Expert Systems With Applications 91 (2018) 127–137

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

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

ARTICLE INFO

Article history:
Received 19 May 2017
Revised 29 August 2017
Accepted 30 August 2017
Available online 1 September 2017

Keywords:
Opinion mining
Aspect Based Sentiment Analysis
Almost unsupervised
Multilingual
Multidomain

ABSTRACT

With the increase of online customer opinions in specialised websites and social networks, automatic systems to help organise and classify customer reviews by domain-specific aspect categories and sentiment polarity are more needed than ever. Supervised approaches for Aspect Based Sentiment Analysis achieve good results for the domain and language they are trained on, but manually labelling data to train supervised systems for all domains and languages is 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 step, performs aspect category classification, aspect-term and opinion-word separation and sentiment polarity classification for any given domain and language. We evaluate its domain aspect and sentiment classification performance in the multilingual SemEval 2016 task 5 (ABSA) dataset. We show competitive results for several domains (hotels, restaurants, electronic devices) and languages (English, Spanish, French and Dutch).

1. Introduction

During the last decade, the Web has become one of the most important sources for customers and providers to compare and evaluate products and services. The vast amount of content generated every day in countless websites and social media keeps growing and requires automated ways to handle and classify all these opinions. As a result, 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, such as customer reviews (Liu, 2012; Pang & Lee, 2008). Aspect Based Sentiment Analysis (ABSA) refers to the systems that determine the opinions or sentiments expressed on different features or aspects of the products and services under evaluation (e.g., battery or performance for a laptop). An ABSA system should be capable of classifying each opinion according to the aspect categories relevant for each domain in addition to classifying its sentiment polarity (usually positive, negative or neutral), as depicted in Fig. 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., 2016; Pontiki, Galanis, Papageorgiou, Manandhar, & Androutsopoulos, 2015; Pontiki et al., 2014). This is the case of deep-learning based systems, that provide very good performance but require a significant amount of labelled data for training (Araque, Corcuera-Platas, Sánchez-Rada, & Iglesias, 2017; Chen, Xu, He, & Wang, 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 (Jo & Oh, 2011; Kim, Zhang, Chen, Oh, & Liu, 2013; Lin, He, Everson, & Rüger, 2011). In addition, most of these works only report results for English.

In this work, we present W2VLDA, an almost unsupervised system for multidomain and multilingual ABSA, that works leveraging large quantities of unlabelled textual data and an initial configuration consisting of a minimal set of seed words. Fig. 2 shows an schema of W2VLDA. Imagine the following scenario. The owners of a famous restaurant want to monitor the opinion of their customers with respect to a set of domain aspects. In particular, they want to know the opinion about its food, service, price, ambience, location, etc. The input to W2VLDA is a corpus of customer reviews and an example word per each domain aspect they want to monitor (e.g., *chicken* for food, *service* for service, etc.). Additionally, W2VLDA also needs an example of a positive and a negative word (e.g., *excellent* and *horrible*) independent of the domain. With this input, W2VLDA produces two main outputs. First, a weighted list

* Corresponding author.
E-mail addresses: agarciap@vicomtech.org, agarcia175@gmail.com (A. García-Pablos), mcuadros@vicomtech.org (M. Cuadros), german.rigau@ehu.eus (G. Rigau).

http://dx.doi.org/10.1016/j.eswa.2017.08.049
0957-4174/

Fig. 1. An example of classical Sentiment Analysis vs. Aspect Based Sentiment Analysis.

