W2VLDA: Almost Unsupervised System for Aspect Based Sentiment Analysis

Aitor García-Pablos[a,*], Montse Cuadros[a], German Rigau[b]

[a] Vicomtech-IK4, Mikeletegi 57, San Sebastian, Spain
[b] IXA Group, EHU, Manuel Lardizabal 1, San Sebastian, Spain

Abstract

With the increase of online customer opinions in specialised websites and social networks, the necessity of automatic systems to help to organise and classify customer reviews by domain-specific aspect/categories and sentiment polarity is more important than ever. Supervised approaches for Aspect Based Sentiment Analysis obtain good results for the domain/language they are trained on, but having manually labelled data for training supervised systems for all domains and languages is usually very costly and time consuming. In this work we describe W2VLDA, an almost unsupervised system based on topic modelling, that combined with some other unsupervised methods and a minimal configuration, performs aspect/category classification, aspect-terms/opinion-words separation and sentiment polarity classification for any given domain and language. We evaluate the performance of the aspect and sentiment classification in the multilingual SemEval 2016 task 5 (ABSA) dataset. We show competitive results for several languages (English, Spanish, French and Dutch) and domains (hotels, restaurants, electronic devices).

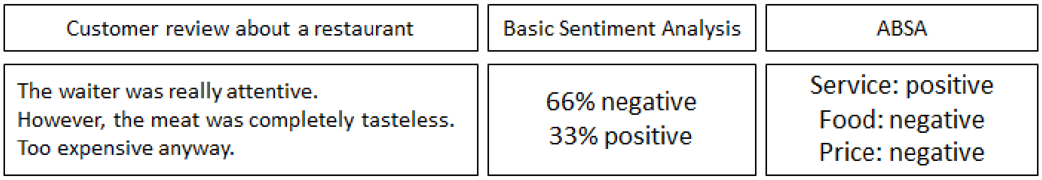
Keywords: Sentiment Analysis, Almost Unsupervised, Multilingual, Multidomain

1. Introduction

During the last decade, the Web has become one of the most important sources for customers and providers to evaluate and compare products

*Corresponding author

Email addresses: agarciap@vicomtech.org (Aitor García-Pablos), mcuadros@vicomtech.org (Montse Cuadros), german.rigau@ehu.eus (German Rigau)



Customer review about a restaurant	Basic Sentiment Analysis	ABSA
The waiter was really attentive. However, the meat was completely tasteless. Too expensive anyway.	66% negative 33% positive	Service: positive Food: negative Price: negative

Figure 1: An example of classical Sentiment Analysis vs. Aspect Based Sentiment Analysis

and services. The vast amount of content generated every day in countless websites and social networks keeps growing and requires automated ways to handle and classify all these opinions. Because of that, many different algorithms and approaches have been developed in the area of Opinion Mining.

Opinion Mining is a subfield of Natural Language Processing (NLP) that deals with the automatic analysis of opinions shared by humans in different contexts, like in customer reviews (Pang and Lee, 2008; Liu, 2012). Aspect Based Sentiment Analysis (ABSA) refers to the systems that determine the opinions or sentiments expressed on different features or aspects of the products/services under evaluation (e.g. battery or performance for a laptop). An ABSA system should be capable of classifying each opinion according to the aspects relevant for each domain in addition to classifying its sentiment polarity (usually positive, negative or neutral), as depicted in figure 1.

Best performing ABSA systems generally use manually labelled data and language specific resources for training on a particular domain and for a particular language (Pontiki et al., 2014, 2015, 2016). This is the case of deep-learning based systems, that provide very good performance but require a significant amount of labelled data for training (Chen et al., 2017; Araque et al., 2017).

On the other hand, weakly-supervised systems do not require labelled data for training, but they usually need some language specific resources, such as carefully curated lists of seed words or language dependent tools to preprocess the input (Lin et al., 2011; Jo and Oh, 2011; Kim et al., 2013). In addition, most of these works only report results for English.

In this work, we present W2VLDA, an almost unsupervised system for multilingual and multidomain ABSA, that works leveraging large quantities of unlabelled textual data and an initial configuration consisting of a minimal set of seed words. Figure 2 shows an schema of W2VLDA. Imagine the following scenario. The owners of a famous restaurant want to monitor the opinion of their costumers with respect to a set of aspects. In particular,

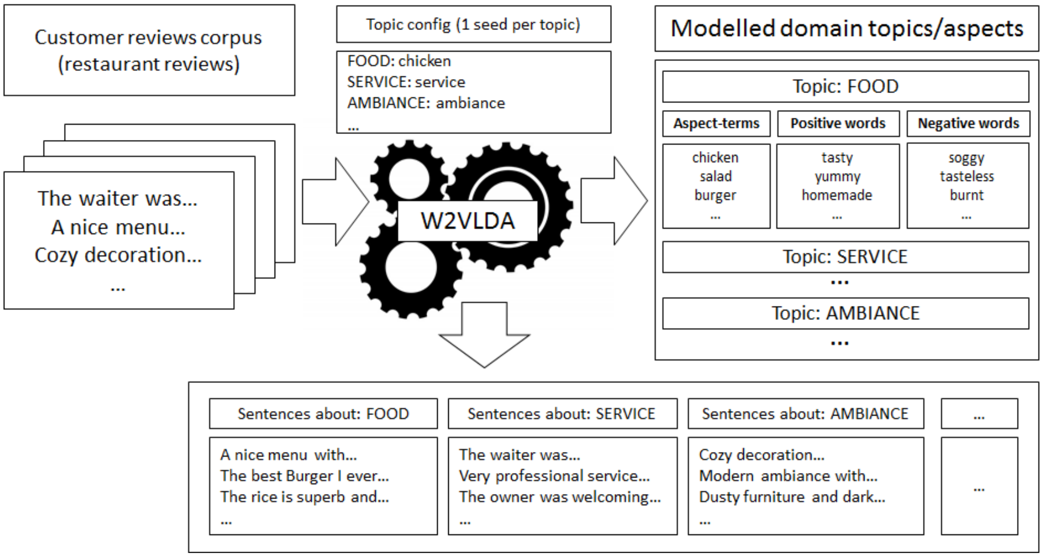


Figure 2: An schema of W2VLDA. The input is an unlabelled corpus o a particular domain and the topic specification. Topics are split into three word distributions: aspect-terms, positive words and negative words to ease the interpretation of each topic. Sentences are modelled by topic/aspect and polarity.

they want to know the opinion about its food, service, price, ambience, location, etc. The input of W2VLDA is a corpus of customer reviews and an example word per aspect they want to monitor (for instance, chicken for the aspect food, service for the aspect service, etc.)[1]. With this input, W2VLDA produces two main outputs. First, a weighted list of words per aspect (for instance, chicken, salad, burger, etc. for the aspect food), a weighted list of positive words (tasty, yummy, homemade, etc.) and weighted list of negative words (soggy, tasteless, burnt, etc.) for every selected aspect. Thus, our system performs at a word level three subtasks simultaneously: aspect classification, aspect-term/opinion-word separation, and sentiment polarity classification. Second, W2VLDA also produces a weighted list of sentences for every selected domain aspect and polarity.

The system is based on a topic modelling approach combined with continuous word embeddings and a Maximum Entropy classifier. It runs over

[1] W2VLDA also needs an example of a positive and a negative word (for instance, excellent and horrible).

an unlabelled corpus of the target language and domain just by defining the desired aspects with a single seed-word per aspect. We show results for different domains (restaurants, hotels, electronic devices) and languages (English, Spanish, French and Dutch). We compare its performance with other topic modelling based approaches, and we evaluate the performance of this approach on the SemEval2016 task 5 dataset, which provides a manually labelled set of restaurant reviews for several languages, including English, Spanish, French and Dutch. The contributions of this work are the minimal need of supervision (just one seed word per aspect/polarity) to perform ABSA over any unlabelled corpus of customer reviews. The lack of language or domain specific requirements allows the system to be readily used for other languages and domains. Another contribution is the automatic separation of the topic words into aspect-terms, positive words and negative words to improve the readability of the generated topics. We will leave the source code publicly available[2].

