2018 IEEE 5th International Conference on Data Science and Advanced Analytics

Portability of Aspect Based Sentiment Analysis: Thirty Minutes for a Proof of Concept

Luca Dini
Innoradiant
Grenoble, France
luca.dini@innoradiant.com

Paolo Curtoni
Innoradiant
Grenoble, France
paolo.curtoni@ innoradiant.com

Elena Melnikova
Innoradiant
Grenoble, France
elena.melnikova@innoradiant.com

***Abstract*—This paper describes a system for aspect based sentiment analysis based on the assumption that domain portability should be achieved with minimal manual configuration. The approach exploits the integration of dependency parsing, graph based extraction rules over dependency trees and distributional semantics techniques. Results are considered satisfying for a "proof of concept" demonstrator.**

Keywords—aspect based sentiment analysis, opinion extraction, unsupervised systems, work embedding, dependency parsing.

I. Introduction

Aspect based sentiment analysis has become a popular topic in recent years, due to the activity of several international conferences and evaluations. At the same time progresses in deep learning based methods have triggered a strong consensus on the adoption of ML based strategies to achieve high performances for the evaluated systems. In this paper we explore a different perspective on the topic, i.e. the one of ABSA in a completely "unseen" domain, where, crucially, no learning set is available. While relatively interesting from a scientific point of view, this is the typical context of application of ABSA in industrial applications, where the context and the product (entity) under study varies from customer to customer and the effort needed to create a learning set is not sustainable.

In this paper we will prove that by coupling symbolic treatment with word embedding expansion strategies, domain portability of ABSA can be achieved in matter of few minutes. On this respect we will use the SemEval 2016 benchmark for French uniquely as a test set, i.e. without training any ML component over gold annotations. After the industrial problem statement in Section II and a state of the art presentation in Section III, we will provide in section IV a description of our (mainly symbolic) system of sentiment and entity detection as it is. In section V we will show how the current baseline in sentiment detection can be improved by exploiting currently available opinion words as keys to retrieve further domain dependent opinionated words, while in section VI we will prove how reasonable detection of entities under judgement can be achieved by providing just a couple of seed words for entity type. Finally in section VII we will show that the methodology can be applied on any domain with comparable results.

II. Problem statement

ABSA is a task which is central to a number of industrial applications, ranging from e-reputation, crisis management, customer satisfaction assessment etc. Here we focus on a specific and novel application, i.e. capturing the voice of the customer in new product development (NPD). It is a well-known fact that the high rate of failure (76%, according to Nielsen France, 2014) in launching new products on the market is due to a low consideration of perspective users' needs and desires. In order to account for this deficiency a number of methods have been proposed ranging from traditional methods such as KANO ([14],[23]) to recent lean based NPD strategies ([17]). All of them are invariantly based on the idea of collecting user needs with tools such as questionnaires, interviews and focus groups. However with the development of social networks, reviews sites, forums, blogs etc. there is another important source for capturing user insights for NPD: users of products (in a wide sense) are indeed talking about them, about the way they use them, about the feelings they raise. Here it is where ABSA becomes central: whereas for applications such as e-reputation or brand monitoring capturing just the sentiment is largely enough for the specific purpose, for NPD it is crucial to capture the entity an opinion is referring to and the specific feature under judgment.

A. Towards a democratization of ABSA for NPD

ABSA for NPD is a novel technique and as such it might trigger doubts on its adoption: given the investments on NPD (198 000 M€ only in the cosmetics sector) it is normal to find a certain reluctance in abandoning traditional methodologies for voice of the customer collection in favor of social network based ABSA. In order to contrast this reluctance two conditions need to be satisfied. On the one hand, one must prove that ABSA is feasible and effective in a specific domain (Proof of Concept, POC); on the other hand the costs of a high quality in-production system must be affordable and comparable with traditional methodologies (according to Eurostat the spending of European PME in the manufacturing sector for NPD will be about 350,005.00 M€ in 2020, and PME usually have limited budget in terms of "voice of the customer" spending).

If we consider the fact that the range of product/services which are possible objects of ABSA studies is immense[1] it is clear that we must rely on almost completely unsupervised technologies for ABSA, which translates in the capability of performing the task *without* a learning corpus.

In the following of this paper we will describe our first steps for achieving such a result on the basis of the Semeval 2016 corpus for French (Museums and Restaurants). In this preliminary phase we will consider only sentiment and entity identification, while reserving feature identification for future works. In Section III we will provide an overview of the Semeval 2016 tasks as well as relevant literature concerning unsupervised ABSA. In section V to VII we will describe our methodology and the evaluation results, while

[1] The site of UNSPC reports more than 40,000 categories of products (https://www.unspsc.org).

978-1-5386-5090-5/18/$31.00 ©2018 IEEE
DOI 10.1109/DSAA.2018.00085

Section VIII will contain comparisons with analogous experiments as well as hints for future research on automatic feature extraction.