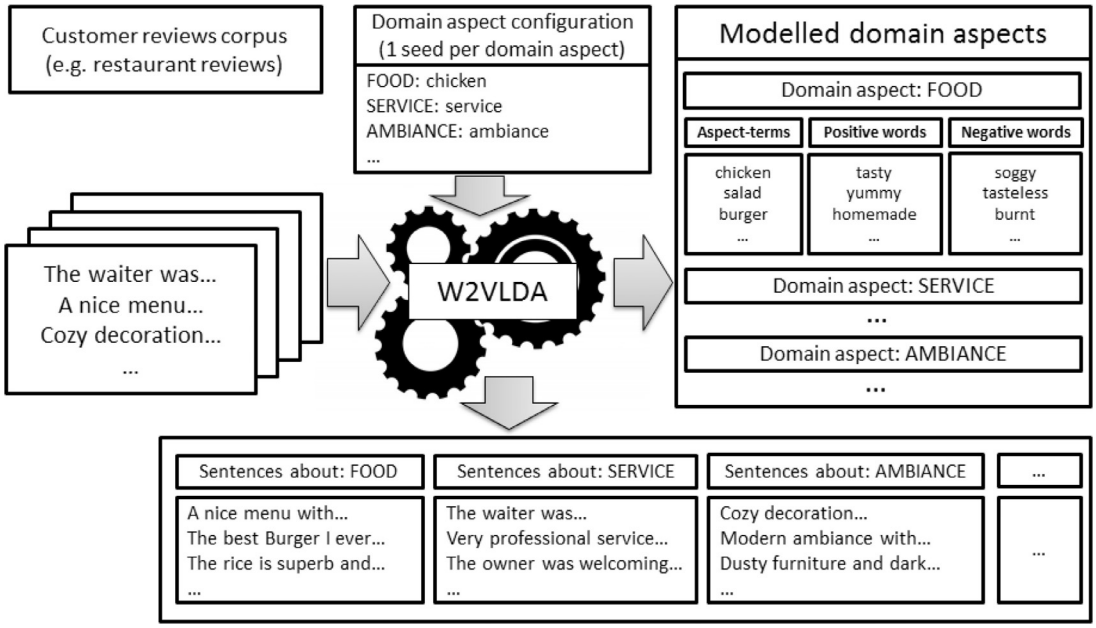


Fig. 2. A schema of W2VLDA. The input is an unlabelled corpus of a particular domain and its domain aspects specification. Domain aspects are split into three word distributions: aspect-terms, positive words and negative words. Sentences are modelled by domain aspect and polarity.

of aspect-terms (e.g., *chicken, salad, burger*, etc.), a weighted list of positive words (e.g., *tasty, yummy, homemade*, etc.) and a weighted list of negative words (e.g., *soggy, tasteless, burnt*, etc.) for each domain aspect (e.g., food). Second, W2VLDA also produces a weighted list of sentences for every 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 an unlabelled corpus of the target domain and language just by defining the desired domain aspects with a single seed word for each of them. We show results for different domains (restaurants, hotels, electronic devices) and languages (English, Spanish, French and Dutch) and compare its performance against other topic modelling based approaches on the SemEval2016 task 5 dataset. The main contribution of this work is that it requires minimal supervision (i.e., just one seed word per domain aspect plus one positive and one negative word independent of the domain) to perform ABSA over any unlabelled corpus of customer reviews. The lack of domain or language dependencies allows the system to be readily used for other domains and languages. Another contribution is the automatic separation of domain aspects into aspect-terms, positive words and negative words which facilitates the interpretation of the modelled topics. The source code will be made publicly available.[1]

After this short introduction, the paper is structured as follows. Previous related work is first reviewed in Section 2. Then, Section 3 describes our system, including the seed word based configuration, aspect-term and opinion-word separation and topic modelling parts. In Section 4, evaluation results are presented. Finally, Section 5 describes the main 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 (Chen, Mukherjee, & Liu, 2014; Liu, Xu, & Zhao, 2012). 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 (Blair-Goldensohn et al., 2008; Hu & Liu, 2004; Popescu & Etzioni, 2007; Qiu, Liu, Bu, & Chen, 2011; Wu, Zhang, Huang, & Wu, 2009). 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 of bootstrapping or polarity propagation leveraging some base dictionaries and pre-existing linguistic resources (Huang, Niu, and Shi, 2014; Jijkoun, de Rijke, & Weerkamp, 2010; Rao & Ravichandran, 2009).

A well-known unsupervised method for modelling text in documents is Latent Dirichlet Allocation (LDA). LDA is a generative model introduced by Blei, Ng, and Jordan (2003) that quickly gained popularity because it is unsupervised, flexible and extensible. 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 sev-

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

eral kinds of tasks, such as dimensionality reduction or unsupervised clustering. Due to its flexibility, LDA has been extended and combined with other approaches, in order to obtain improved topic models or to model additional information (Mcauliffe & Blei, 2008; Ramage, Hall, Nallapati, & Manning, 2009).

Topic models have been applied to Sentiment Analysis to jointly model topics and the sentiment of words (Alam, Ryu, & Lee, 2016; Jo & Oh, 2011; Kim et al., 2013; Lin et al., 2011; Lin, Road, & Ex, 2009; Lu, Ott, Cardie, & Tsou, 2011). A usual way to guide a topic modelling process towards a particular objective is to bias the LDA hyper-parameters using certain a priori information. When modelling the polarity of the documents, this 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 domain polarity words by just a single domain independent positive word (e.g., *excellent*) and a single domain independent negative word (e.g., *horrible*).