After this short introduction, the paper is structured as follows. First, section 2 reviews previous related work. Then, section 3 describes our system, including the seed-word based configuration, the aspect-term/opinion-word separation and the topic modelling part. After that, section 4 shows the results and evaluation. Finally, section 5 describes the conclusions and future work.

2. Related work

During the last decade the research community has addressed the problem of analysing user opinions, particularly focused on online customer reviews (Liu et al., 2012; Chen et al., 2014). The problem of customer opinion analysis can be divided into several subtasks, such as detecting the aspect (aspect classification) and detecting the opinion about the aspect of the product being evaluated.

A common approach in the literature is to identify frequent nouns, lexical patterns, dependency relations applying supervised machine learning approaches (Hu and Liu, 2004; Popescu and Etzioni, 2007; Blair-Goldensohn et al., 2008; Wu et al., 2009; Qiu et al., 2011). Some works focus on automatically deriving the most likely polarity for words, constructing a so-called sentiment lexicon (Mostafa, 2013). The typical approaches use different variants

[2] `https://bitbucket.org/aitor-garcia-p/w2vlda-last/overview`

of bootstrapping or polarity propagation leveraging some base dictionaries and pre-existing linguistic resources (Rao and Ravichandran, 2009; Jijkoun et al., 2010; Huang et al., 2014).

A well-known unsupervised method for text modelling documents is Latent Dirichlet Allocation (LDA). LDA is a generative model introduced by (Blei et al., 2003) that quickly gained popularity because it is an unsupervised, flexible and extensible technique to model documents. LDA models documents as multinomial distributions of so-called topics. Topics are multinomial distributions of words over a fixed vocabulary. Topics can be interpreted as the categories from which each document is built up, and they can be used for several kinds of tasks, like dimensionality reduction or unsupervised clustering. Due to its flexibility, LDA has been extended and combined with other approaches, obtaining topic models that improve the resulting topics or that model additional information (Mcauliffe and Blei, 2008; Ramage et al., 2009).

Topic models have been applied to Sentiment Analysis to jointly model topics and sentiment of words (Lin et al., 2009, 2011; Jo and Oh, 2011; Lu et al., 2011; Kim et al., 2013; Alam et al., 2016). A usual way to guide a topic modelling process towards a particular objective is to bias the LDA hyperparameters using certain apriori information. In the case of modelling the polarity of the documents, it usually means using a carefully selected set of seed words. Our method follows this idea, but replaces the need for a carefully crafted list of language or domain polarity words by only a single domain independent positive word (e.g. *excellent* for English) and a single domain independent negative word (e.g. *horrible* for English).

In general, topics coming from a topic modelling approach are anonymous word distributions, requiring an additional step to map them to a meaningful domain category. This task requires a manual inspection by an expert or a mapping calculation to an existing resource (Bhatia et al., 2016). Our approach relies on a minimal topic configuration to define the topics for the target domain the user wants to monitor. Thus, the resulting topics match the ones defined initially by the user. This is done by leveraging semantic word similarities to guide the topic modelling towards the defined topics. This semantic word similarity is obtained using continuous word embeddings over the domain words. Continuous word embeddings are known for capturing semantic regularities of words (Mikolov et al., 2013a; Collobert and Weston, 2008). Some works have made use of this fact to improve the resulting topics (Das et al., 2015; Nguyen et al., 2015; Qiang et al., 2016), but their objective is to improve the unsupervised modelling of a corpus instead of guiding the model towards a predefined set of topics. There are works that exploit word embeddings in a supervised machine learning setting to perform sentiment analysis (Tang et al., 2014; Giatsoglou et al., 2017).

Some authors have also attempted an automatic aspect-term/opinion-word separation within the topic modelling process (Zhao et al., 2010; Mukherjee and Liu, 2012). Aspect terms are the words that are used to speak about the aspect being evaluated (e.g. *waiter* or *waitstaff* when speaking about the *service* of a restaurant). On the other hand, opinion words express the sentiment about an aspect, such as *attentive* or *terrible*. The separation of these two kinds of words might be useful because it eases the interpretation of the resulting topics, and the sentiment classification can be focused on the opinion-words which are more likely to bear sentiment information. Zhao et al. (2010) attempted this separation training a supervised classifier on a small manually labelled dataset and using Part-of-Speech tagging. Mukherjee and Liu (2012) elaborated on this idea trying a similar approach but substituting the manually labelled dataset with an existing lexicon of opinion words for English. Instead, we apply Brown clustering (Brown et al., 1992) to a set of training instances from an unlabelled corpus in order to train an aspect-term/opinion-word classifier that is later integrated into the topic modelling process. Following this approach, no additional language-dependent resources are required, and the full process could be applied to any language and domain.

In summary, combining topic modelling, continuous word embeddings and a minimal topic definition, our proposed approach can model customer reviews in different languages and domains performing three subtasks at the same time: aspect classification, sentiment classification and aspect-terms/opinion-words separation. To our knowledge, no other almost unsupervised system tries to perform these three tasks at the same time and without requiring any pre-existing language or domain dependent resource.

3. System description

The main objective of the W2VLDA system is to perform the three tasks (detecting aspects, opinions and their polarity) of Aspect Based Sentiment Analysis at the same time. That is, to classify pieces of text into a predefined set of domain aspects and classify their sentiment polarity as positive or negative. In addition, our system separates opinion words from aspect terms

Domain aspect or Polarity	Seeds (English)	Seeds (Spanish)	Seeds (French)
food	chicken	pollo	poulet
service	service	servicio	service
ambience	ambience	ambiente	ambiance
drinks	drinks	bebidas	boissons
location	location	ubicación	emplacement
positives	excellent	excelente	excellent
negatives	horrible	horrible	épouvantable

Table 1: Example of seed words (one per domain aspect) used to monitor certain aspects of restaurant reviews in several languages, including the general polarity seeds

without requiring additional resources or supervision. The system at its core consists of an LDA-based topic model extended with additional variables, with biased topic modelling hyperparameters based on continuous word embeddings, and combined with unsupervised pre-trained classification model for aspect-term/opinion-word separation.

3.1. Topics and sentiment configuration

W2VLDA only requires a minimal domain aspect and sentiment polarity configuration per language and domain. The configuration consists of a single seed to define each desired domain aspect, plus a single general positive seed word and a single general negative seed word valid for all domain aspects. This simple configuration is the only language and domain dependent information required by W2VLDA [3]. Therefore, a simple translation of the seeds should suffice to make the system work for another language or domain, as long as each translated seed has an equivalent meaning and use in the target language. Table 1 shows an example of a domain aspect and polarity configuration for the restaurant domain in several languages.

3.2. Aspect-term and opinion-word separation

Part of the outcome of the system consists of the aspect-term/opinion-word separation into differentiated word classes. In order to achieve this separation without adding any language dependent tool or resource, the system

[3] A list of general stopwords for each target language is also necessary in order to obtain better results. We use the stopword lists from Apache Lucene.