III. SEMEVAL 2016 AND PREVIOUS WORKS ON UNSUPERVISED ABSA

A. Semeval2016's overview

SemEval (Semantic Evaluation) is "an ongoing series of evaluations of computational semantic analysis systems "[2], organized since 1998. Its purpose is to evaluate semantic analysis systems. ABSA (Aspect Based Sentiment Analysis) was one of the tasks of this event introduced in 2014. This type of analysis, carried out the opinionated texts, provides information about consumer opinions on products and services which can help companies to evaluate the satisfaction and improve their business strategies. A generic ABSA task consists to analyze a corpus of unstructured texts and to extract fine-grained information from the user reviews. The goal of the ABSA task within SemEval is to directly compare different test datasets, approaches and methods to extract such information ([19]).

In 2016, ABSA provided 39 training and testing datasets for 8 languages and 7 domains. Most datasets come from customer reviews (especially for the domains of restaurants, laptops, mobile phones, digital camera, hotels and museums), only one dataset (telecommunication domain) comes from tweets. The subtasks of the sentence-level ABSA, were intended to identify all the opinion tuples encoding three types of information: Aspect category, Opinion Target Expression (OTE) and Sentiment polarity. Aspect is in turn a pair (E#A) composed of an Entity and an Attribute. Entity and attributes, chosen from a special inventory of entity types (e.g. "restaurant", "food",…) and attribute labels (e.g. "general", "prices"…) are the pairs towards which an opinion is expressed in a given sentence. Each E#A can be referred to a linguistic expression (OTE) and be assigned one polarity label.

The French contribution for ABSA ([1]) focused on the analysis of the restaurants and museums domains. For the restaurant domain both training and test data were provided for sentence-level ABSA subtask. As for museum reviews, only test data was available as a gold-standard for evaluation in the out-of-domain subtask.

The evaluation assesses whether a system correctly identifies the aspect categories towards which an opinion is expressed. The categories returned by a system are compared to the corresponding gold annotations and evaluated according to different measures (precision (P), recall (R) and F-1 scores). System performance for all slots is compared to baseline score. Baseline System selects categories and polarity values using Support Vector Machine (SVM) based on bag-of-words features ([1]).

B. Related works on unsupervised ABSA

Traditionally, in ABSA context, one problematic aspect is represented by the fact that, given the non negligible effort of annotation, learning corpora are not as large as needed, especially for languages other than English. This fact, as well as extension to "unseen" domains, pushed some researchers to explore unsupervised methods. For instance [9], to which we will compare in the last section, explores new architectures that can be used as feature extractors and classifiers for Aspect terms unsupervised detection.

Such unsupervised systems can be based on syntactic rules for automatic aspect terms detection ([14]), or graph representations ([8]) of interactions between aspect terms and opinions, but the vast majority exploits resources derived from distributional semantic principles (concretely, word embedding).

The benefits of word embedding used for ABSA were successfully shown in [22]. This approach, which is nevertheless supervised, characterizes an unconstrained system (in the Semeval jargon a system accessing information not included in the training set) for detecting Aspect Category, Opinion Target expression and Polarity. The used vectors were produced using the skip-gram model with 200 dimensions and were based on multiple ensembles, one for each E#A combination. Each ensemble returns the combinations of the scores of constrained and unconstrained systems. For Opinion Target expression, word embedding based features extend the constrained system. The resulting scores reveal, in general, rather high rating position of the unconstrained system based on word embedding.

In subsection VIII.A we will also compare with systems which represent a compromise between supervised and unsupervised ABSA, i.e. semi-supervised ABSA systems, such as [12], an almost unsupervised system based on topic modelling and W2V, and W2VLDA ([8]). The former uses human annotated datasets for training, but enrich the feature space by exploiting large unlabeled corpora. The latter combines different unsupervised approaches, like word embedding and Latent Dirichlet Allocation (LDA, Blei et al. 2003) to classify the aspect terms into three Semeval categories. The only supervision required by the user is a single seed word per desired aspect and polarity. Because of that, the system can be applied to datasets of different languages and domains with almost no adaptation.

In general, in the literature, different unsupervised methods have been proposed to obtain vector based word meaning representations. They rely on algorithms like HAL and COALS and are available as open source in packages such as S-Space ([13]). Other popular algorithms are GloVe[3], CBOW and Skip-gram ([16] ; [18]). In the experiments described in this paper we exploit only the Skip-gram approach based on the Word2Vec[4] implementation. It is important to notice that this choice is not due to a principled decision but to non functional constraints related the fact that that algorithm has a java implementation, is reasonably fast and it is already integrated with Innoradiant NLP pipeline.