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 manual inspection by an expert or a mapping calculation to an existing resource (Bhatia, Lau, & Baldwin, 2016). Our approach relies on a minimal topic configuration step, in which the user defines the domain aspects to monitor in the target domain. Thus, the resulting topics match the domain aspects defined initially by the user. This is done by leveraging semantic word similarities to guide the topic modelling process towards the defined domain aspects. Semantic word similarity is obtained using continuous word embeddings over the domain words. Continuous word embeddings are known for capturing semantic regularities of words (Collobert & Weston, 2008; Mikolov, Chen, Corrado, & Dean, 2013a). Some works have exploited this fact to improve topic modelling results (Das, Zaheer, & Dyer, 2015; Nguyen, Billingsley, Du, & Johnson, 2015; Qiang, Chen, Wang, & Wu, 2016), but their objective is to enhance the unsupervised modelling of a corpus instead of guiding the model towards a predefined set of topics. There are also works that exploit word embeddings in a supervised machine learning setting to perform sentiment analysis (Giatsoglou et al., 2017; Tang et al., 2014).

Some authors have also attempted to separate aspect-terms and opinion-words automatically within the topic modelling process (Mukherjee & Liu, 2012; Zhao, Jiang, Yan, & Li, 2010). 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 *slow*. 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, Desouza, Mercer, Pietra, & Lai, 1992) to a set of training instances from an unlabelled corpus in order to train an aspect-term and 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 can be applied to any domain and language.

In summary, combining topic modelling, continuous word embeddings and a minimal topic definition, our system can model customer reviews in different domains and languages performing three subtasks at the same time: domain aspect classification, aspect-terms/opinion-words separation and sentiment polarity classification. To our knowledge, no other system performs these three tasks with such a minimal configuration of seed words, without requiring manually annotated training examples, pre-existing language or domain dependent resources.

3. System description

The main objective of the W2VLDA system is to perform the three main tasks 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 aspect-terms from opinion-words without requiring additional resources or supervision. The system at its core consists of an LDA-based topic model, extended with biased topic modelling hyper-parameters based on continuous word embeddings and combined with an unsupervised pre-trained classification model for aspect-term and opinion-word separation.

3.1. Topics and sentiment configuration

W2VLDA only requires a minimal domain aspects and sentiment polarity configuration per target domain and language, which consists on defining a single seed aspect-term for each desired domain aspect, plus a single positive seed word and a single negative seed word independent of the domain. These features are the only domain and language dependent information required by W2VLDA.[2] Therefore, a simple translation of the seed words in one language is enough to make the system work in another language, as long as each translated seed has an equivalent meaning and use in the target language. Table 1 shows an example of domain aspects and sentiment polarity configuration for the restaurants domain in several languages.

3.2. Aspect-term and opinion-word separation

Part of the outcome of the system is the separation of aspect-terms and opinion-words into differentiated word classes. In order to achieve such separation without any additional language dependent tool or resource, the system uses Brown clusters (Brown et al., 1992) to model examples of aspect-terms and opinion-words and train a Maximum Entropy (MaxEnt) based classification model. Brown clusters have been used as unsupervised features with good results in supervised Part-of-Speech tagging (Turian, Ratinov, & Bengio, 2010) and Named Entity Recognition (Agerri & Rigau, 2016). Brown clusters are computed[3] from the unlabelled domain corpus without additional supervision. They are then used as features for the context words of the instances used to train the MaxEnt classifier. The training instances are obtained leveraging the occurrences of the initial configuration with domain aspects and polarity seed words, assuming that the domain aspect seed words are aspect-terms and the polarity seed words are opinion-words.

Fig. 3 describes the process to obtain the classification model. First, domain aspect seed terms 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 unlabelled domain corpus and modelled according to their two word, [−2,+2], context window. Next, context words are replaced by their corresponding Brown cluster to build each training instance. Finally, a MaxEnt model is trained using these training instances.