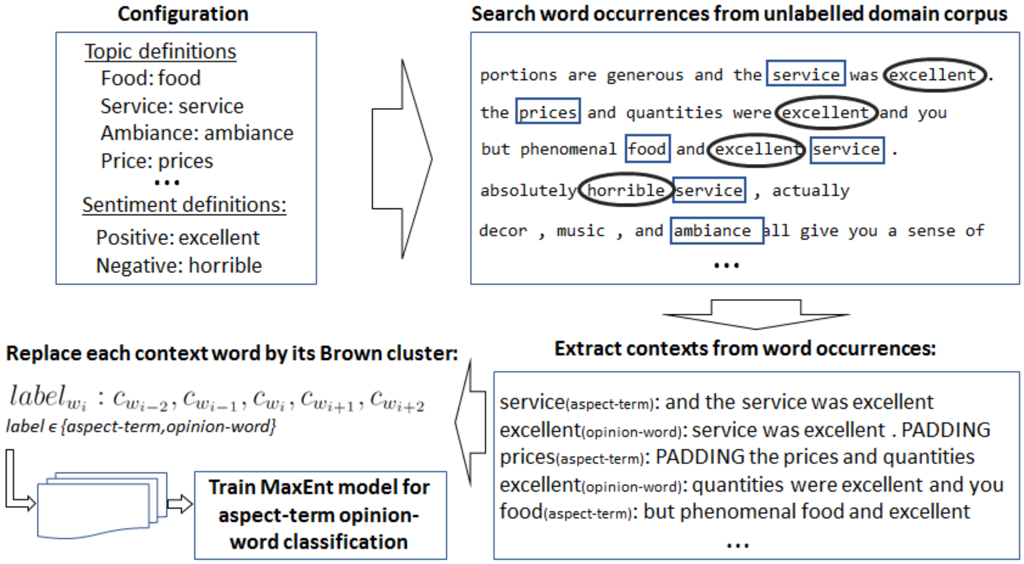

Configuration
Topic definitions
Food: food
Service: service
Ambiance: ambiance
Price: prices
...
Sentiment definitions:
Positive: excellent
Negative: horrible
Search word occurrences from unlabelled domain corpus
portions are generous and the service was excellent .
the prices and quantities were excellent and you
but phenomenal food and excellent service .
absolutely horrible service , actually
decor , music , and ambiance all give you a sense of
...
Extract contexts from word occurrences:
service(aspect-term): and the service was excellent
excellent(opinion-word): service was excellent . PADDING
prices(aspect-term): PADDING the prices and quantities
excellent(opinion-word): quantities were excellent and you
food(aspect-term): but phenomenal food and excellent
...
Replace each context word by its Brown cluster:
$label_{w_i} : c_{w_{i-2}}, c_{w_{i-1}}, c_{w_i}, c_{w_{i+1}}, c_{w_{i+2}}$
label ϵ {aspect-term,opinion-word}
Train MaxEnt model for aspect-term opinion-word classification

Figure 3: Process to obtain the MaxEnt model for aspect-term/opinion-word separation.

uses Brown clusters (Brown et al., 1992) to model examples of aspect-terms and opinion-words and train a MaxEnt-based classification model. Brown clusters have been used as unsupervised features with good results in supervised Part-of-Speech tagging (Turian et al., 2010) and Named Entity Recognition (Agerri and Rigau, 2016). Brown clusters are computed[4] from the domain unlabelled corpus with no additional supervision, and are used as the features for the two words context window, [-2,+2], of each training example. The training instances are obtained leveraging the occurrences of the initial configuration with aspects and polarity seed words, assuming that domain aspect seed words are aspect-terms and polarity-words are opinion-words.

Figure 3 describes the process to obtain the classification model. First domain aspect seed words and polarity seed words are used as gold aspect-terms and gold opinion-words respectively. Then the occurrences of these words are bootstrapped from the domain corpus and they are modelled according to their context window. Next, context words are replaced by their corresponding Brown cluster to build each training instance. Finally, a MaxEnt model

[4] We use the Brown clustering implementation at `https://github.com/koendeschacht/brown-cluster`

is trained using these generated training instances.

We have experimented with a different number of Brown clusters (100, 200, 500, 1000 and 2000) but the impact of this parameter was negligible for this purpose. The reported results have been obtained using 200 clusters.

A drawback of this approach is that every word in the vocabulary will be classified as aspect-term or as opinion-word. There are words that do not belong to any of these categories. It would be interesting to have a third class (e.g. *"other"*), but it would require labelling training instances for that additional class, introducing a manual supervision that we want to keep to a minimum. We assume that the words that are not clearly aspect-terms or opinion-words will be spread across both classes, losing relevance during the topic modelling process.

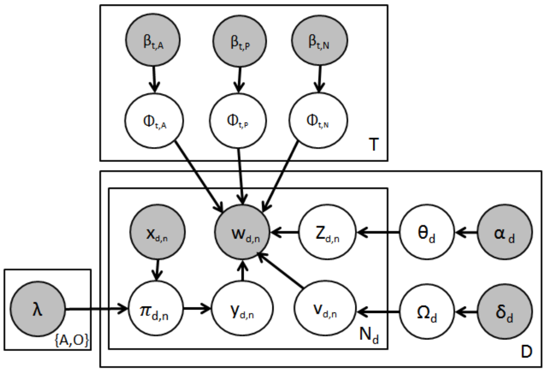
3.3. Combining everything in a topic model

The core of the system consists of an LDA-based topic model, extended to include the aspect-term/opinion-word separation and the positive/negative separation for each topic. In addition, the aspect-term/opinion-word separation is guided by a pre-trained classifier as explained at section 3.2, while the topic and polarity modelling are guided by biasing certain hyper-parameters according to the given topic configuration.

Figure 4 shows the proposed model in plate notation and the generative story modelled by the algorithm.

The generative hypothesis described by the model is the following. For each document d a distribution of topics, θ_d, is sampled from a Dirichlet distribution with parameter α_d, which is a vector with asymmetric topic priors for that document. Note that in this context each *document* corresponds to individual sentences instead of full texts. Then for each word n in document d a topic value is drawn: $z_{d,n} \sim Multi(\theta_d)$, $z \in \{1..T\}$. Then a aspect-term/opinion switch variable is sampled: $y_{d,n} \sim Bernoulli(\pi_{d,n})$, $y \in \{A, O\}$. Depending on $y_{d,n}$, the word $w_{d,n}$ is emitted from the topic aspect terms distribution ($\phi_{z_{d,n},A}$) or else, a polarity value $v_{d,n}$ is sampled from Ω_d to choose if the word has to be drawn from $\phi_{z_{d,n},P}$ or $\phi_{z_{d,n},N}$ (positive and negative words respectively).

The model guides the topic and polarity modelling towards the desired values by biasing the hyper-parameters that govern the Dirichlet distributions from which the topics and words are sampled. In a standard LDA setting those hyper-parameters (commonly named α and β) are symmetric



```
For each topic $t \in \{1..T\}$:
    sample $\phi_t^A \sim \text{Dirichlet}(\beta_t^A)$
    sample $\phi_t^P \sim \text{Dirichlet}(\beta_t^P)$
    sample $\phi_t^N \sim \text{Dirichlet}(\beta_t^N)$
For each document $d \in \{d_1..d_M\}$:
    sample $\theta_d \sim \text{Dirichlet}(\alpha_d)$
    sample $\Omega_d \sim \text{Dirichlet}(\delta_d)$
    For each word $w \in \{w_{d,1}..w_{d,N}\}$:
        sample $\pi_{d,n} \sim \text{MaxEnt}(\lambda, x_{w_{d,n}})$
        draw $z_{d,n} \sim \text{Multinomial}(\theta_d)$
        draw $y_{d,n} \sim \text{Bernoulli}(\pi_{d,n})$
        if $y_{d,n} = A$:
            sample $w_{d,n} \sim \text{Multinomial}(\phi^A_{z_{d,n}})$
        else if $y_{d,n} = O$:
            sample $v_{d,n} \sim \text{Bernoulli}(\Omega_d)$
            if $v_{d,n} = P$:
                sample $w_{d,n} \sim \text{Multinomial}(\phi^P_{z_{d,n}})$
            else if $v_{d,n} = N$:
                sample $w_{d,n} \sim \text{Multinomial}(\phi^N_{z_{d,n}})$
```

Figure 4: Proposed model in plate notation and its generative process algorithm.

because no apriori information about the topic and word distributions is assumed. In our model, these hyper-parameters are biased using a similarity calculation among the words of the domain corpus and the topic seed words of the initial configuration. This similarity measure is based on the cosine distance between the dense vector representation of the topic defining seeds and each word of the vocabulary. Such a dense vector representation of the words over a particular vocabulary, commonly referred as word embeddings, could be obtained using any distributional semantics approach, but in this work we stick to the well-known word2vec (Mikolov et al., 2013a). Word embeddings are a very popular way of representing words as the input for a variety of machine learning techniques and are known for encoding interesting syntactic and semantic properties (Mikolov et al., 2013b). In this case, we exploit the semantic similarity among words that can be calculated using the cosine distance of the resulting word vectors. The similarity, sim, is the value between a word and a set of words (e.g. some topic defining seeds), and it is calculated using 1.

$$\text{sim}(w,t) = \underset{v\in t}{\text{argmax}}\,\text{sim}(w,v) \tag{1}$$

Where w is any word found in the domain corpus, v is any of the seed words chosen to define topic t, and sim stands for the cosine distance between two word vectors.

