and punctuation marks) and returns a KAF document. An example of tokenisation represented in KAF is shown in Fig. 3. The attribute "wid" is a unique token identifier assigned for later reference. The attribute "sent" and "para" indicate the number of sentence and paragraph in the context of the analysed text. The attributes “offset” and “length” indicate the beginning of the current token (measured by the number of characters from the start of the analysed text) and the number of characters of the current token respectively.



```
<wf wid="w1" sent="1" para="1" offset="0" length="1">I</wf>
<wf wid="w2" sent="1" para="1" offset="2" length="4">have</wf>
<wf wid="w3" sent="1" para="1" offset="7" length="4">been</wf>
<wf wid="w4" sent="1" para="1" offset="12" length="2">at</wf>
<wf wid="w5" sent="1" para="1" offset="15" length="7">Albergo</wf>
<wf wid="w6" sent="1" para="1" offset="23" length="10">Acquarello</wf>
<wf wid="w7" sent="1" para="1" offset="34" length="5">hotel</wf>
<wf wid="w8" sent="1" para="1" offset="40" length="2">at</wf>
<wf wid="w9" sent="1" para="1" offset="43" length="6">Lugano</wf>
```

Fig. 3 Resulting tokens for a fragment of the sentence, represented in KAF



```
<!--I-->
<term tid="t1" type="close" lemma="i" pos="Q" morphofeat="PRP">
  <span>
    <target id="w1" />
  </span>
</term>
<!--have-->
<term tid="t2" type="open" lemma="have" pos="V" morphofeat="VBP">
  <span>
    <target id="w2" />
  </span>
</term>
<!--been-->
<term tid="t3" type="open" lemma="be" pos="V" morphofeat="VBN">
  <span>
    <target id="w3" />
  </span>
</term>
<!--at-->
<term tid="t4" type="close" lemma="at" pos="P" morphofeat="IN">
  <span>
    <target id="w4" />
  </span>
</term>
<!--Albergo-->
<term tid="t5" type="close" lemma="Albergo" pos="R" morphofeat="NNP">
  <span>
    <target id="w5" />
  </span>
</term>
<!--Acquarello-->
<term tid="t6" type="close" lemma="Acquarello" pos="R"
```

Fig. 4 Fragment of PoS-tagging information of the analysed sentence represented in KAF

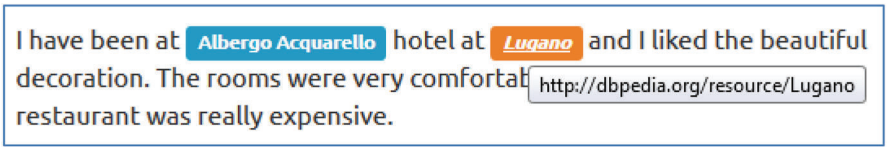
Such document is the input for the Part-of-Speech tagger module, which annotates each word as being a noun, a verb, an adjective, etc. and lemmatises them. An illustrative representation of the result can be found at Fig. 4. The attribute "tid" is a unique identifier for later reference. The attribute "lemma" is the lemma of the corresponding word. The attribute "pos" is the Part-of-Speech of the word according to KAF notation (e.g. 'Q' for pronouns, 'V' for verbs, 'N' for nouns, 'P' for prepositions, 'R' for proper nouns, etc.). The attribute "morphofeat" is again the Part-of-Speech of the word, but using an arbitrary notation that depends on the implementation of the module (i.e. different languages and tools use different tagsets to represent the Part-of-Speech and morphological information). The example shows that the English Pos-taggers outputs the morphofeat attribute using the *Penn TreeBank* tagset[6] commonly used for English. The reason to include both "pos" and "morphofeat" attributes is that the first one ("pos") is forced to fit the KAF notation maximizing the compatibility among modules, while the second one ("morphofeat") allows keeping the original Part-of-Speech tag that may contain additional information useful for other purposes.

The output KAF is sent to the Named Entity Recognition module to detect Named Entities. As shown in Fig. 5, the analysis detects two entities in the text: *Albergo Acquarello* and *Lugano*. The former has been classified as an "organisation" (the *Albergo Acquarello hotel*), while the latter has been defined as a geo-spatial location (Lugano, Switzerland). Once the result is sent to the Named Entity Linking module, the mention to *Lugano* has been linked to its entry in DBpedia. This allows determining which "*Lugano*" entity the text talks about (in case there is more than one possible "*Lugano*" in the world) and obtaining additional metadata about the entity if available (e.g. geo-coordinates, population, country, etc.). In Fig. 6 the corresponding KAF representation (only of the named entities layer) is shown.

If the Polarity-tagger module is invoked, the analysis of the sentiment and opinion-related information are obtained. As illustrated in Fig. 7, the module assigns a polarity (positive, negative) to the words in the text according to a sentiment lexicon, which is a dictionary that states the most probable polarity for a word inside the given sector. The detected positive and negative words have been highlighted with different colours, as well as the intensifiers (i.e. the words that intensify the polarity of the surrounding words). In Fig. 8 a fragment of the KAF representation for the polarity annotation is shown, including the polarity information for the words "really" and "expensive", which are an intensifier and a negative word respectively.