IV. SYSTEM DESCRIPTION

In this section we will provide a detailed description of the methodology we put in place to achieve POC-level ABSA. What we mean for POC-level ABSA is an entity based sentiment analysis in any given domain which could

[2] https://aclweb.org/aclwiki/SemEval_Portal, seen on 05/24/2018

[3] Available at http://www-nlp.stanford.edu/projects/glove/

[4] Available at https://code.google.com/p/word2vec/

be shown to a potential partner to prove the quality of ABSA in her specific domain, thus triggering investments for further system development. At this level we must produce a system which is not exhaustive but delivers reasonably precise results for specific entities.

In this section we will provide a generic description of one of the systems currently in use at Innoradiant (https://innoradiant.com). In subsection V we will describe how domain adaptation is achieved for sentiment analysis, while in section VI we will describe our approach to unsupervised entity enrichment.

A. System Description

The experiments have been performed by using Innoradiant's Architecture for Language Analytics (henceforth IALA). The platform implements a standard pipelined architecture composed of classical NLP modules: Sentence Splitting → Tokenization → POS tagging → lexicon access → Dependency Parsing → Feature identification → Attitude analysis. Most layers are based on a mix of machine learning based algorithms and rule based filters, in charge of performing error correction and domain adaptation. Attitude analysis and Entity identification represent an exception to this practice, as they are currently completely symbolic. In the following of this section we will describe these components in their *default settings*.

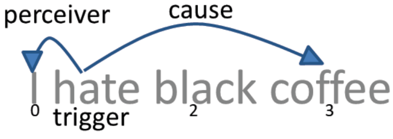
B. Default Sentiment Analysis

Inspired by ([3] and [4]) sentiment analysis in IALA is mainly symbolic. The basic idea is that dependency representations are an optimal input for rules computing sentiments. Contrary to the HOLMES system ([5]), however, here the rules take into account *at the same time* both the dependency structure *and* the linear word order. The reason underlying such a choice is that texts coming from social networks are often ill-formed or with a poor syntax, which causes errors in the parsing. Considerations of linear order help then to increase the robustness for all the cases where the dependency tree does not contain correct information (e.g. too many nodes pointing to root).

The rule language formalism is inspired by [21] and thanks to its template filling capability, in several cases, the grammar is able to identify the **perceiver** of a sentiment and, most importantly, the **cause** of the sentiment, represented by a word in an appropriate syntactic dependency with the sentiment-bearing lexical item. For instance the representation of the opinion in *I hate black coffee.* would be something such as:

```
 <Opinion trigger=″1″ perceiver=″0″ cause=″3″>.
```

(where integers represent position of words in a CONLL like structure) or in a graphical representation:



The set of triggers used by the grammars has been built over the time by using manual scrutiny of French social network corpora. Each lexical trigger is associated with an *arity* (FrameNet like argument structure) which allows the correct assignment of roles such as **cause** and **perceiver**.

In this specific experiment we adopted the standard sentiment grammar for French, which is composed of about 70 phrasal rules. Lexical information on polarity (and other semantic features such as agentivity, stativity etc.) is contained in external lists of lexemes: in the domain independent setting, the lexicons contain 314 negative items and 200 positive ones. We will see in section V that one of the goals of the experiment is to enrich such lists with domain dependent information.

It goes without saying that, given this architecture, IALA is able to extract multiple opinions per sentence, contrary to most commercial systems, which are limited to a by-sentence tonality classification.

C. Default Entity Identification

By default, entities (which are normally products and services under analysis) are identified since early processing phases by means of regular expressions. This choice is rooted in the fact that by acting at this level multiword entities (such as *hydrating cream*) are captured as single words since early stages. Moreover character based regular expressions are a handy tool for productively capturing misspelled forms, which are very frequent in languages such as French.

At the entity recognition phase IALA keeps all possible interpretation *ambiguities* and *under-specifications*. We consider for instance the word *cream* ambiguous with respect to *cooking cream* and *cosmetic cream* and underspecified when used in a cosmetic context without the specification of the specific kind of cream (such as its second occurrence in *"I love hydrating creams, but those creams are not easily affordable"*.[5] Disambiguation and/or resolution of under specification are performed at higher level of analysis via discourse analysis and word vector distance techniques. We do not dig more into this subject as entity disambiguation was not necessary for the task described in the rest of this paper.

D. External Resources

We base the automatic configuration algorithm described in the following of this paper on an unsupervised resource, i.e. an in-domain corpus without any manual annotation. In the case of the museums task we harvested from the TripAdvisor site all reviews concerning museums in French. After the analysis of all documents we ended up with a corpus of 487 818 sentences and 8,620 million lemmas.