The α hyper-parameters control the topic probability distribution for each document as in the original LDA. But instead of having a single symmetric α value, each document has a biased α for each topic, based on semantic word similarity, as described in 2.

$$\alpha_{d,t} = \frac{\sum_{i}^{N_d} \operatorname{sim}(w_{d,i}, t)}{\sum_{t'}^{T} \sum_{i}^{N_d} \operatorname{sim}(w_{d,i}, t')} * \alpha_{base} \tag{2}$$

On the other hand, the β hyper-parameters, which control the distribution of words for each topic, are calculated in a similar way, as shown in 3 and 4.

$$\beta_{t,w} = \operatorname{sim}(w, t) * \beta_{base} \tag{3}$$

$$\beta_{q,w} = \operatorname{sim}(w, q) * \beta_{base} \quad q \in \{P, N\} \tag{4}$$

Finally, the δ hyper-parameters control the polarity distribution for each document, and they are calculated for each document as shown in 5.

$$\delta_{d,q} = \frac{\sum_{i}^{N_d} \operatorname{sim}(w_{d,i}, q)}{\sum_{q' \epsilon \{P,N\}} \sum_{i}^{N_d} \operatorname{sim}(w_{d,i}, q')} * \delta_{base} \tag{5}$$

In the formulas $w_{d,i}$ is the *i-th* word of the document d, N_d is the number of words in that document, t is a topic from the set of defined topics T. Similarly q is a pre-defined polarity words set, P for positives and N for negatives (in our experiments P only contains *excellent* and N only contains *horrible* for English, or their equivalents for other languages).

α_{base}, β_{base} and δ_{base} are configurable hyper-parameters, analogous to the symmetric α and β in the original LDA model.

In addition to the bias of these hyper-parameters, the distribution π that governs each binary aspect-term/opinion-word switching variable, y, is set from the pre-trained aspect-term/opinion-word classifier probabilities applied to each word and its context features as described in section 3.2.

The posterior inference of the model is obtained via Gibbs sampling (Griffiths and Steyvers, 2004). Let $w_{d,n}$ be the n-th word of the d-th document, given the assignment of all other variables, its topic assignment $z_{d,n}$ is sampled using (6). Analogously, the aspect-term/opinion-word assignment $y_{d,n}$ and the polarity of the opinion-words, $v_{d,n}$ are sampled using (7) and (8) respectively.

$$p(z_{d,n} = t | z_{-d,n}, y_{-d,n}, v_{-d,n}, \cdot) \propto \frac{n^{t,A}_{w_{d,n}} + \beta^{t,A}_{w_{d,n}}}{\sum\limits_{v}^{V} n^{t,A}_{v} + \beta^{t,A}_{v}} \times \frac{n^{t,P}_{w_{d,n}} + \beta^{t,P}_{w_{d,n}}}{\sum\limits_{v}^{V} n^{t,P}_{v} + \beta^{t,P}_{v}} \times \frac{n^{t,N}_{w_{d,n}} + \beta^{t,N}_{w_{d,n}}}{\sum\limits_{v}^{V} n^{t,N}_{v} + \beta^{t,N}_{v}} \times (n_{d,t} + \alpha_{d,t}) \tag{6}$$

$$p(y_{d,n} = u | z_{d,n} = t, \cdot) \propto \frac{n^{t,u}_{w_{d,n}} + \beta^{t,u}_{w_{d,n}}}{\sum\limits_{v}^{V} n^{t,u}_{v} + \beta^{t,u}_{v}} \times \frac{exp(\lambda_u \times x_{d,n})}{\sum_{u' \in \{A,O\}} exp(\lambda_{u'} * x_{d,n})} \tag{7}$$

$$p(v_{d,n} = q | z_{d,n} = t, \cdot) \propto \frac{n^{t,q}_{w_{d,n}} + \beta^{t,q}_{w_{d,n}}}{\sum\limits_{v'}^{V} n^{t,q}_{v'_{d,n}} + \beta^{t,q}_{v'_{d,n}}} \times (n_{d,q} + \delta_{d,q}) \tag{8}$$

In these formulas, $n^{t,u}_{w_{d,n}}$ is the number of times the vocabulary term corresponding to $w_{d,n}$ has been assigned to topic t and word-type $u \in \{A, O\}$ (i.e. Aspect-terms or Opinion-words), $n_{d,t}$ is the number of words in the document d assigned to topic t, λ_u are the pre-trained aspect-term/opinion-word classifier model weights for word-type u and $x_{d,n}$ is the feature vector for $w_{d,n}$, composed by the Brown clusters of the context words. Analogously, $n^{t,q}_{w_{d,n}}$ is the number of times $w_{d,n}$ has been assigned to topic t and polarity $q \in \{P, N\}$ and $n_{d,q}$ is the number of words in the document d assigned to polarity q.

4. Evaluation

We evaluate W2VLDA for the three different subtasks that it performs: topic (aspect) classification, sentiment classification, and aspect-term/opinion-word separation. First, we compare W2VLDA with other LDA-based methods. Then, we also evaluate W2VLDA in a multilingual ABSA dataset comparing its performance classifying topics (aspects) and sentiment with some supervised machine learning approaches trained on labelled data.

Language:Domain	Domain topic	Aspect-terms	Positive words	Negative words
English: restaurant reviews	Food	chicken, beef, pork, tuna, egg, onions, shrimp, curry	moist, goat, smoked, seared, roasted, red, crispy, tender	undercooked, dry, drenched, overcooked, soggy, chewy
	Service	staff, workers, employees, chefs, hostess, manager, owner	helpful, polite, knowledgeable, efficient, prompt, attentive	inattentive, rude, unfriendly, wearing, making, packed
	Ambiance	lighting, wall, interior, vibe, concept, ceilings, setting, decor	modern, beautiful, chic, nice, trendy, cozy, elegant, cool	bad, loud, uninspired, expensive, big, noisy, dark, cramped
English: electronic devices reviews	Warranty	warranty, support, repair, service, answer, center, policy	worked, lucky, owned, big, exchange, extended, longer	called, contact, broken, faulty, defective, expired, worthless
	Design	plastic, wheel, style, handle, pocket, design, exterior, wheels	adjustable, clean, good, versatile, attractive, lightweight, stylish	ugly, odd, awkward, tight, felt, weird, cute, stupid, flimsy
	Price	money, store, item, bucks, price, regret, deal, gift	paying, reasonable, penny, worth, delivered, stars, inexpensive,	disappointed, paid, cheaper, skeptical, pricey, overpriced

Table 2: Resulting topic words distributions for English in two different domains. The topics are automatically split into three different word distributions: topic aspect terms, topic positive words and topic negative words.

We show results for several datasets, demonstrating how the system works for different languages and domains just by changing the topic configuration, composed of a single seed word per each desired topic, language and domain.

For instance, table 2 shows some of the resulting words for several domains (restaurants and electronic devices), topics (food, service, ambience for restaurants, and warranty, design and price for electronic devices) for English customer reviews, including the automatic separation of aspect-terms from positive and negative words per topic. Table 3 shows the equivalent information for restaurants and hotel reviews in Spanish and French.