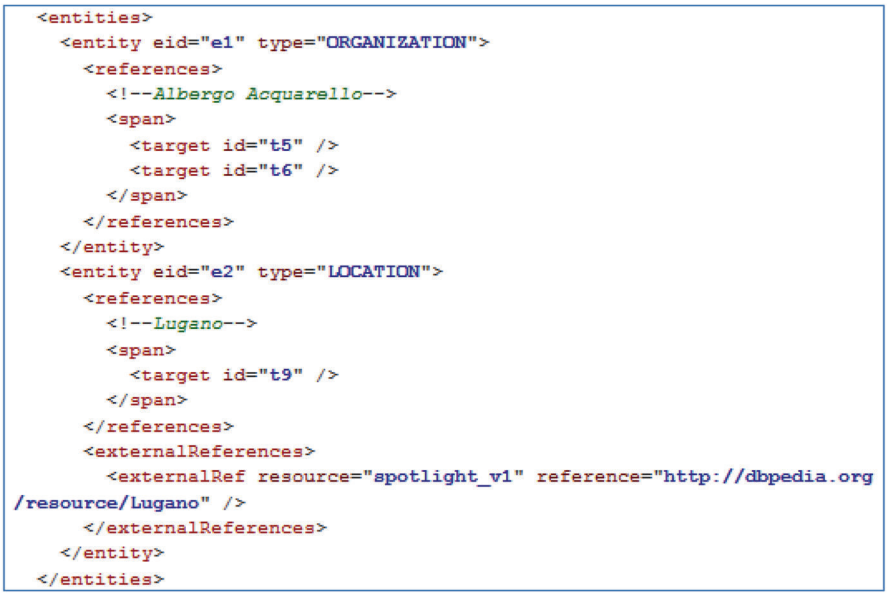
The polarity information is a first step to get an insight about the sentiment of the review. The Opinion detector module further detects complete expressions which include several words; classifies them as being positive or negative taking into account the overall expression; and finds the target of that expression, such as the particular object or feature which the opinion is about. The module uses machine learning techniques to classify which parts of a sentence are opinion expressions.

[6] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.



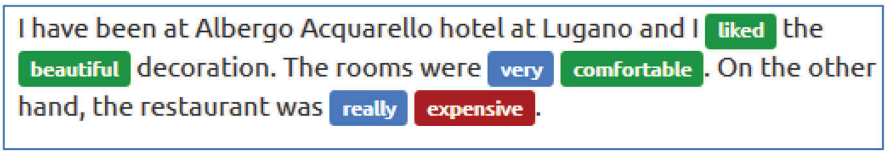
I have been at Albergo Acquarello hotel at Lugano and I liked the beautiful decoration. The rooms were very comfortab http://dbpedia.org/resource/Lugano restaurant was really expensive.

Fig. 5 Representation of the Named Entity Recognition result: Albergo Acquarello as an "organisation", and Lugano as a "location", highlighted in different colours



```
  <entities>
    <entity eid="e1" type="ORGANIZATION">
      <references>
        <!--Albergo Acquarello-->
        <span>
          <target id="t5" />
          <target id="t6" />
        </span>
      </references>
    </entity>
    <entity eid="e2" type="LOCATION">
      <references>
        <!--Lugano-->
        <span>
          <target id="t9" />
        </span>
      </references>
      <externalReferences>
        <externalRef resource="spotlight_v1" reference="http://dbpedia.org
/resource/Lugano" />
      </externalReferences>
    </entity>
  </entities>
```

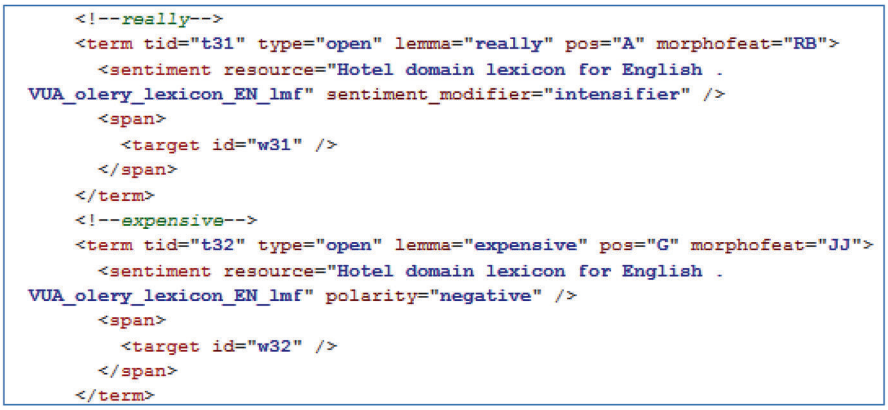
Fig. 6 Named Entity Recognition result represented in KAF



I have been at Albergo Acquarello hotel at Lugano and I liked the beautiful decoration. The rooms were very comfortable. On the other hand, the restaurant was really expensive.