From this corpus we obtained a Word2Vec resource by using the DL4j library (skip-gram).[6] The resource (W2VM(useums), henceforth) was obtained by using

[5] The problem could also be analyzed as a case of nominal anaphora.

https://deeplearning4j.org/word2vec.html lemmas rather than surface forms. Relevant training parameters for reproducing the model are:

- number of features in the word vector:200
- window size:5
- iterations for each mini-batch:5

Moreover, in order to filter noise, we skipped sentences with less than three and more than fifty words.

V. Unsupervised Sentiment Tuning

It is a well-known fact ([11]) that polarity words may vary from domain to domain: a word such as "aggressive" might be negative in the domain of call centers but positive in the domain of car design. "Delicate" is positive in the domain of food, but negative in the domain of heavy duty hardware. For these reasons the base grammar for French, which is assumed to work across multiple domains, only includes polarity items which can considered definitely positive or definitely negative, irrespective of the domain. The goal of this experiment is to find an unsupervised way to expand such a list of opinionated terms (triggers) according to the specific domain of analysis.

Our baseline in the museums domain is obtained by running the standard IALA pipeline for sentiment analysis without any modification. Under these conditions the system achieves a precision of 0.8543689, a recall of 0.40552995, and an F-measure of 0.54999995[7], which characterize a high precision system, with a low recall.

In order to increase the recall, and according to a methodology already applied, for instance by [22], we use W2VM to obtain a list of in-domain polarity words out of polarized seed items (opinion triggers already present in the grammar). Rather than expanding our lexical lists, however, we make use of only those words that IALA considered as polarity words *in the specific corpus*. We obtain a set of 80 positive and 55 negative lexical items that we use for expansion (polarity seeds).

The first and simplest expansion strategy we tested is just to add to the lexical polarity list the ten words closest to each polarity seed in W2VM, without any filtering. Closeness is just defined by cosine similarity among vectors. With this simple assumption we already obtain a non-marginal improvement of the baseline system: recall increases by 29 % (0.523) while precision only decreases by 14% (0.730) , with an F-MEASURE of 0.61.

One of the drawbacks of this method is the tendency of W2VM to show among the closest lemmas of a polarized word its opposite (antonym), a fact that causes the observed decrease in precision (for instance the closest word to "heureusement" (happily) is "dommage" (too bad)). In other words, with this method we do not exploit the fact that a good candidate expansion should be, for instance, close to a positive lemma, but also distant from all negative ones.

A more precise strategy is therefore to compute two abstract vectors POS and NEG, each one representing the average of the vectors of polarity seeds of a given polarity and the ones characterized by the inverse polarity multiplied by -1. With this vector we apply again the vector similarity strategy to obtain the N closest world to POS or NEG.

Of course, under the latter hypothesis, it is crucial to know what is the best value of N in order to optimize our measures. Fig. 1 shows the variation of precision, recall and F-measure in relation to N (with steps of 30 words starting from 10 up to 1180). We observe that the F-measure tends to increase also for high values, which would suggest, in an evaluation context, to keep a rather high top N value (e.g. around 900, which gives an average F-measure of 0.65). However, as stated elsewhere, our goal is to create a method for fast prototyping, not entering into a competition: as the prototype must be operated by a potential customer, our experience shows that, in general, she ignores what F-measure is, she is very sensitive to errors, and she tends to ignore missed values. As a consequence, in the following of this paper we adopt a TopN value of 250, which allows keeping the precision value above 0.7, with an F-measure of 0.63).

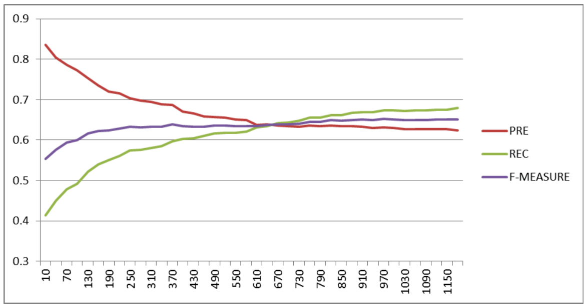


Fig. 1. Variation of quality measures with different top N candidates.

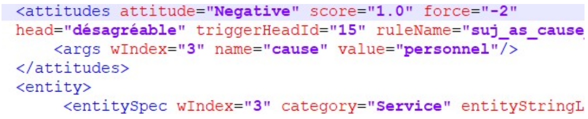
VI. Minimally Supervised Entity Tuning

As stated in section IV.C in IALA we associate entities to linguistic expressions which are recognized at early stages in processing. They are normally associated to opinions via a syntactic link: if a word playing the role of **cause** in a sentiment has been recognized as an entity, then we associate that opinion to that entity. Moreover we normalize single entities as *categories* of entity. For instance "personnel" and "reception" could bear the category "Service". In Fig. 2 we provide an example of a fragment of information delivered by IALA for the sentence "Par contre le personnel qui surveille les salles (une femme surtout) est très désagréable." ("On the contrary the personnel watching the halls (particularly a woman)is very unpleasant"), where the link between "unpleasant" and "personnel" is made by virtue of the fact that the argument **cause** of the attitude and index of the entity are the same (3).