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

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

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

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	Epouvantable

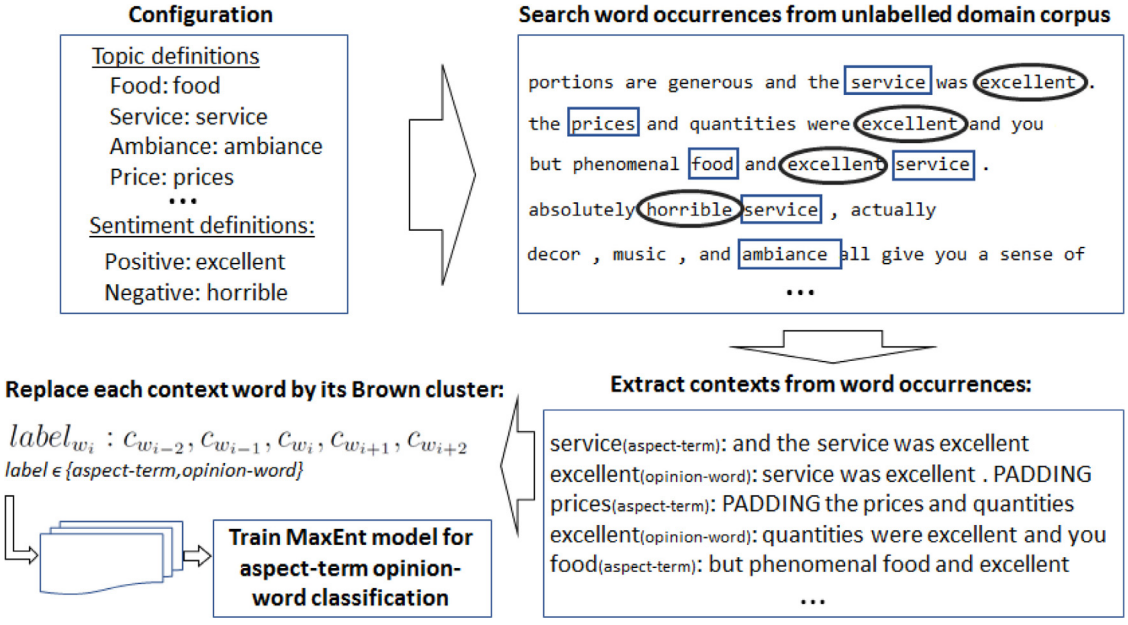


Fig. 3. Process to obtain the MaxEnt model for aspect-term and opinion-word separation.

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 aspect-term and opinion-word separation. 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 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.

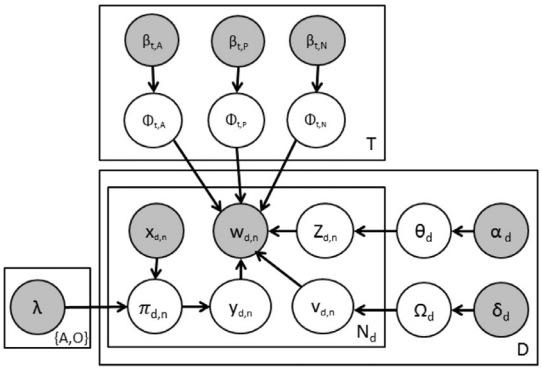
3.3. Combining everything in a topic model

The core of the system is a LDA-based topic model, extended to include aspect-term and opinion-word separation and sentiment polarity classification for each defined domain aspect. While aspect-term and opinion-word separation is guided by a pre-trained classifier as described in Section 3.2, topic (i.e., domain aspect) and polarity modelling are done by biasing certain hyper-parameters according to the given domain aspects configuration.

Fig. 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. For each word n in document d a topic value is drawn: $z_{d,n} \sim Multi(\theta_d)$, $z \in \{1..T\}$. Then, an 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 because no a priori 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 domain aspect seed terms 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, Yih, & Zweig, 2013b). In this case, we exploit the semantic similarity among words that can be calculated using the cosine



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 $w_{d,n} \sim \text{Multinomial}(\phi^N_{z_{d,n}})$

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

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 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).

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

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).

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

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

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

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

In the formulas $w_{d,i}$ is the ith 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 set of polarity words, 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 and 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 & Steyvers, 2004). Let $w_{d,n}$ be the nth word of the dth 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.

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

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

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

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 and 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: domain aspect classification, aspect-term and opinion-word separation and sentiment polarity classification. First, we compare W2VLDA with other LDA-based methods. Then, we also evaluate W2VLDA on a multilingual ABSA dataset comparing its

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

Language:domain	Domain aspect	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 3
Resulting domain aspect word distributions for two languages, Spanish and French, and for different domains. The domain aspects are automatically split into three different word distributions: aspect terms, positive words and negative words.

Language:domain	Domain aspect	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

domain aspect and sentiment polarity classification performance against other supervised machine learning approaches trained on labelled data.

To illustrate the final output of the system, we show results for several datasets, demonstrating how the system works in different domains and languages just by changing the initial configuration, composed of a single seed aspect-term for each desired domain aspect, plus a single domain independent positive seed word and a single domain independent negative seed word.