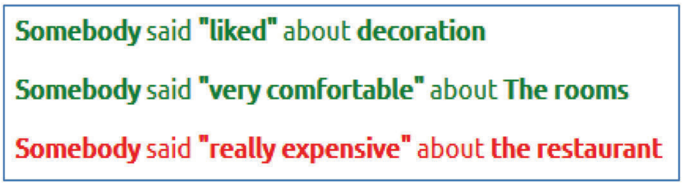
Fig. 7 Detected polarity of the words highlighted with *different colours*

For example, Fig. 9 shows a visual representation of the triplet of information the opinion detector identifies. The first part of each opinion is the opinion holder. In a standard hotel review, the opinion holder is the author of the review implicitly. Because there is no explicit opinion holder, it appears as "Somebody" in the example. The second part of the triplet is the opinion expression itself, a word or group of words that comprises an opinion or a particular sentiment towards something. An opinion expression can be positive, negative or neutral.



```
    <!--really-->
    <term tid="t31" type="open" lemma="really" pos="A" morphofeat="RB">
      <sentiment resource="Hotel domain lexicon for English .
VUA_olery_lexicon_EN_lmf" sentiment_modifier="intensifier" />
      <span>
        <target id="w31" />
      </span>
    </term>
    <!--expensive-->
    <term tid="t32" type="open" lemma="expensive" pos="G" morphofeat="JJ">
      <sentiment resource="Hotel domain lexicon for English .
VUA_olery_lexicon_EN_lmf" polarity="negative" />
      <span>
        <target id="w32" />
      </span>
    </term>
```

Fig. 8 The terms “really” and “expensive” represented in KAF



Somebody said "liked" about decoration

Somebody said "very comfortable" about The rooms

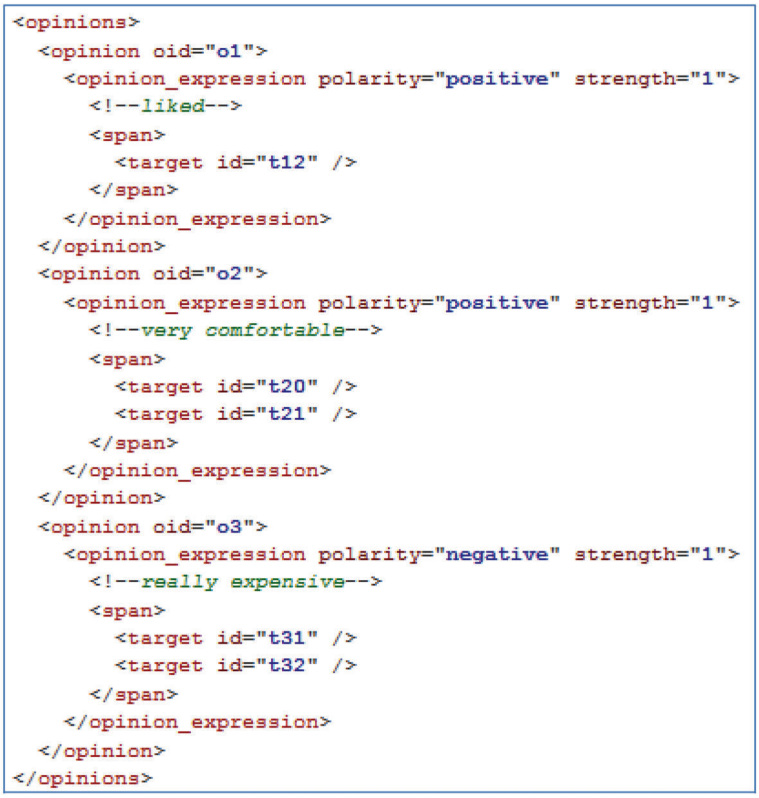
Somebody said "really expensive" about the restaurant

Fig. 9 An inline representation of the information obtained by the Opinion detector

Finally, the opinion target is the object or the feature being reviewed or assessed by the review. The opinion target (also called aspect term, feature term, etc.) is crucial to obtain a fine grained sentiment score and to aggregate the opinions on a per-feature basis to assess the strengths and weaknesses of a product or service. For example, while hotel rooms can be positively perceived, breakfast service is negatively evaluated. In Fig. 10 an example of KAF representation for the opinions layer is shown.