[7] The Semeval 2016 evaluation script was run with parameters "Eval -vrb n -evs 1 -phs A -sbt SB1" with sentiments or entity#sentiment pairs considered as categories. For instance, according to the annotation guidelines, "…for the sentence "*The prices were CHEAP compared to the quality of service and food*" the following annotations should be included in the xml file.

```
<Opinion category="RESTAURANT#PRICES"/>
<Opinion category="SERVICE#GENERAL />
<Opinion category="FOOD#QUALITY"/>
```

The obvious consequence of such an assumption is that the evaluation program will check gold/system correctness at the *set level*,i.e. after removal of duplicates.



```
<attitudes attitude="Negative" score="1.0" force="-2"
head="désagréable" triggerHeadId="15" ruleName="suj_as_cause
    <args wIndex="3" name="cause" value="personnel"/>
</attitudes>
<entity>
     <entitySpec wIndex="3" category="Service" entityStringL
```

Fig. 2. Syntactic association of sentiment and entities..

In a normal context a relevant configuration effort must be devoted in coding interesting entities and associating categories to them. The Semeval 2016 corpus gives us the opportunity to test to which extent this operation of entity coding and entity categorization could be automated.

A. The Baseline

Of course we cannot assume that everything can be obtained without supervision: indeed the corpus/evaluation comes with a *predefined set of categories* and there is no automatic way of associating these labels to entities without using the learning corpus or manual coding. As the challenge of this experiment was *not* to use the gold annotations, we devised a methodology of term expansion on the basis of a minimal human input. A linguist was therefore asked to associate two words to each Semeval entity on purely intuitive basis. The coding of these entities took less than half an hour and the associations for the Museums corpus are summarized in TABLE I.

TABLE I. SEED WORDS FOR THE MUSEUM DOMAIN

Categories	Seeds
Museum	musée, galerie
Location	lieu, vue
Service	service, accueil
Facilities	restaurant, boutique
TourGuiding	guide, tour
Collections	collection, oeuvre

We can consider this minimal configuration work as our baseline. Of course if we perform an evaluation over this minimal configuration we obtain absolutely risible results. Over a set of extracted opinions only 555 have an explicit **cause**, and in only 60 cases the cause corresponds to a coded entity (lemma equivalence with the "seeds" column of TABLE I). In all other cases (769) we categorized opinions with an entity which was *randomly* chosen from the 6 available Museum/Semeval entities. The choice of the random case is very important for making sense of this kind of evaluation. Indeed, one could decide to consider the random case as the most frequent category. However the concept of "most frequent category" only makes sense in an already annotated corpus and it is in principle not available on a brand new corpus/domain: hence the exclusion of a fixed default category strategy and the choice of a random assignment in case of non-decision by the system.[8] It should also be noticed that in all our evaluation experiments we work on entity assignment on the basis *of the output of IALA enhanced sentiment assignment*. This means that our overall F-measure on entity detection could never be higher than the 63% reached for unsupervised sentiment extraction and described in section V (i.e. a system with perfect entity assignment would produce an F-measure of 63%). We understand, of course, that from a scientific point of view this all-in-one evaluation obfuscates the responsibility of the different algorithms, and this might be the reason why such an evaluation was not proposed in Semeval 2016. On the other hand, in real life analysis there is no other possibility than working with entity and sentiments at the same time, given the fact that retrieving entities only is of little interest to the customer.

On these bases, the baseline for opinion/entity detection is: precision: 0.123 recall: 0.095 and F-measure: 0.107

B. Entity Expansion

The obvious candidate methodology for increasing the performance of the system is of course the automatic population of entity lists on the basis of the manually coded entities (seed entities). Thus our algorithm would just take the top N words which are closest to an entity seed, assign them the same category and add them to the list of entities. In this phase some additional filtering is performed, namely:

- Words are automatically normalized to account for common misspells, such as accentuation.
- Words are filtered according to their POS: only nouns are retained.
- If the same word appears in the expansion of two or more seeds of different categories, we only keep the instance with highest association to the seed (Cosine distance).

In order to determine the optimal top N number we could not, for internal system reason, perform evaluation on a large range of top numbers as we did in Fig. 1. Therefore we performed only evaluation at 10 (F-measure: 0.148), 20 (F-measure:0.154), 30 (F-measure:0.145), 50 (F-measure:0.147), 100 (F-measure:0.141) 200 (F-measure:0.135) and 400 (F-measure:0.135). For the following we consider 20 top N expansions a good value. We also tried to increase the quality by imposing a minimal similarity between the seed and the expansion (0.4), but this did not end up in any improvement.