For instance, Table 2 shows some of the resulting words for two domains in English: the restaurants and electronic devices reviews. The examples include the automatic separation of aspect-terms from positive and negative words per domain aspect. Table 3 shows the equivalent information for the restaurant and hotel reviews domain in Spanish and French, respectively.

Likewise, Table 4 shows examples of sentences classified under different domain aspects for the restaurant reviews domain in English and Spanish and the hotel reviews domain in French.

4.1. Resources and experimental setting

In order to evaluate W2VLDA, we use the following resources. For topic or aspect category classification, we use the dataset from Ganu, Elhadad, and Marian (2009) which contains restaurant reviews labelled with domain-related aspects (e.g., food, staff, ambience) in English. For sentiment classification, we use the Laptops and DIGITAL-SLR dataset (Jo & Oh, 2011), which consists 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 aspects and polarity for six languages.

In order to compute the topic model and the required word embeddings, we have automatically gathered additional restaurant reviews from some popular customer review websites. These unlabelled domain corpora consist of a few thousand restaurant reviews in English, Spanish, French and 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[4] implementation with default parameters to compute the domain-based word embeddings.

Table 1 shows the domain aspects and polarity definition used in the experiments for the restaurants domain. Unless stated otherwise, the polarity seeds used for every domain are *excellent* and *horrible* for English and their equivalents in other languages.

The values for α_{base}, β_{base} and δ_{base} mentioned in Section 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 & 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 domain aspect classification setting using the restaurant reviews dataset from Ganu et al. (2009). This dataset contains a few thousand restaurant reviews 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 & Elhadad, 2010) and ME-LDA (Zhao et al., 2010).

[4] http://spark.apache.org/mllib/.

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

Lang: domain	Domain aspect	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 min à pied de buckingham palace et saint james park , 10 à 15 min 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 5
Comparison of domain aspect classification against other LDA based approaches in the restaurant reviews domain.

Method	Topics											
	Staff			Food			Ambiance			Overall		
	Prec.	Rec.	F1	Prec.	Rec.	F1	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**

LocLDA and ME-LDA are LDA-based approaches, and thus, unsupervised. However, the reported results involved some supervision as described in Zhao et al. (2010). First, the authors computed a topic model of 14 topics. Then they examined each topic and manually set a label according to their judgment. W2VLDA provides a set of labelled topics according to the domain aspects defined in the initial configuration step, 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 (i.e., domain aspect) with highest posterior probability.

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

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

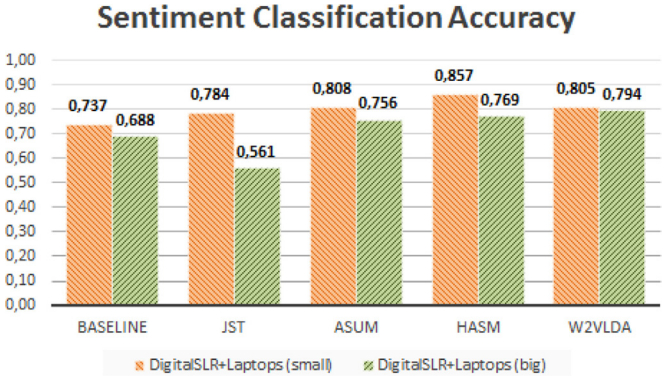


Fig. 5. Comparison of sentiment classification accuracy with other LDA based approaches in the electronic devices reviews domain.

Fig. 5 shows the sentiment classification comparison results. 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 the SemEval2016 dataset

We use the SemEval 2016 task 5 dataset (Pontiki et al., 2016) to perform a multilingual evaluation of W2VLDA. The SemEval 2016 dataset consists of restaurant reviews in several languages. Reviews are split by sentence and labelled with explicit aspect-term mentions, the coarse-grained domain aspect or category they belong to and their polarity for that aspect or category. Fig. 6 shows an example of SemEval 2016 restaurant reviews dataset for English.

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

```
<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>
```

Fig. 6. SemEval 2016 task 5 English restaurant reviews dataset example.

Table 6
SemEval 2016 dataset category distribution after filtering out sentences with more than one annotation and low occurrence categories.

	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 7
SemEval 2016 dataset polarity distribution after filtering out sentences with more than one annotation.

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

The SemEval 2016 restaurant reviews dataset is annotated for six coarse-grained categories: food, service, ambience, drinks, location and restaurant. The last 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. Because such an abstract concept cannot be represented by a seed word, we omit this category from the evaluation. To avoid ambiguities and simplify sentence classification, we only keep those sentences with a single category label. Finally, we filter out the *drinks* and *location* categories because their occurrence in the dataset is below 5% of the instances. Thus, we carry out multilingual evaluation of W2VLDA on the three main categories of the SemEval 2016 restaurant reviews dataset: *food, service* and *ambience*.