3.4 Further functionalities provided by the platform

Apart from the described analysis components, there are other pre-processing modules that provide a potentially valuable output to further analyse customer reviews. The main difference with the previously described modules is that these extra functionalities do not provide a KAF document or any other result that could be directly piped into another module, but provide outputs than can be leveraged in different ways, acting as a pre-processing step for further exploitation or for domain customisation, like the word-category pairs to be used by the property-tagger module (see Sect. 3.4.2).



```
<opinions>
  <opinion oid="o1">
    <opinion_expression polarity="positive" strength="1">
      <!--liked-->
      <span>
        <target id="t12" />
      </span>
    </opinion_expression>
  </opinion>
  <opinion oid="o2">
    <opinion_expression polarity="positive" strength="1">
      <!--very comfortable-->
      <span>
        <target id="t20" />
        <target id="t21" />
      </span>
    </opinion_expression>
  </opinion>
  <opinion oid="o3">
    <opinion_expression polarity="negative" strength="1">
      <!--really expensive-->
      <span>
        <target id="t31" />
        <target id="t32" />
      </span>
    </opinion_expression>
  </opinion>
</opinions>
```

Fig. 10 Opinion layer of the KAF document for the given example

3.4.1 Multiword term generation module

A multiword term can be defined as a term formed by more than one words, like idioms, expressions, locutions or usual word collocations. Some examples of multiword terms could be *hot dog*, *Italian food*, *hotel chain*, *train station*, *public transport* or *wireless connection*. Depending on the type and requirements of text processing task, it is useful to detect multiword terms to be analysed as a single word. For example, when detecting sentiment polarity of customer reviews, it is important to distinguish the positive word *happy* from the expression *happy hour*.

The platform provides a multiword generation utility, which only requires a set of unlabelled texts of the target domain (for example, customer reviews about hotels). No manual annotation or labelling is necessary, just the raw texts analysed with the PoS-tagger module to convert them into KAF files. The multiword generation module runs on these KAF files and generates a list of multiword terms, ranked by the likelihood of being a correct multiword.

In order to generate a multiword term list, the module generates sequences of *n* consecutive words occurring in the input texts (word *n*-grams) and computes the

Log-Likelihood Ratio (LLR) of the words co-occurring in the *n*-grams. LLR is a common measure in the literature to estimate if two events (two words in this case) co-occur by chance or if they are truly correlated (Dunning 1993). Then, the module ranks the n-grams by their LLR score and outputs a list of multiword terms with higher LLR score.

Additionally, to prevent obtaining a noisy list in which many candidate collocations are composed by stopwords (determiners, pronouns, and other undesired words), the module uses Part-of-Speech information of individual words that compose the candidate multiword terms to filter out the candidate combinations that do not follow certain patterns (e.g. noun + noun, adj + noun, noun + prep + noun, etc.). Table 1 shows the multiword terms that emerged automatically from a set of hotel reviews, which helps revealing relevant domain specific concepts that are expressed with more than one word. Such a list can be used for further processing, for example in the word categorisation module described in the following section.

3.4.2 Word categorisation module

The word categorisation module is a complement to the property-tagger module. The property-tagger module classifies words into certain properties or categories of the Named Entity or domain being reviewed, helping to summarise opinions by category, grouping sentences or reviews by topics, etc. For example, a sentence of a hotel review mentioning shower and towels probably can be classified under the broader category *bathroom*.

Table 1 Example of multiword terms obtained for hotels sorted by LLR score

Multiword terms from hotel reviews
Animation team
Walking distance
Train station
Aqua park
Metro station
Railway station
City center
Air conditioning
Public transport
Business trip
Public transportation
Swimming pool
Special thanks
Wireless internet
Shopping mall
Flat screen
Subway station

The property-tagger uses dictionaries of *word-category* pairs to assign the corresponding category to words appearing in texts (i.e. in a domain specific review). If a word in an analysed text appears in the *word-category* dictionary, the pairing category is assigned to it. This information outputted by the property-tagger module is represented and stored in the KAF document in a specific layer for later uses.

The *word-category* dictionary can be created in different ways. One possibility is to manually create the dictionary, assigning a category word by word in a supervised way. This approach should be the most precise one, as it only depends on the quality of the human annotation process and results. The main drawback is the difficulty of creating such a dictionary manually for a new domain or language. Moreover, manually crafting and maintaining such dictionary for each new possible domain and language is expensive and time consuming, which makes it unfeasible most of the times. In addition, if the categories inventory changes (i.e. the possible categories of interest) then the dictionary should be manually revised and updated.

To overcome such drawbacks, the platform provides an additional tool, a word-categorisation module which employs a semi-supervised approach to generate *word-category* dictionaries. The approach is based on the generation of a vector space model to represent the domain words, so that the implicit semantics of the words for the processed domain are captured. In this case, it is compulsory to have a large amount of customer reviews of the target domain. Raw texts are pre-processed with the tokenisation and Part-of-Speech modules and only nouns, verbs, adjectives and adverbs are kept. The resulting set of pre-processed documents is processed using the SemanticVectors[7] library (Widdows and Cohen 2010). Documents are indexed and the obtained index is used to create a vector representation of each word, which is condensed into a more compact representation (i.e. lower dimensional vectors) using a Random Projection algorithm (Sahlgren 2005).

Each desired domain category or domain topic, for example *room*, *location*, *staff* and *price* for hotels, must be defined in some way. In this case, a category is defined providing few representative seed words that, according to an expert judgement, fall inside that category in the domain under analysis and act as topic indicators. Table 2 shows some examples of categories for the hospitality domain and their corresponding seed words (only three seeds per category).