C. Entity Matching

The quality of the results obtained in the previous section is evidently influenced by the fact that only opinions with a syntactically identified cause are taken into account and we have seen that for 769 opinions the system could not detect any cause. To this we must add the fact that in the phase of polarity words expansion some error in cause detection was introduced as well, for instance due to the incapacity of making the distinction between verbs such as *hate* and *disgust* as exemplified in "I hate **cheese**" and "**cheese** disgusts me" (in bold the **cause** argument as identified by the grammar).

In order to mitigate this effect we implemented an additional entity computation algorithm based on the consideration of the right and left context of an opinion bearing element (e.g. "désagréable" in the example in Fig. 2.) In this case we take

[8] For the sake of completeness we report that a fake run always returning "Museum" as default would have achieved an F-measure of 0.34

the N left and N right full words adjacent to the opinion word and we compute their average vector. We compare this vector with the average vector obtained from the two seeds in the same category (category vector). The winning category is simply assigned as the entity to which the opinion refers. This method was tested with textual windows ranging from 5 to 45, and best results were obtained with a window of 35 words: precision: 0.412, recall: 0.338, F-measure: 0.371. We tried to increase the precision of the results by imposing a minimal threshold for assigning an opinion to a specific entity, but the results where deceiving (F-measure: 0.32): this is evidently due to the fact that a non-assignment triggers the default random assignment which is always worse than W2VM window-based assignment, even below the target threshold.

Finally we measured if the W2V window-based method alone was not performing better than our hybrid approach, which privileges first lexical matching of a cause and then, if it provides no result, applies a window based method. The results prove that the hybrid method is still superior, as without considering causal role in opinion we obtain a decrease both in precision (0.403) and recall(0.313).

VII. Testing on a new Corpus

The goal of this research was to study how far a completely unsupervised method could improve our baseline grammar-based analysis. Our methodology, however, was partially biased: indeed we used anyway a gold standard for tuning a few parameters of our system, such as top N expansion, window size, etc. In order to prove that the system is really unsupervised we applied the same parameters to a completely new corpus, i.e. the Semeval 2016 corpus on *restaurants*. Gold annotations were used for evaluation only. We also trained a new W2V resource (W2VR) using the same approach as with museums: this time the corpus is much bigger, totaling to of 3,834,240 sentence and 65,088,072 lemmas.

As for seeding entities, TABLE II. shows the lexical choices made by our linguists.

TABLE II. Seed words for the Restaurant Domain

Categories	Seeds
RESTAURANT	restaurant, bistrot
FOOD	cuisine, pizza
DRINK	vin,bière
SERVICE	service, accueil
LOCATION	lieu, vue
AMBIENCE	ambiance, cadre

Procedurally, the first step was to test sentiment detection results using the same parameters as in section V. The results are quite conclusive as we get a precision of 0.781 and a recall of 0.545 (F-measure: 0.642), which are completely in line with the results obtained on the Museums corpus.

The same holds concerning Sentiment#Entity pairs recognition: by applying the same parameters (expansion Top N=25 and windows size=35) we obtain results that are even better than the ones of the Museums corpus: precision: 0.4791, recall: 0.488 and F-measure: 0.399. We tend to attribute this effect to the size of the corpus over which the W2VR resource was computed, which was eight times larger.

Finally we performed the same experiment with the Restaurants corpus but using a generic W2V resource in order to understand the influence of the corpus specificity. The generic W2V was obtained from a balanced 140 million words corpus (newspapers, social networks and fiction). As expected, we observe a dramatic decrease at the level of entity/sentiment pair detection (precision=0.287, recall=0.193, F-measure=0.231) as well a deterioration with respect to the sentiment only detection, for which the F-measure descends from 0.64 to 0.60. We notice, parenthetically, that these results are even inferior to the grammar only threshold, which was 0.61379516. These results are in line with [6], even though the delta between generic and domain specific W2V is less impressing.

VIII. Evaluation and Future Research

A. Comparison with State of the Art

TABLE III. Summary of the results achieved with expansion vs. grammar only baseline

	Museums					
	Sentiment			Sentiment#Entity		
	P	R	F	P	R	F
Baseline	0.85	0.40	0.54	0.12	0.09	0.10
With expansion	0.70	0.57	0.63	0.41	0.33	0.37
	Restaurants					
	Sentiment			Sentiment#Entity		
	P	R	F	P	R	F
Baseline	0.86	0.39	0.53	0.12	0.08	0.09
With expansion	0.78	0.54	0.64	0.47	0.48	0.40

It would be interesting to compare the results we achieved, reported in table III for convenience, with previous experiments, especially in the Semeval 2016 context. Unfortunately this is difficult, as in that exercise the task of detecting aspect (entity+attribute) was separated from the one of detecting sentiments: indeed for the latter task the participants could rely on an annotated text where aspect was already identified, which makes results hardly comparable. Still we notice that on the restaurant dataset our semi-unsupervised approach to entity detection obtains an F-measure of 62.5 *when computed on correctly recognized sentiment only*. This compares quite positively with the baseline of SemEval-2016, which was 52.6, and even with the best performing system (61.2), even though it is evident that the fact of recognizing entity only (vs. entity+aspect) introduces definitely a positive bias on our system.