Likewise, table 4 shows examples of sentences classified under different topics (food, service, ambience for restaurants, and staff, ambience and location for hotels) for several domains (restaurants and hotels) and languages (Spanish and French).

Language:Domain	Domain topic	Aspect-terms	Positive words	Negative words
Spanish: restaurant reviews	Food	crema, tartar, ensaladas, sopa, brasa, patatas, salsas, alcachofas	caprese, sublime, destacar, casera, tierna, trufada, ahumada	aguada, mojar, congeladas, quemadas, fritos, rancias, reseco
	Service	camareros, camarero, maitre, dueño, encargado, metre	eficiente, eficaz, atentos, correcta, cercano, diligente	lento, pésimo, desagradable, prepotente, maleducado
	Ambiance	toques, atmósfera, material, mobiliario, bancos, modernidad	tranquilo, relajado, cálido, buena, amplio, luminoso, precioso,	cutre, insoportable, pequeño, tanta, oscuro, poca, normalita
French: hotel reviews	Food	nourriture, sauce, produits, pâte, bouffe, saveur, risotto	raisonnable, michelin, excellents, merveilleuse, veritable, superbe	correcte, cuit, idem, passable, excessif, moleculaire, difficile
	Staff	personnel, ècoute, staff, gentillesse, concierge, membres	sympathique, attentionné, efficace, compètent, professionnel	dèplorable, antipathique, dèbordè, distant, constamment
	Ambiance	impression, couloirs, odeurs, personnages, hiver, escaliers	vieillissant, grand, rènovè, boone, typiquement, cosy, agrèablement	froide, vètuste, forte, incendie, bruyants, inexistante, complète

Table 3: Resulting topic words distributions for two Spanish and French and for different domains. The topics are automatically split into three different word distributions: topic aspect terms, topic positive words and topic negative words.

4.1. Resources and experimental setting

In order to evaluate W2VLDA, we use the following resources. For topic classification we use the dataset from (Ganu et al., 2009) which contains restaurant reviews labelled with domain-related categories (e.g. food, staff, ambience) for English. For sentiment classification, we use the Laptops and DIGITAL-SLR dataset (Jo and Oh, 2011), consisting of English reviews of electronic products with their corresponding 5-star rating.

Additional multilingual experiments have been performed using the SemEval-2016 task 5 datasets (Pontiki et al., 2016). In particular, the restaurant reviews datasets which are labelled with domain-related categories and polarity for six languages.

In order to compute the topic model and the word embeddings, we have automatically gathered additional customer reviews about restaurants from some popular customer review websites. These unlabelled domain corpora consist of a few thousand restaurant reviews in English, Spanish, French and

Lang: Domain	Domain Topic	Examples of sentences with high posterior probability for different topics
English: restaurant reviews	Food	Appetizer was grilled pizza dough topped with fig jam, prosciutto, arugula, cherry tomatoes... Four of us enjoyed sizziling rice seafood soup, the most savory garlic string beans.
	Service	Seated promptly, waiter arrived at 6:10 brought us our drink order 6:15. Bartenders are friendly and quick to be helpful
	Ambiance	The atmosphere as a restaurant though is very nice: cute decor, quieter, and dim lighting... The ambiance of the restaurant is very nice, the decor and lighting set a great atmosphere
Spanish: restaurant reviews	Food	Probamos las croquetas melosas de jamón, milhoja de tomate y mozzarella con salsa de miel. Paté de perdiz, tartar de bonito, steak tartar, paté de cabracho, brocheta de pollo y postres
	Service	El servicio a los clientes deja bastante que desear El trato es magnífico, camareros muy simpáticos y amables, un trato educado y exquisito
	Ambiance	Cena agradable en un lugar de ambiente tranquilo, cosmopolita, con buena música El local es feo decorado como un bar de carretera en EEUU o un autobús
French: hotel reviews	Staff	Service de qualitè et personnel extrêmement agreable, aux petits soins, disponible et serviable! Le personnel est rèactif, serviable, disponible, toujours prêt à rèpondre aux attentes des clients.
	Ambiance	L'hotel est une attraction en soi, il y a un adventure park a l'interieur, on se croirait a disneyland. Le bâtiment a un certain charme, certaines tapisseries sont dèfraîchies, se sent londonien
	Location	A 5 minutes à pied de buckingham palace et saint james park , 10 à 15 minutes de big ben. Hotel à 15 min de la gare à pied, 15 min d'oxford street, à 40 min du centre ville à pied.

Table 4: Some examples of sentences with the highest posterior probability for several topics, domains and languages.

Dutch.

We use word2vec to compute the word embeddings that are used for the word similarity calculation. In particular, we use the Apache Spark MLlib [5] implementation with default parameters to compute the domain-based word embeddings.

Table 1 shows the topic definition used in the experiments for the domain of restaurants, just one word per topic. Unless stated otherwise, the polarity seeds for every domain are *excellent* and *horrible* or their equivalents in other languages.

The values for α_{base}, β_{base} and δ_{base} mentioned in 3.3, which play a similar role to α and β in the original LDA, are set to the values commonly recommended in the literature (Griffiths and Steyvers, 2004): 50/T for α_{base} and δ_{base} being T the number of topics, and 0.01 for β_{base}. The topic modelling process runs for 500 iterations in every experiment with a burn-in period of 100 iterations and a sampling lag of 10 iterations.

4.2. Comparison with other LDA based approaches

First, we evaluate W2VLDA in a topic classification setting using the restaurant reviews dataset from (Ganu et al., 2009). This dataset contains few thousand reviews from restaurants, classified into several categories but the authors report results only for the three main categories: *food*, *ambience* and *staff*. We compare W2VLDA against the results reported in (Zhao et al., 2010) for two LDA-based approaches, LocLDA (Brody and Elhadad, 2010) and ME-LDA (Zhao et al., 2010).

LocLDA and ME-LDA are LDA-based approaches, and thus, unsupervised. But the results reported in the experiment involved some supervision as described in Zhao et al. (2010). First, the authors computed a topic model of 14 topics. Then the authors examine each topic and manually set a label according to their judgment. W2VLDA provides already named topics at the end of the process, so no manual topic inspection and labelling are required. In order to assign a topic label to a particular sentence, we use the resulting topic distribution for that sentence (θ_d) selecting the topic with highest posterior probability.

Table 5 shows the results of the experiment and the comparison with the other systems. Despite not requiring human intervention to relabel the obtained topics unlike the other two systems, W2VLDA obtains slightly better overall results.

[5] `http://spark.apache.org/mllib/`

Method	Topics											
	Staff			Food			Ambiance			Overall		
	Prec.	Rec.	F-1	Prec.	Rec.	F-1	Prec.	Rec.	F1	Prec.	Rec.	F1
LocLDA	**0.80**	0.59	0.68	0.90	0.65	0.75	0.60	0.68	0.64	0.77	0.64	0.69
ME-LDA	0.78	0.54	0.64	0.87	**0.79**	**0.83**	**0.77**	0.56	**0.65**	**0.81**	0.63	0.70
W2VLDA	0.61	**0.86**	**0.71**	**0.96**	0.69	0.81	0.55	**0.75**	0.63	0.70	**0.77**	**0.72**

Table 5: Comparison against other LDA based approaches on restaurants domain

We also evaluate the ability of W2VLDA to assign correct polarities to customer reviews. We use the estimated polarity distribution of a sentence (Ω_d) to assign to a review the polarity with the highest probability. We compare our polarity classification results with respect to those from JST (Lin et al., 2011), ASUM (Jo and Oh, 2011) and HASM (Kim et al., 2013). The evaluation runs over the laptops and digital SLRs subset obtained from the Amazon Electronics dataset[6]. As explained at (Kim et al., 2013) two datasets are used, a *small* dataset containing 1000 reviews with 1 star rating (strong negative) and 1000 5 stars (strong positive), and a *large* dataset with additional 1000 reviews of 2 stars (negative) as well as 1000 reviews of 4 stars (positive). The baseline consists of a simple polarity seed word count, using the polarity seed words from (Turney and Littman, 2003), assigning to the sentence the polarity with the greatest proportion. As stated in previous sections, W2VLDA uses just a single polarity seed for each sentiment polarity, *excellent* and *horrible* respectively.