With the domain words represented as a vector space model and categories defined by some representative words, the method assigns the most likely category to a new word, comparing it against the category seed words. The comparison is based on the cosine distance of vector representation of each *word-to-classify* and the *category-words*. The most similar category based on this metric is assigned to the classified word. At the end, a dictionary of *word-category* pairs is generated using the input words and their assigned categories.

Table 3 shows some examples of automatically assigned categories for new words, using categories and seed words defined at Table 2. Although the table shows category assignments that seem intuitively correct, a further experiment has been conducted in the Sect. 4 to evaluate this semi-supervised approach.

[7] https://code.google.com/p/semanticvectors/.

Table 2 Example of categories defined for hospitality domain using few seed words

Category to define	Seed words
Room	Room, bed, pillow
Staff	Staff, service, worker
Restaurant	Restaurant, food, rice
location	Location, street, place
Value for money	Expensive, money, price

Table 3 Example of words automatically classified into domain categories

Unclassified words	Automatically assigned category
Bedroom	Room
Walking distance	Location
Dinner	Restaurant
Suite	Room
Wine	Restaurant
Airport	Location
Bill	Value for money
Budget	Value for money
Friendliness	Staff
Personnel	Staff

4 Evaluation in the accommodation domain

A set of hotel reviews has been manually annotated with sentiment and opinion related information to evaluate the described tools in the hospitality domain. The reviews were extracted from online customer review websites, mainly from Zoover[8] and HolidayCheck.[9] Several variables were taken into account in order to choose the reviews in six languages (English, Spanish, French, Italian, Dutch and German), such as the home country of the reviewer or the motivation for the stay at that hotel (work or leisure), or the 5-star rating given by the original authors in the source website, in order to obtain a balanced dataset. Such data is usually available as metadata annexed to the reviews. After discarding some reviews with no useful or incorrectly annotated content, nearly 200 reviews were selected for each language.

As a first step, two human annotators per language (native speakers or with a deep knowledge of the language they were annotating), tagged the reviews according to certain annotation guidelines with the help of a customised annotation tool. The selected 200 reviews for each of the six addressed languages were annotated. Per each review, the opinion expressions and when possible, the corresponding opinion holders and opinion targets were annotated. Further valuable information was manually tagged, like the polarity of the words or the general category of the opinion target (e.g. both "coffee" and "orange juice" belong to the

[8] http://www.zoover.com.

[9] http://www.holidaycheck.com/.

"breakfast" category, while "towel" and "shower" belong to the "bathroom" category).

80 % of the annotated hotel reviews were then used to train the models of the Opinion detector module of the platform, which is based on machine learning techniques like Conditional Random Fields (CRF) (Sutton and McCallum 2012) and Support Vector Machines (SVM) (Brereton and Lloyd 2010) that must be trained over a previously annotated dataset.

The remaining subsets containing a balanced number of positive and negative opinions were used to perform a formal evaluation of the resulting opinion detection models. The results of this evaluation are shown in Table 4. The first clear conclusion is that the results vary for each language. The different complexity of the languages; the sparse vocabulary due to the limited size of the training set; the annotation quality; or the accumulated errors stemming from the different per-language analysis pipelines (tokenisation, Part-of-Speech tagging, lemmatisation) could be potential explanations for this disparity among languages. Using other linguistic resources such as opinion lexicon with the polarity of the words could improve and better tune words with different sentiments in several domains.

Regarding the training and evaluation of the Named Entity Recognition and Classification (NERC) module, manually annotated hotel reviews did not include sufficient amount of Named Entities. In many of the cases, authors did not mention the name of the hotel or the location explicitly because it was implicit in the context

Table 4 Opinion detector evaluation results

Tool	Language	Precision (%)	Recall (%)	F-score (%)	Method	Dataset	Total opinion expressions[a]
Opinion detector	De	75.64	48.88	59.38	CRF + SVM	OpeNER manual hotel annotations	2103
Opinion detector	En	85.52	58.45	69.44	CRF + SVM	OpeNER manual hotel annotations	2075
Opinion detector	Es	74.41	46.55	57.27	CRF + SVM	OpeNER manual hotel annotations	2194
Opinion detector	Fr	70.94	46.28	56.02	CRF + SVM	OpeNER manual hotel annotations	1626
Opinion detector	It	65.47	40.39	49.96	CRF + SVM	OpeNER manual hotel annotations	1525
Opinion detector	Nl	82.8	51.77	63.71	CRF + SVM	OpeNER manual hotel annotations	2098

[a] This column refers to the total amount of opinion expressions manually annotated by human annotators in each dataset

of the review. Thus, the module has been trained and evaluated using general domain datasets that do not include specific vocabulary related to the accommodation domain. NERC models have been trained for the six languages targeted in the evaluation.

Table 5 shows the results of the evaluation for NERC modules of the platform in the six targeted languages. The results also are clearly different for each language. When available, datasets widely used in the field have been employed, like CoNLL 2002 and CoNLL 2003 for English, Spanish, German and Dutch, and Evalita 2007 dataset for Italian. For French, the proprietary ESTER corpus has been used.