On the other hand Semeval offered an Out-of-domain ABSA task described in [19] as “In SB3 participants had the opportunity to test their systems in domains for which no training data was made available; the domains remained unknown until the start of the evaluation period. Test data for SB3 were provided only for the museums domain in

French.". Unfortunately, according to the same paper, there was no submission for this task.[9]

A more pertinent comparison might be made with results mentioned in [7], which describes an unsupervised experiment on the dataset of Semeval 2015 (In English): the approach is comparable with ours, with a massive use of W2V to expand lexical seeds, even if the layer of domain independent rules is absent. Unfortunately, the results cannot be compared as i) they perform entity *and* aspect identification, which is more difficult than entity only; ii) they perform opinion identification on the basis of a gold standard on annotated entities, which is an easier task than unrestricted opinion extraction. The consequence of the latter assumption is that sentiment analysis is viewed as a classification problem, whereas in our case is seen as an information extraction problem: The accuracy measure they report on the restaurant domain (0,69) is therefore not exactly comparable with our F-measure; nevertheless we observe that if in the same domain (but different language) we simply compute accuracy just as a ratio between correctly recognized sentiments and retrieved sentiments, we obtain an accuracy of 0.78, which is probably explained by the use of domain independent polarity grammars. As for entity detection, the F-measure they report (0.41) is quite similar to ours (0.39), but it is unclear if they are comparable, given the fact that their system detect Entity#Attribute pairs, whereas we identify Entity#Opinion pairs.

The same holds for [8] where the authors improve the system presented at Semeval 2015 by coupling the original W2V core with guided topic modelling based on LDA and prove the language independence of the system by applying it to different Semeval 2016 languages. Unfortunately the results are still incomparable given the separation they operate between entity detection and opinion classification. Moreover they operate a simplification of data by retaining (in the restaurant domain) only the FOOD, SERVICE and AMBIENCE categories and only sentences with just one assignment, which makes any comparison impossible. However, we retain from that paper the benefits provided by the use of LDA. We also value the important conclusion that, for aspect identification, the passage from two to three words in the entity seed increases the accuracy by 3% (from 0.73 to 0.76).

Another inspiring approach is represented by [10], which achieves unsupervised Aspect Term Extraction (ATE) by training a CRF based classifier over a completely automatically annotated corpus, obtaining results (F of 0.42 and 0.59 for laptops and restaurants respectively) which are higher than the Semeval 2014 supervised baseline. Again, a comparison with our approach is almost impossible, as the task is limited to identify aspect terms present in a sentence without any further characterization (cf. also [9]). Still, also in this case, their methodology contains interesting hints such as the use of a terminology extraction system ([20]) to filter relevant candidates.

B. Conclusion and Further Works

In this paper we presented a method for achieving ABSA in new domains with minimal supervision (less than half a hour for encoding the seeds). The goal was not to obtain a full-fledged ABSA system, but a proof of concept quality level system, i.e. a "demonstrator" showing the potentialities of these technologies to a perspective customer. While we provided formal evaluation against available gold standards, it is unclear to determine if the POC quality level was achieved. [10] This might depend from factors which are external to the evaluation context. For instance both sets of labels contain one generic label (Museum and RESTAURANT) which we treated exactly as any other category (random assignment in case of system's uncertainty). In a real context it would be wiser to privilege these categories for all cases where the system has no evidence for a more specific assignment: a user is more likely to tolerate underspecification than errors. Also, in a real context some additional time could be devoted to manually filtering polarity and entity expansions words which are clearly system errors. All this would contribute, in our view, to a prototype convincing enough to trigger a phase of further development.

Concerning this phase, it is worth to mention the fact that the present methodology is heavily based on a skeleton of symbolic rules and expansion lists which can be both manually revised: they represent a solid starting point for the development of a full-fledged ABSA, even in absence of a learning set.

From a more scientific point of view, there is room for improvements into many directions. We have already mentioned the integration with LDA and the use of a terminology extraction system for a more precise identification of candidates. It would be interesting also to investigate the effects of the substitution of the plain textual window we used, with more constrained strategy, such as distance on a dependency tree, may be with weights on arc traversal. Finally the system would benefit of a better identification of the lexical **cause** of opinions: in the Museums experiment only 555 opinions over 1324 have an explicit cause, which evidently forces the system to rely on the less precise window-based method.

References

[1] M. Apidianaki, X. Tannier and C. Richart, "Datasets for Aspect-Based Sentiment Analysis in French", Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia, May 2016. 2016.