Figure 5 shows the result of this comparison. W2VLDA obtains comparable results for the small dataset and better results for the big dataset despite using only a single seed word to define each polarity.

4.3. Multilingual evaluation on SemEval2016 dataset

We use the SemEval 2016 task 5 datasets (Pontiki et al., 2016) in order to perform a multilingual evaluation of W2VLDA. SemEval 2016 datasets consist of restaurant reviews in several languages. The reviews are split by sentence and labelled with the explicit aspect term mentions, the coarse-grained category they belong to, and the polarity for that category.

SemEval 2016 restaurants datasets are annotated for six coarse-grained categories: food, service, ambience, drinks, location, and restaurant. The last

[6] Available at `http://uilab.kaist.ac.kr/research/WSDM11/`

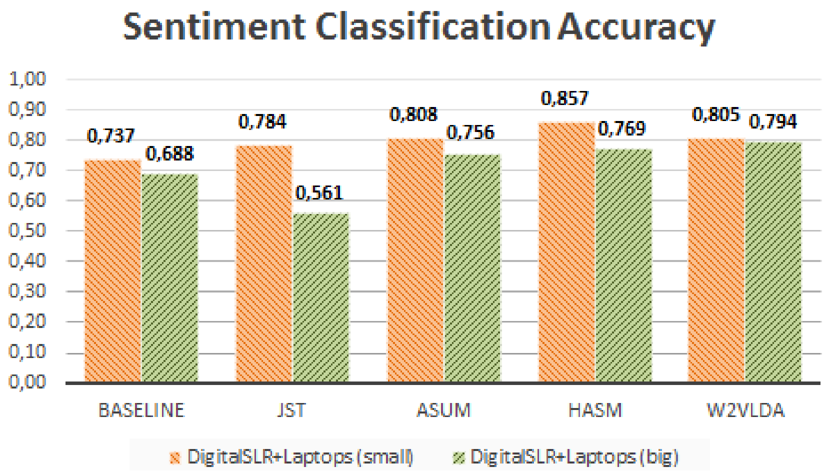


Figure 5: Sentiment classification accuracy comparison with other LDA based approaches in a electronic devices reviews dataset

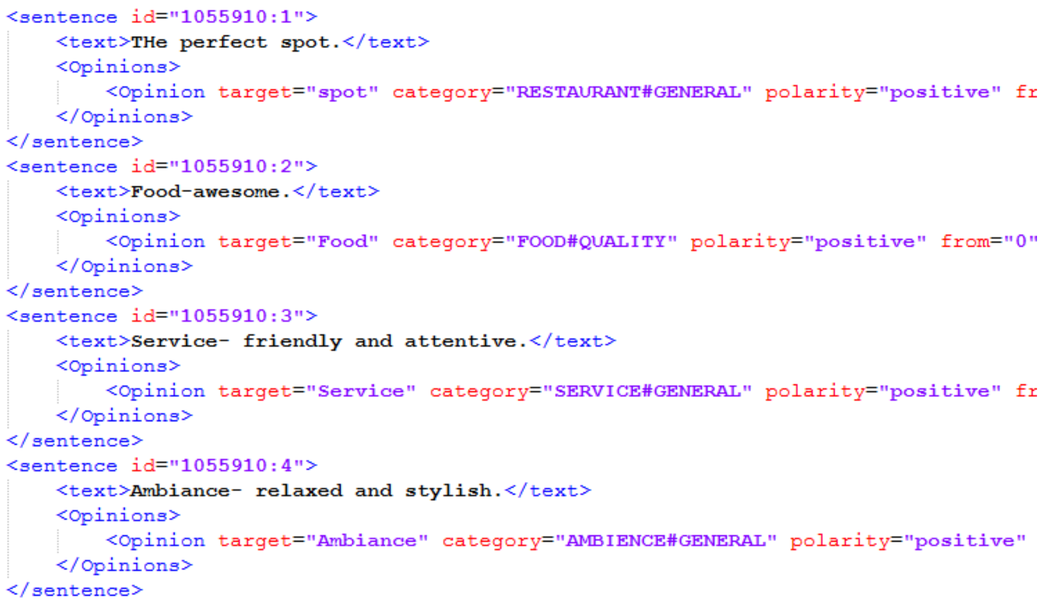
	EN	ES	FR	NL
Food	486	364	370	374
Service	328	233	290	350
Ambience	110	145	98	117
Total	924	742	758	841

Table 6: SemEval 2016 dataset category distribution after filtering unwanted categories and sentences with more than one annotation.

category, *restaurant* acts as a miscellaneous category that is used when the sentence does not refer to any other specific category but to the restaurant as a whole. Such an abstract concept cannot be represented by a seed word, so we omit this category from the evaluation. To avoid ambiguities and simplify the classification of a sentence, we only keep sentences with a single category label. Finally, since the categories *drinks* and *location* have very little representation in the datasets (below the 5% of the instances), we keep only the three main categories: *food*, *service* and *ambience*.

Table 6 and table 7 show the distribution of categories and polarities respectively for the resulting datasets, for four languages: English, Spanish, French and Dutch.

Since W2VLDA is a topic modelling, it needs a reasonable amount of domain documents to build the statistical model. To cope with this require-



```
<sentence id="1055910:1">
    <text>THe perfect spot.</text>
    <Opinions>
        <Opinion target="spot" category="RESTAURANT#GENERAL" polarity="positive" fr
    </Opinions>
</sentence>
<sentence id="1055910:2">
    <text>Food-awesome.</text>
    <Opinions>
        <Opinion target="Food" category="FOOD#QUALITY" polarity="positive" from="0"
    </Opinions>
</sentence>
<sentence id="1055910:3">
    <text>Service- friendly and attentive.</text>
    <Opinions>
        <Opinion target="Service" category="SERVICE#GENERAL" polarity="positive" fr
    </Opinions>
</sentence>
<sentence id="1055910:4">
    <text>Ambiance- relaxed and stylish.</text>
    <Opinions>
        <Opinion target="Ambiance" category="AMBIENCE#GENERAL" polarity="positive"
    </Opinions>
</sentence>
```

Figure 6: SemEval 2016 task 5 restaurants dataset example (for English).

	EN	**ES**	**FR**	**NL**
Positive	551	417	300	405
Negative	326	273	413	369
Total	877	690	713	774

Table 7: SemEval 2016 dataset polarity distribution after filtering unwanted categories and sentences with more than one annotation.

ment, we have implemented a script to automatically extract restaurant reviews of the required languages from an online customer reviews website. Due to copyright permissions, we cannot share these reviews, but table 8 shows the number of reviews used to feed the algorithm. The polarity mentioned on the table is based on the number of the stars from the 5-star rating (as usual, 1-2 stars meaning negative and 4-5 starts meaning positive). As it can be observed in the table, for some languages the script has not found an equal number of positive and negative reviews. We tried to compensate this fact with oversampling, to pair the number of positive and negative reviews before running the algorithm. In this case we oversample negative examples for each language until they equal in number the positive ones (i.e. 10k). Note that in the case of Dutch this may lead to an excessive oversampling

Restaurant customer reviews downloaded from a website				
	EN	ES	FR	NL
Positives (4 or 5 stars)	10000	10000	10000	10000
Negatives (1 or 2 stars)	10000	8400	5500	830
Total reviews	20000	18400	15500	10830

Table 8: Downloaded reviews distribution per language and polarity (using 5-star rate). The automatic script could not find the same number of negative reviews for all the languages. We try to alleviate this problem oversampling negatives reviews.

due to the small number of available negatives examples. Also note that these polarities are just to get an insight of the polarity distribution of the datasets, but they are not used for any sort of supervised training.