Looking at the evaluation results, it can be concluded that the results for English and Spanish are within or close to state-of-the-art of NERC systems, obtaining also reasonable performance for other languages.

Regarding the evaluation of the word categorization module, it has been performed using a list of manually labelled words. During the manual annotation, human annotators tagged which category the hotel reviews belonged to from a predefined category inventory. The manually tagged word-category pairs have been used to assess the agreement between the automatic category assigned to each word by the tool and the human judgement.

No labelling or manual annotation is required to train the word categorization module, only texts from the target domain. About ten thousand raw hotel reviews were gathered again from online customer review websites, mainly from Zoover and HolidayCheck. Three seed words for each category were randomly picked from the list of word-category pairs obtained during the manual annotation process described above. Then, the remaining words in the test list were assigned a category automatically using the word categorization module. The automatically obtained category was then compared to the category assigned by human annotators.

Table 6 shows the confusion matrix after the automatic classification of the hospitality related words for English. The rows are the categories assigned by

Table 5 Named Entity Recognition evaluation results

Tool	Lang	Precision (%)	Recall (%)	F-Score (%)	Method	Dataset
Ner-base	De	84.02	58.56	69.02	Perceptron	CoNLL 2003
Ner-base	En	89.39	85.19	87.24	Perceptron + dictionaries	CoNLL 2003
Ner-base	Es	79.91	80.58	80.24	Maximum entropy	CoNLL 2002
Ner-base	Fr	86.15	75.69	80.58	Maximum entropy	ESTER corpus
Ner-base	It	81.15	62.70	70.74	Perceptron + dictionaries	Evalita 2007
Ner-base	Nl	79.85	75.41	77.57	Perceptron	CoNLL 2002

Table 6 Confusion matrix of automatic and manual (gold) categories assigned to words in English hotel reviews (rows are gold categories; columns are automatically assigned categories using the tool)

Gold\auto	Location	Room	Bathroom	Restaurant	Noise	Staff	Price	
Location	*42*	0	0	0	1	0	1	44
Room	1	*23*	3	0	1	1	0	29
Bathroom	1	11	*25*	0	0	0	0	37
Restaurant	0	0	0	*23*	1	6	8	38
Noise	5	4	0	1	*20*	4	1	35
Staff	0	0	0	1	0	*56*	0	57
Price	3	1	0	2	1	7	*22*	36
Total words	52	39	28	27	24	74	32	276
Accuracy[a]	0.80	0.58	0.89	0.85	0.83	0.75	0.68	

[a] Here accuracy measures the ratio of correct word-category assignments and total assignments

human annotators while the columns represent the categories assigned automatically by the module. The diagonal, in which the row label is equal to the column label, shows the correct category assignments in italics (i.e. when the predicted category for a word matches the category assigned by human annotators). As shown, the method is rather accurate despite its simplicity. Most of the categorisation errors appear in semantically related categories like *room* and *bathroom*, which contain words that are usually difficult to separate with automatic methods.

Finally, it should be mentioned that the platform can be further used to build added-value applications to enhance decision making process. Figure 11 displays

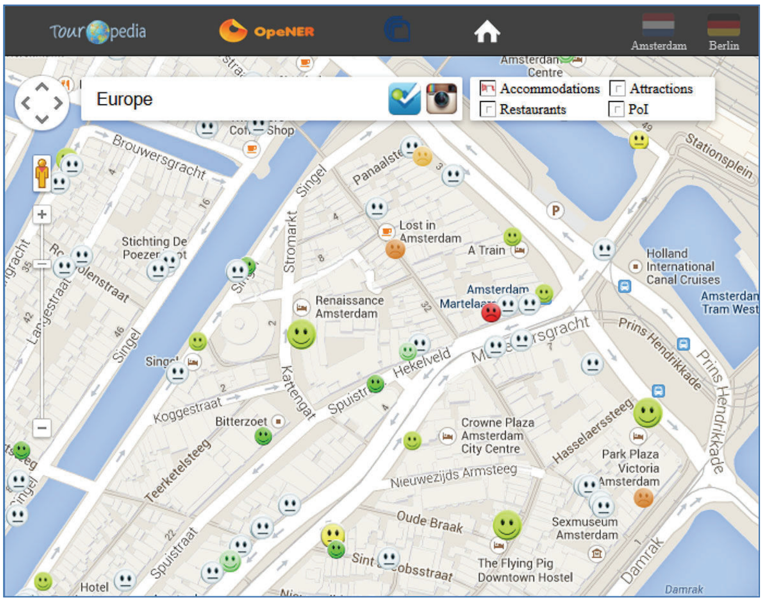


Fig. 11 User interface of Tour-pedia based on some of the proposed tools

the graphical interface of Tour-pedia[10] (Bacciu et al. 2014), an application that geolocates the results of the Sentiment Analysis of hotel reviews using emoticons to provide a quick overview of the feedback provided by customers in their reviews on the social media.