Tables 6 and 7 show respectively the distribution of domain aspects and sentiment polarities for the datasets used in our evaluation for four languages: English, Spanish, French and Dutch.

Since W2VLDA follows a topic modelling approach, it needs a reasonable amount of domain documents to build the statistical model. To cope with this requirement, we have implemented a script that automatically extracts restaurant reviews in the required languages from an online customer reviews website.[6] Table 8 shows the amount of downloaded restaurant reviews used to feed the algorithm. Their polarity has been derived from the star ratings (as usual, 1–2 stars meaning negative and 4–5 starts meaning positive). As it can be seen in the table, the number of positive and negative reviews is unmatched for some languages. In order to try to compensate this fact, we oversample negative examples for each language until they equal the positive ones in number (i.e., until they reach 10 k). Note that for Dutch, this may lead to excessive oversampling given the small number of available negatives examples. Also, note that these polarity labels are just used to get an insight of the polarity distribution of the datasets, but are not used for any sort of supervised training.

Table 8
Language and polarity (5-star rating) distribution of the downloaded restaurant reviews. We could not download the same number of positive and negative reviews for all languages. We try to alleviate this problem oversampling negative reviews.

Restaurant customer reviews downloaded from a website				
	EN	ES	FR	NL
Positives (4 or 5 stars)	10,000	10,000	10,000	10,000
Negatives (1 or 2 stars)	10,000	8,400	5,500	830
Total reviews	20,000	18,400	15,500	10,830

The evaluation experiment is done as follows. For each language, we use the downloaded restaurant reviews to run the W2VLDA algorithm. This involves calculating the domain word embeddings, the Brown clusters and the topic model estimation. Using the model generated for each language, the domain aspect and polarity distributions θ and Ω are then estimated for each of the sentences in the evaluation set. The topic with the highest probability in the estimated topic distribution for that sentence is assigned as the domain aspect category label. Analogously, the polarity with the highest probability in the estimated polarity distribution for that sentence is assigned as the polarity label. The assigned domain aspect label 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.

[6] Due to copyright restrictions, we cannot share these reviews.

Table 9
Multilingual domain aspect classification results on the SemEval2016 dataset. NB and MLP are the supervised baselines, NaiveBayes and MultiLayer perceptron, respectively. Majority baseline shows the result of simply choosing the most frequent class. W2VLDA_NO is the proposed approach without word embeddings. W2VLDA is the proposed approach.

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 10
Multilingual sentiment polarity classification results on the SemEval 2016 dataset. NB and MLP are the supervised baselines, NaiveBayes and MultiLayer perceptron, respectively. Majority baseline shows the result of simply choosing the most frequent class. W2VLDA_NO is the proposed approach without word embeddings. W2VLDA is the proposed approach.

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**

The accuracy achieved is compared to several baselines. First, two supervised baselines are used. One is a Naive–Bayes classifier (NB), trained on the labelled sentences. The training sentences are transformed to bag-of-words vectors with a vocabulary size of 80 k 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, which shows the accuracy that can be obtained when 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 used to bias the topic modelling hyper-priors. Instead of using the word-embedding 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 multilingual domain aspect classification results on the SemEval2016 evaluation dataset. Since the evaluation dataset is 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. Results on each individual subset are then obtained using the average accuracy of 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 95% confidence for all the languages except for Dutch, for which despite obtaining better results than the baselines it only achieves 80% confidence in the statistical significance test.

Table 10 shows the multilingual sentiment polarity classification results on the SemEval2016 dataset. Result calculations and statistical significance tests have been performed as for domain aspect classification evaluation. Again, W2VLDA outperforms the baselines with 95% confidence in the statistical test, except for Dutch. A possible reason for this is that the oversampling performed on the downloaded Dutch reviews for topic modelling was excessive, or that the data contained in the downloaded Dutch corpus was less representative than for other languages (see Table 8). Studying the lower bound limits of the amount of unlabelled data required to train the W2VLDA approach is an interesting problem that we leave for future research work.

4.4. Assessing the impact of the seed words

Since the proposed approach heavily relies on the defined 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 using different seed words and seed word combinations.

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, we only change the seed words that define the domain aspects in each run. The polarity seed words remain the same.