The evaluation experiment is done as follows. For each language, we use the downloaded reviews to run the algorithm. It includes calculating the domain word embeddings, Brown clusters and the topic model estimation. Using the generated model for each language the topic and polarity distributions, θ and Ω, are estimated for each of the sentences of the evaluation set. The topic with the highest probability in the estimated topic distribution for that sentence is assigned as the category label (i.e. domain aspect). Analogously, the polarity with the highest probability in the estimated polarity distribution for that sentence is assigned as the polarity label. The assigned category is compared to the gold category, and the accuracy (ratio of correctly labelled examples) is calculated. The same process is followed to calculate the polarity classification accuracy.

The obtained accuracy is compared to several baselines. First, two supervised baselines are used. One is a Naive-Bayes classifier (NB), trained using the labelled sentences. The sentences are transformed to bag-of-words vectors with a vocabulary size of 80k words and normalised using tf-idf weights. The other supervised baseline is a Multilayer Perceptron algorithm (MLP), with two hidden layers, and the same tf-idf vector as input. Another baseline is the majority baseline, that shows the accuracy that can be obtained in the case of choosing the most frequent class. This is only to ensure that the datasets are not excessively unbalanced and the algorithms are really learning relevant information. Finally, the last baseline (W2VLDA_NO) is the same W2VLDA but replacing the word-embeddings similarity mechanism to bias the topic modelling hyper-priors. Instead of using the word-embedding

Domain aspect classification				
	EN	**ES**	**FR**	**NL**
NB	0.492	0.497	0.472	0.457
MLP	0.554	0.564	0.496	0.464
Majority baseline	0.333	0.333	0.333	0.333
W2VLDA_NO	0.313	0.374	0.356	0.315
W2VLDA	**0.781**	**0.633**	**0.586**	**0.473**

Table 9: Domain aspect classification results. NB and MLP are the supervised baselines, NaiveBayes and MultiLayer perceptron respectively. Majority baseline shows which would be the result of simply choosing the most frequent class. W2VLDA_NO is the proposed approach without word embeddings. W2VLDA is the proposed approach.

similarity to calculate a bias for every word, only the configured seed words receive a strong bias for their corresponding topic or polarity.

Table 9 shows the evaluation results for the domain aspects classification (food, service, ambience). Since the evaluation datasets are not completely balanced for each of the domain aspects (see table 6), we run the evaluation on several balanced subsets created by random sampling the base datasets for each language. Each balanced subset contains 100 sentences from each domain aspect. We do this five times generating five different subsets, and we use these subsets to evaluate the baselines and W2VLDA. The results on each individual subset are obtained using the average accuracy applying a 10-fold cross validation. We calculate the average and standard deviation of the results on each subset to perform a t-test of statistical significance. W2VLDA outperforms the baselines with a 95% of confidence for all the languages except for Dutch, which despite obtaining better results than the baselines it only achieves a 80% on confidence in the statistical significance test.

Table 10 shows the evaluation results for the polarity classification (positive and negative). The calculation of the results and the statistical significance tests have been performed in the same way than for the domain aspect classification. Again, W2VLDA outperforms the baselines with a 95% on confidence in the statistical test, except for Dutch. A possible reason for this is that the oversampling performed for the unlabelled Dutch reviews for the topic modelling was excessive, or the data contained in it was less representative than for other languages (see table 8). Studying which are the lower

Sentiment polarity classification				
	EN	**ES**	**FR**	**NL**
NB	0.672	0.577	0.587	0.563
MLP	0.711	0.602	0.583	0.577
Majority baseline	0.500	0.500	0.500	0.500
W2VLDA_NO	0.531	0.552	0.534	0.523
W2VLDA	**0.773**	**0.723**	**0.628**	**0.623**

Table 10: Sentiment polarity classification results. NB and MLP are the supervised baselines, NaiveBayes and MultiLayer perceptron respectively. Majority baseline shows which would be the result of simply choosing the most frequent class. W2VLDA_NO is the proposed approach without word embeddings. W2VLDA is the proposed approach.

bounds of the required amount of data would be an interesting problem that we let for future research.

4.4. Assessing the seed words impact

Since the proposed approach heavily relies on the seed words (i.e. seeds words are the only source of supervision to guide the algorithm to the desired goal), it is interesting to evaluate the impact of different seed words and their combination.

We perform some experiments for English using the SemEval 2016 restaurant reviews dataset and several combinations of seed words for the target domain aspects and sentiment polarities. In the first experiment group, for each run we only change the seed words that define the domain aspects. The polarity seeds remain the same.

We use three different seed words for each domain aspect, in particular: *food*, *chicken* and *burger* for domain aspect *FOOD*; *service*, *staff* and *waiter* for domain aspect *SERVICE*; and *ambience*, *atmosphere* and *décor* for domain aspect *AMBIENCE*. We try different permutations and combinations of the seed words, including pairs of seed words for each domain aspect, and finally also the combination of the three seed words together. Table 11 show the results for this experiment. The results show that the accuracy is stable across all the combinations regardless of the chosen seed words. As expected, some combinations perform better than others but overall the average if high and the standard deviation is below 5% of the accuracy. The best result is obtained using all the seed words at the same time. This last fact is not

Aspects:{FOOD},{SERVICE},{AMBIENCE}	**Aspects acc.**	**Polarity acc**
{food},{service},{ambience}	0,709	0,738
{chicken},{staff},{atmosphere}	0,653	0,729
{burger},{waiter},{décor}	0,662	0,731
{food,chicken},{service,staff},{ambience,atmosphere}	0,735	**0,742**
{food,burger},{service,waiter},{ambience,décor}	0,724	0,721
{chicken,burger},{staff,waiter},{atmosphere,décor}	0,673	0,725
All the 3 seeds for every aspect	**0,761**	0,722
Average	0,702	0,730
Standard deviation	0,041	0,008

Table 11: Impact of different seed words combination for the domain aspect classification.

surprising, since with more seeds the semantic coverage to guide the algorithm to the desired domain aspects is increased (as long as the seed words are semantically coherent with the domain aspect they are defining).

Another fact that can be observed in the table is that domain aspect seed words do not affect the polarity results, as it would be expected. The polarity results show minor variations among the experiments, but the standard deviation is only a 0.8% of the accuracy.

Analogously to the domain aspect seed words, we have performed some experiments with the polarity words. We have tested several combinations of seeds with opposed polarity: excellent - horrible, awesome - awful, etc. Table 12 show the results. Even with seed words of less extreme polarity, like good - bad, the results are quite stable. We also test combining more than a single word for each polarity, and as the results table shows, combining the three seed words for each polarity achieves the best result. The standard deviation for all the experiment runs is just a 1.2% of the accuracy. Similarly to what was observed for the domain aspects, the polarity seed words do not seem to affect the domain aspect classification accuracy, with only a 1.2% on standard deviation for all the runs.

Finally, in order to perform a sanity check to evaluate if the sentiment polarity classification is really depending on the correct selection of the polarity seed words, we perform two more runs using misleading words as polarity seeds. In particular, we use *cat* and *waitress* for positives and *dog* and *waiter* for negatives. The use of these words as polarity seeds is obviously incorrect,

Polarity:{POSITIVE},{NEGATIVE}	**Aspects acc.**	**Polarity acc.**
{excellent},{horrible}	0,701	0,724
{terrific},{terrible}	0,712	0,736
{awesome},{awful}	0,691	0,745
{nice},{poor}	0,704	0,735
{good},{bad}	0,684	0,712
{affordable},{expensive}	**0,716**	0,729
{excellent,terrific},{horrible,terrible}	0,683	0,726
{excellent,terrific,awesome},{horrible,terrible,awful}	0,692	**0,747**
Average	0,698	0,732
Standard deviation	0,012	0,012

Table 12: Impact of different polarity seeds words for the sentiment polarity classification.