Reviews and other metadata (e.g. location metadata on a map) from customers have been extracted from different sources like Facebook, Google Places or FourSquare. Such reviews have been processed with the platform to obtain polarity measurements. Then, each location spot has been placed on an interactive map using an emoticon to show the overall aggregated sentiment. The emoticons (i.e. analysed locations) can be clicked to read the raw customer reviews. Tour-pedia is an illustrative example of how to build an added-value service on top of the text processing capabilities provided by the platform.

5 Conclusions

The large amount of customer-generated content emerging everyday over the Web is both a big challenge and a large opportunity. Specialised websites to write reviews and provide feedback allow customers to publish their opinion about products and services, clearly impacting the hospitality domain.

This paper describes a free Open Source NLP platform which aims at bringing text processing technologies a step closer to SMEs and other kind of end-users interested in analysing textual content. The ready-to-use tools and modules allow the creation of a customised analysis pipeline with Named Entity Recognition, Sentiment Analysis and Opinion Mining capabilities. The presented platform is based on a single data representation format (KAF) to enable a simple integration between the different modules and ease the extension and development of new modules and components. The provided tools work for six languages (English, Spanish, French, Italian, German and Dutch), but due to the modularity and interoperability provided by the use if KAF, it should be easy to extend or to include new tools by anyone, for example to add new languages or functionalities, as long as KAF format is respected.

The evaluation of the different tools composing the platform has been validated in the hospitality domain, in particular, on the basis of hotel reviews written by customers. During the development and customization of the platform to the hospitality domain, a set of hotel reviews has been manually annotated with sentiment and opinion related information. Such reviews have been then used to train and test the models that enable the work of the different modules.

Some of the provided tools help improving or further customising the platform for hospitality domain and also may serve for other domains of interest. The Open Source nature of the platform provides a good entry point to the language processing technologies and enables SMEs to extend the provided software and build their own analysers and products upon it. It can also help researchers to study customer

[10] http://tour-pedia.org/about/.

reviews to detect trends and patterns without the need of developing their own text processing tools from scratch.

Future research will be oriented towards increasing the domain adaptability of different tools, like the Sentiment Analysis and Opinion Detection modules. In particular, in order to build tools that can be easily used or ported in different languages, approaches that do not require language specific resources should be developed. In this way, such tools would be of great interest for a domain like hospitality, which is intrinsically multilingual.

References

Agerri R, Cuadros M, Gaines S, Rigau G (2013) OpeNER: Open Polarity Enhanced Named Entity Recognition. In: Proceedings of the 29th annual meeting of Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN'13. Madrid, España. Procesamiento del Lenguaje Natural, vol. 51, pp 215–218

Bacciu C, Lo Duca A, Marchetti A, Tesconi M (2014) Accommodations in Tuscany as Linked Data. In: Proceedings of the 9th edition of the language resources and evaluation conference

Bagga A, Baldwin B (1999) Cross-document event coreference: Annotations, experiments, and observations. In: Proceedings of the workshop on coreference and its applications

Bosma W, Vossen P, Soroa A (2009) KAF: a generic semantic annotation format. In: Proceedings of the GL2009 Workshop on semantic annotation

Brants T (2000) TnT: a statistical part-of-speech tagger. In: Proceedings of the sixth conference on Applied natural language processing, vol 1

Brereton RG, Lloyd GR (2010) Support vector machines for classification and regression. Analyst 135:230–267

Browning V, So KKF, Sparks B (2013) The influence of online reviews on consumers' attributions of service quality and control for service standards in hotels. J Travel Tour Mark 30(1–2):23–40

Cambria E, White B (2014) Jumping NLP curves: a review of natural language processing research [review article]. Comput Intell Mag IEEE 9(2):48–57

Cambria E, Schuller B, Xia Y, Havasi C (2013) New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 2:15–21

Collins M (2002) Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In: Proceedings of the ACL-02 conference on empirical methods in natural language processing, pp 1–8

Derczynski L, Ritter A, Clark S, Bontcheva K (2013) Twitter part-of-speech tagging for all: overcoming sparse and noisy data. In: Proceedings of the recent advances in natural language processing, September, pp 198–206

Doan A, Ramakrishnan R, Halevy AY (2011) Crowdsourcing systems on the World-Wide Web. Commun ACM 54(4):86–96

Dunning T (1993) Accurate methods for the statistics of surprise and coincidence. Comput Linguist 19(1):61–74

Filieri R, McLeay F (2014) E-WOM and accommodation: an analysis of the factors that influence travelers' adoption of information from online reviews. J Travel Res. 53(1):44–57

Ghose A, Ipeirotis P, Li B (2009) The economic impact of user-generated content on the Internet: Combining text mining with demand estimation in the hotel industry. In: Proceedings of the 20th workshop on information systems and economics (WISE)

Giesbrecht E, Evert S (2009) Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German Web as Corpus. Web Corpus Workshop WAC 5:27

Gräbner D, Zanker M, Fliedl G, Fuchs M (2012) Classification of customer reviews based on sentiment analysis. In: Proceedings of the 19th conference on information and communication technologies in tourism (ENTER), pp 460–470