[2] M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation", in The Journal of machine Learning research, Vol. 3, pp. 993–1022.

[3] L. Dini, A. Bittar, C. Robin, F. Segond and M. Montaner. "SOMA: The Smart Social Customer Relationship Management", in Sentiment Analysis in Social networks. Chapter 13. 2017, pp.197-209. DOI: 10.1016/B978-0-12-804412-4.00013-9.

[9] "No submissions were made for sb3-muse-fr & sb1-telctu."

[10] We understand that the concept of POC quality level is quite fuzzy. On this respect we are trying to set up experiments with our customer to have a more precise characterization of the "minimal viable quality" and how it relates with more scientific measures.

[4] L. Dini and A. Bittar, "Emotion Analysis on Twitter: The Hidden Challenge", Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.

[5] M. Dupuch, F. Segond, A. Bittar, L. Dini, L. Soualmia, S. Darmoni, Q. Gicquel and M.H. Metzger, "Separate the grain from the chaff: make the best use of language and knowledge technologies to model textual medical data extracted from electronic health records", Proceedings of the 6th Language & Technology Conference, 2013.

[6] Dusserre and M. Padró, "Bigger does not mean better ! We prefer specificity". Poceedings of the 12th International Conference on Computational Semantics (IWCS). 19-22 September 2017, Montpellier (France), 2017.

[7] García-Pablos M. Cuadros and G. Rigau, "V3: Unsupervised Aspect Based Sentiment Analysis for SemEval2015 Task 12." Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, June 4-5, 2015, pp. 714–718.

[8] García-Pablos, M. Cuadros and G. Rigau, "W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis", Expert Systems with Applications, Vol. 91, 2018, pp. 127-137.

[9] Giannakopoulos, D. Antognini, C. Musat, A. Hossmann and M. Baeriswyl "Dataset Construction via Attention for Aspect Term Extraction with Distant Supervision.", 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 2017, pp. 373-380.

[10] Giannakopoulos, C. Musat, A. Hossmann and M. Baeriswyl "Unsupervised Aspect Term Extraction with B-LSTM & CRF using Automatically Labelled Datasets", in press, arXiv:1709.05094v1, [cs.CL], 15 Sep 2017, WASSA@EMNLP, 2017.

[11] W.L. Hamilton, K. Clark, J. Leskovec, D. Jurafsky, "Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora.", CoRR, 2016

[12] T. Hercig, T. Brychcín, L. Svoboda, M. Konkol and J. Steinberger, "Unsupervised Methods to Improve Aspect-Based Sentiment Analysis in Czech", Computación y Sistemas, vol. 20, No. 3, 2016, pp. 365-375.

[13] Jurgens and K. Stevens, "The S-Space package: An open source package for word space models", in System Papers of the Association of Computational Linguistics, 2010.

[14] Kano N., Seraku Nk, Takahashi F., Tsuji. "Attractive quality and must be quality". Quality,14:39-48, 1984.

[15] K. Liu, L. Xu and Zhao, "Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Jun 2014, 2014.

[16] T. Mikolov, W.-t. Yih and G. Zweig, "Linguistic regularities in continuous space word representations". Proceedings of NAACL-HLT, 2013, pp. 746–751.

[17] D. Olsen, "The Lean Product Playbook: How to Innovate with Minimum Viable Products and Rapid Customer Feedback", 2015.

[18] J. Pennington, R. Socher and C. D. Manning, "Glove: Global vectors for word representation", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532-1543.

[19] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clecq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S.M. Jiménez-Zafra, G. Eryiğit. "SemEval-2016 Task 5: Aspect Based Sentiment Analysis", in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016). San Diego, USA, June 2016.

[20] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss and J. Han, "Automated Phrase Mining from Massive Text Corpora", IEEE Transactions on Knowledge and Data Engineering, in press, 2018.

[21] M. A. Valenzuela-Escárcega, G. Hahn-Powell and M. Surdeanu, "Description of the Odin Event Extraction Framework and Rule Language", arXiv:1509.07513v1 [cs.CL], 24 Sep 2015, version 1.0, 2015.

[22] D. Xenos, P. Theodorakakos, J. Pavlopoulos, P. Malakasiotis, I. Androutsopoulos, "AUEB-ABSA at SemEval-2016 Task 5: Ensembles of Classifiers and Embeddings for Aspect Based Sentiment Analysis", Proceedings of SemEval-2016, San Diego, California, June 16-17, 2016, pp. 312–317.

[23] L. Witell, M. Löfgren, J. Dahlgaard, "Theory of attractive quality and the Kano methodology – the past, the present, and the future", in Total Quality Management & Business Excellence, Vol. 24, issue 11-12, 2013, pp. 1241-1252. DOI: 10.1080/14783363.2013.791117.