We use three different seed words for each domain aspect, in particular: *food, chicken* and *burger* for the *FOOD* domain aspect; *service, staff* and *waiter* for the *SERVICE* domain aspect; and *ambience, atmosphere* and *décor* for the *AMBIENCE* domain aspect. We try different permutations and combinations of the seed words, including the use of pairs of seed words for each domain aspect, and finally also the combination of the three seed words together. Table 11 shows the results of this experiment. As it can be seen, the obtained accuracy is stable across all combinations regardless of the chosen seed words. As expected, some combinations perform better than others but overall the average is high and the standard deviation is below 5%. The best result is achieved using all the seed words at the same time. This is not surprising since the semantic coverage to guide the algorithm to the desired domain aspects increases with the amount of seed words, as long as these 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 polarity results, as it would be expected. The polarity results show minor variations across the experiments and their standard deviation is only a 0.8%.

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

Finally, in order to perform a sanity check and evaluate if sentiment polarity classification really depends 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* as positives and *dog* and *waiter* as negatives. The use of these words as polarity seeds is obviously incorrect but what we want to check is if using such meaningless polarity words leads to bad polarity classification results. Table 13 shows the results for this experiment, confirming that the selection of representative polarity seed words is important to correctly guide the algorithm.

Table 11
Impact of different seed word combinations on domain aspects classification.

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 12
Impact of different polarity seeds word combinations on sentiment polarity classification.

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 13
Sentiment polarity classification results using misleading words as polarity seeds, to check to what extent sentiment polarity classification depends on the validity of the chosen polarity seeds.

Polarity:{POSITIVE},{NEGATIVE}	Aspects acc.	Polarity acc.
{cat},{dog}	0.642	0.447
{waitress},{waiter}	0.635	0.419

4.5. Aspect-term/opinion-word separation evaluation

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

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

The experiment now involves running the W2VLDA algorithm 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 a 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 and opinion-word separation is correct, the proportion of correctly classified aspect-terms and opinion-words should be high.

We perform several experiments varying the number of Brown clusters involved in the process (see Section 3.2) in order to evaluate its impact on aspect-term and opinion-word separation. Fig. 7 shows the resulting proportions of correctly assigned aspect-terms and opinion-words. In general, they are high compared to random assignment, which indicates that automatic aspect-term and opinion-word separation performs correctly most of the times. Interestingly, aspect-terms are better distinguished than opinion-words.

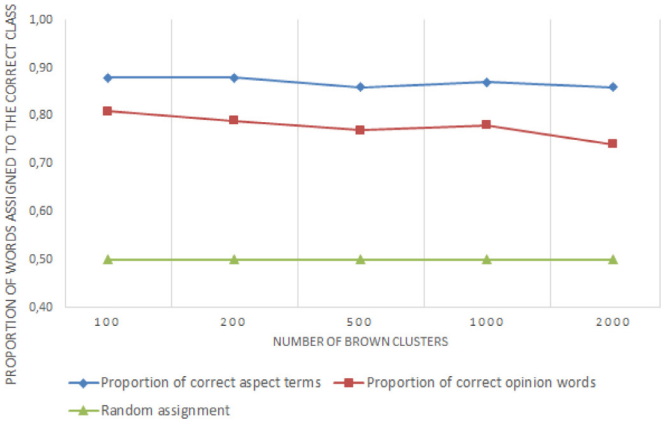


Fig. 7. Results of aspect-term and opinion-word separation for English. Each point indicates the proportion (percentage) of aspect-terms or opinion-words that have been correctly classified. Random assignment is the random guess baseline.

5. Conclusions and future work

In this document, we have presented W2VLDA, a system that performs aspect and sentiment classification almost with no supervision and without the need of domain or language specific resources. In order to do that, the system combines several unsupervised approaches, such as Latent Dirichlet Allocation (LDA) or word embeddings, to bootstrap information from a domain corpus. The only supervision required by the user is a single seed word per desired domain aspect and per polarity. Because of that, the system can be applied to datasets in different domains and languages almost with no adaptation. The resulting topics and polarities are directly paired with the domain aspects defined by the user at the beginning, so the output can be used to perform Aspect Based Sentiment Analysis (ABSA). In addition, the system separates aspect-terms and opinion-words automatically, facilitating the interpretation of the resulting domain aspect vocabulary. We evaluate W2VLDA for domain aspect and polarity classification using customer reviews in several domains and com-

pare it against other LDA-based approaches, achieving slightly better overall results despite using lower supervision. We also evaluate its performance using a subset of the multilingual SemEval 2016 task 5 ABSA dataset, outperforming the baselines with high confidence in statistical significance for all languages in both, domain aspects and polarity classification. As future work, it would be interesting to include an automated way to deal with stopwords 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 results. On the other hand, it would be interesting to study if more specialised word embeddings related to sentiment analysis (Rothe, Ebert, & Schütze, 2016) would have a positive impact on the results, always 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). Thanks to Dr Arantza del Pozo for a final proofreading of the article. We also thank the anonymous reviewers for their useful comments and suggestions.