Hu M, Liu B (2004) Mining opinion features in customer reviews. AAAI. 4(4):755–760

Kasper W, Vela M (2011) Sentiment analysis for hotel reviews. Computational linguistics-applications conference, pp 45–52

Kim EEK, Mattila AS, Baloglu S (2011) Effects of gender and expertise on consumers' motivation to read online hotel reviews. Cornell Hosp Q. 52(4):399–406

Kiyavitskaya N, Zeni N, Cordy JR, Mich L, Mylopoulos J (2009) Cerno: light-weight tool support for semantic annotation of textual documents. Data Knowl Eng 68(12):1470–1492

Lau K, Lee K, Ho Y (2005) Text mining for the hotel industry. Cornell Hotel Restaur Adm Q 46(3):344–362

Lee H, Peirsman Y, Chang A, Chambers N, Surdeanu M, Jurafsky D (2011) Stanford' s multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: Proceedings of the fifteenth conference on computational natural language learning: shared task. Association for Computational Linguistics, pp 28–34

Lee MJ, Singh N, Chan ESW (2011b) Service failures and recovery actions in the hotel industry: a text-mining approach. J Vacation Mark 17(3):197–207

Litvin SW, Goldsmith RE, Pan B (2008) Electronic word-of-mouth in hospitality and tourism management. Tour Manag 29(3):458–468

Liu B (2010) Sentiment analysis and subjectivity. Handb Nat Lang Process 2:627–666

Liu Z, Park S (2015) What makes a useful online review? Implication for travel product websites. Tour Manag 47:140–151

Liu S, Law R, Rong J, Li G, Hall J (2013) Analyzing changes in hotel customers' expectations by trip mode. Int J Hosp Manag 34:359–371

Marrero M, Urbano J, Sánchez-Cuadrado S, Morato J, Gómez-Berbís JM (2012) Named entity recognition: fallacies, challenges and opportunities. Comput Stand Interfaces

Montejo-Ráez A, Díaz-Galiano MC, Martinez-Santiago F, Ureña-López LA (2014) Crowd explicit sentiment analysis. Knowl Based Syst 69:134–139

Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Lingvisticae Investig 30(1):3–26

O'Connor P (2008) User-generated content and travel: a case study on TripAd-visor.com. In: O'Connor P, Höpken W, Gretzel U (eds) Information and communication technologies in tourism, vol 2008. Springer, Vienna, pp 47–58

O'Reilly T (2005) What Is Web 2.0? Design patterns and business models for the next generation of software, September 30. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. Accessed 14 Dec 2015

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135

Park S-Y, Allen JP (2013) Responding to online reviews: problem solving and engagement in hotels. Cornell Hosp Q 54(1):64–73

Popescu A, Etzioni O (2005) Extracting product features and opinions from reviews. Nat Lang Process Text Min (October), pp 339–346

Ramanathan U, Ramanathan R (2011) Guests' perceptions on factors influencing customer loyalty: an analysis for UK hotels. Int J Contemp Hosp Manag 23(1):7–25

Rao D, McNamee P, Dredze M (2013) Entity linking: Finding extracted entities in a knowledge base. In: Poibeau T, Saggion H, Piskorski J, Yangarber R (eds) Multi-source, multilingual information extraction and summarization, part II. Springer, Berlin, Heidelberg, pp 93–115

Řehůřek R, Kolkus M (2009) Language identification on the web: extending the dictionary method. In: Gelbukh A (ed) Computational linguistics and intelligent text processing. Springer, Berlin, Heidelberg, pp 357–368

Sahlgren M (2005) An introduction to random indexing. In: Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE, vol. 5

Sil A, Cronin E, Nie P, Yang Y, Popescu A-M, Yates A (2012) Linking named entities to any database. EMNLP-CoNLL 2012, pp 116–127

Sun L, Mielens J, Baldridge J (2014) Parsing low-resource languages using Gibbs sampling for PCFGs with latent annotations. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2002, pp 290–300

Sutton C, McCallum A (2012) An introduction to conditional random fields. Found Trends Mach Learn 4:267–373

Webster JJ, Kit C (1992).Tokenization as the initial phase in NLP. Proceedings of COLING-92, pp 1106–1110

Widdows D, Cohen T (2010) The semantic vectors package: New algorithms and public tools for distributional semantics. In Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on IEEE, pp 9–15

Xiang Z, Schwartz Z, Gerdes JH, Uysal M (2015) What can big data and text analytics tell us about hotel guest experience and satisfaction? Int J Hosp Manag 44:120–130

Ye Q, Zhang Z, Law R (2009) Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Exp Syst Appl, 36(3):6527–6535 **(Elsevier Ltd)**

Ye Q, Law R, Gu B, Chen W (2011) The influence of user-generated content on traveler behavior: an empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Comput Hum Behav 27(2):634–639

Zhang Z, Wang F, Law R, Li D (2013) Factors influencing the effective-ness of online group buying in the restaurant industry. Int J Hosp Manag 35:237–245