Polarity:{POSITIVE},{NEGATIVE}	**Aspects acc.**	**Polarity acc.**
{cat},{dog}	0,642	0,447
{waitress},{waiter}	0,635	0,419

Table 13: Results using misleading words as polarity seeds to check to which extent the sentiment polarity classification depends on the validity of the chosen polarity seeds.

and what we want to check is if using such meaningless words (for polarity) leads to bad polarity classification results. Table 13 shows the results for this experiment, confirming that the election of representative polarity seed words is relevant to correctly guide the algorithm.

4.5. Aspect-term/Opinion-word separation evaluation

Finally we experiment with the aspect-term and opinion-word separation. As described in section 3.2, W2VLDA models the domain words into separated word distributions: aspect terms or opinion words.

In order to evaluate the accuracy of this words separation, we use Bing Liu's polarity lexicon for English (Hu and Liu, 2004). Since sentiment lexicons contain words bearing some specific sentiment, we treat the words contained in this lexicon as a ground-truth for opinion-words. In addition, we use the gold aspect-terms labelled in the SemEval 2016 dataset as a ground-truth for aspect-terms.

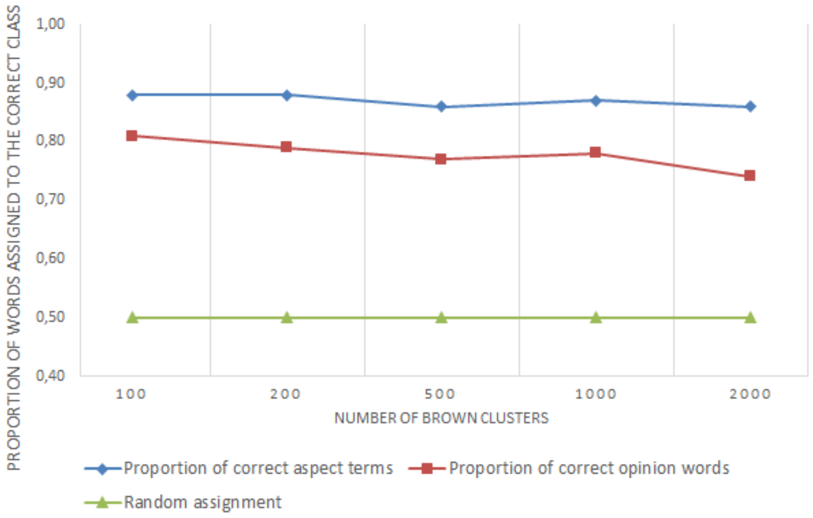


Figure 7: Result of aspect term and opinion word separation for English. Each point indicates the correct proportion (percentage) of aspect terms or opinion words that have been correctly classified. Random assignment is the random guess baseline.

The experiment now consists of running the W2VLDA again on the restaurant review dataset and counting how many times a word from the opinion words ground-truth is classified as an opinion word, and how many times each word from the aspect terms ground-truth is classified as an aspect term. Then the proportion of correct assignments is calculated. If the automatic aspect-term / opinion-word separation is correct, the proportion of opinion words and aspect terms correctly classified should be high.

We perform several experiments varying the number of Brown clusters involved in the process (see section 3.2) to evaluate if it has a noticeable impact on the word separation. Figure 7 shows the resulting proportions of correctly assigned aspect terms and opinion words for English. In general, the correct proportions are high compared to a random assignment, which indicates that the aspect-term/opinion-word separation performs correctly most of the times. Interestingly, aspect-terms are better distinguished than opinion-words.

5. Conclusions and future work

In this document, we have presented W2VLDA, a system that performs aspect and sentiment classification with almost no supervision and without the need of language or domain specific resources. In order to do that, the system combines different unsupervised approaches, like word embeddings or Latent Dirichlet Allocation (LDA), to bootstrap information from a domain corpus. 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. The resulting topics and polarities are directly paired with the aspect names selected by the user at the beginning, so the output can be used to perform Aspect Based Sentiment Analysis. In addition, the system tries to separate automatically aspect terms and opinion words, providing more clear information and insight to the resulting domain aspects vocabulary. We evaluate W2VLDA for aspect classification using customer reviews of several domains and compare it against other LDA-based approaches. We also evaluate its performance using a subset of the multilingual SemEval 2016 task 5 ABSA dataset. As future work, it would we interesting to include an automated way to deal with stop-words and other words that do not carry information for the ABSA task. A better-integrated handling of multi-word and negation expressions could also improve the results. On the other hand, the are more specialised word embeddings related to sentiment analysis (Rothe et al., 2016), and it would be interesting to study if different word embeddings bring improvements to the method keeping a minimal supervision.

Acknowledgements

This work was supported by Vicomtech-IK4 and by the project TUNER - TIN2015-65308-C5-1-R (MINECO/FEDER, UE).

References

References

Agerri, R. and Rigau, G. (2016). Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63 – 82.

Alam, M. H., Ryu, W. J., and Lee, S. K. (2016). Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews. *Information Sciences*, 339:206–223.

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., and Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246.

Bhatia, S., Lau, J. H., and Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. *arXiv preprint arXiv:1612.05340*.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., and Reynar, J. (2008). Building a sentiment summarizer for local service reviews. In *WWW workshop on NLP in the information explosion era*, volume 14, pages 339–348.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Brody, S. and Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (June):804–812.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Chen, T., Xu, R., He, Y., and Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, 72:221–230.

Chen, Z., Mukherjee, A., and Liu, B. (2014). Aspect extraction with automated prior knowledge learning. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 347–358.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Das, R., Zaheer, M., and Dyer, C. (2015). Gaussian LDA for Topic Models with Word Embeddings. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*, pages 795–804.

Ganu, G., Elhadad, N., and Marian, A. (2009). Beyond the stars: Improving rating predictions using review text content. In *WebDB*, volume 9, pages 1–6. Citeseer.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., and Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760.

Huang, S., Niu, Z., and Shi, C. (2014). Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems*, 56:191–200.

Jijkoun, V., de Rijke, M., and Weerkamp, W. (2010). Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 585–594. Association for Computational Linguistics.

Jo, Y. and Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.

Kim, S., Zhang, J., Chen, Z., Oh, A., and Liu, S. (2013). A Hierarchical Aspect-Sentiment Model for Online Reviews. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 526–533.

Lin, C., He, Y., Everson, R., and Rüger, S. (2011). Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, 24:1134–1145.

Lin, C., Road, N. P., and Ex, E. (2009). Joint Sentiment / Topic Model for Sentiment Analysis. *Cikm*, pages 375–384.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Liu, K., Xu, L., and Zhao, J. (2012). Opinion target extraction using word-based translation model. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (July):1346–1356.

Lu, B., Ott, M., Cardie, C., and Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 81–88.

Mcauliffe, J. D. and Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems*, pages 121–128.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, pages 1–12.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. *Proceedings of NAACL-HLT*, pages 746–751.

Mostafa, M. M. (2013). More than words: Social networks text mining for consumer brand sentiments. *Expert Systems with Applications*, 40(10):4241–4251.

Mukherjee, A. and Liu, B. (2012). Aspect extraction through semi-supervised modeling. *ACL '12 Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, (July):339–348.

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., and Eryiğit, G. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado*, pages 486–495.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland*, pages 27–35.

Popescu, A.-M. and Etzioni, O. (2007). Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28. Springer.

Qiang, J., Chen, P., Wang, T., and Wu, X. (2016). Topic Modeling over Short Texts by Incorporating Word Embeddings. *arXiv preprint arXiv: 1609.08496v1*, page 10.

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics.

Rao, D. and Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics.

Rothe, S., Ebert, S., and Schütze, H. (2016). Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). Learning Sentiment-Specific Word Embedding. *Acl*, pages 1555–1565.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Wu, Y., Zhang, Q., Huang, X., and Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 3, pages 1533–1541. Association for Computational Linguistics.

Zhao, W. X., Jiang, J., Yan, H., and Li, X. (2010). Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid. *Computational Linguistics*, 16(October):56–65.