References

Agerri, R., & Rigau, G. (2016). Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence, 238*, 63–82. http://dx.doi.org/10.1016/j.artint.2016.05.003.

Alam, M. H., Ryu, W. J., & Lee, S. K. (2016). Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews. *Information Sciences, 339*, 206–223. doi:10.1016/j.ins.2016.01.013.

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & 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., & Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)* (pp. 953–963).

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Brody, S., & Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. In *Proceedings of the annual conference of the North American chapter of the association for computational linguistics*, 804–812. (June).

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

Chen, T., Xu, R., He, Y., & 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., & Liu, B. (2014). Aspect extraction with automated prior knowledge learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 347–358.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the twenty-fifth international conference on machine learning* (pp. 160–167). ACM.

Das, R., Zaheer, M., & Dyer, C. (2015). Gaussian LDA for topic models with word embeddings. In *Proceedings of the fifty-third annual meeting of the association for computational linguistics* (pp. 795–804).

Ganu, G., Elhadad, N., & Marian, A. (2009). Beyond the stars: Improving rating predictions using review text content. In *Proceedings of the Webdb: vol. 9* (pp. 1–6). Citeseer.

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

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

Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the AAAI: 4* (pp. 755–760).

Huang, S., Niu, Z., & 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., & Weerkamp, W. (2010). Generating focused topic-specific sentiment lexicons. In *Proceedings of the forty-eight annual meeting of the association for computational linguistics* (pp. 585–594). Association for Computational Linguistics.

Jo, Y., & 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* (pp. 815–824). ACM.

Kim, S., Zhang, J., Chen, Z., Oh, A., & Liu, S. (2013). A hierarchical aspect-Sentiment model for online reviews. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. pp. 526–533.

Lin, C., He, Y., Everson, R., & Rüger, S. (2011). Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering, 24*, 1134–1145. doi:10.1109/TKDE.2011.48.

Lin, C., Road, N. P., & Ex, E. (2009). Joint sentiment / topic model for sentiment analysis. In *Proceedings of the conference on Information and knowledge management*, (pp. 375–384). doi:10.1145/1645953.1646003.

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

Liu, K., Xu, L., & Zhao, J. (2012). Opinion target extraction using word-based translation model. In *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., & Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, 81–88. doi:10.1109/ICDMW.2011.125.

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Proceedings of the advances in neural information processing systems* (pp. 121–128).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, 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., & Liu, B. (2012). Aspect extraction through semi-supervised modeling. In *Proceedings of the fiftieth annual meeting of the association for computational linguistics: Long papers, vol. 1*(July), 339–348.

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

Pang, B., & 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., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the tenth international workshop on semantic evaluation (Semeval-2016)* (pp. 19–30). San Diego, CA: Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the ninth international workshop on semantic evaluation* (pp. 486–495). Denver, CO: Association for Computational Linguistics.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the eight international workshop on semantic evaluation* (pp. 27–35). Dublin, Ireland: Association for Computational Linguistics.

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

Qiang, J., Chen, P., Wang, T., & Wu, X. (2016). Topic Modeling over Short Texts by Incorporating Word Embeddings. (p. 10). arXiv:1609.08496v1. 10.1145/1235.

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

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the conference on empirical methods in natural language processing: vol. 1* (pp. 248–256). Association for Computational Linguistics.

Rao, D., & Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. In *Proceedings of the twelfth conference of the European chapter of the association for computational linguistics* (pp. 675–682). Association for Computational Linguistics.

Rothe, S., Ebert, S., & Schütze, H. (2016). Ultradense word embeddings by orthogonal transformation. In *Proceedings of the North American chapter of Association for Computational Linguistics (NAACL-HLT 2016)* (pp. 767–777).

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-Specific word embedding. *ACL*, 1555–1565.

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

Turney, P. D., & 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., & Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the conference on empirical methods in natural language processing: vol. 3* (pp. 1533–1541). Association for Computational Linguistics.

Zhao, W. X., Jiang, J., Yan, H., & Li, X. (2010). Jointly modeling aspects and opinions with a maxent-LDA hybrid. *Computational Linguistics, 16*(October), 56–65.