Journal of Artificial Intelligence Research 58 (2017) 591–626 Submitted 6/16; published 3/17

Explicit Document Modeling through Weighted Multiple-Instance Learning

Nikolaos Pappas NIKOLAOS.PAPPAS@IDIAP.CH
Andrei Popescu-Belis ANDREI.POPESCU-BELIS@IDIAP.CH
Idiap Research Institute, Rue Marconi 19
CH-1920 Martigny, Switzerland

Abstract

Representing documents is a crucial component in many NLP tasks, for instance predicting aspect ratings in reviews. Previous methods for this task treat documents globally, and do not acknowledge that target categories are often assigned by their authors with generally no indication of the specific sentences that motivate them. To address this issue, we adopt a weakly supervised learning model, which jointly learns to focus on relevant parts of a document according to the context along with a classifier for the target categories. Derived from the weighted multiple-instance regression (MIR) framework, the model learns decomposable document vectors for each individual category and thus overcomes the representational bottleneck in previous methods due to a fixed-length document vector. During prediction, the estimated relevance or saliency weights explicitly capture the contribution of each sentence to the predicted rating, thus offering an explanation of the rating. Our model achieves state-of-the-art performance on multi-aspect sentiment analysis, improving over several baselines. Moreover, the predicted saliency weights are close to human estimates obtained by crowdsourcing, and increase the performance of lexical and topical features for review segmentation and summarization.

1. Introduction

Many NLP tasks such as document classification, question answering, and summarization heavily rely on how well the contents of the document are represented in a given model. In particular, when classifying the sentiment of documents towards an item, the attitude of the author generally results from the ratings of several specific aspects of the item. For instance, the author of a review might have a rather positive overall sentiment about a movie because they have particularly liked the plot and the setting, but not too much the actors. Determining the rating of each aspect automatically is a challenging task, mainly because it is difficult to find a fixed-length document representation which works well for all the aspects. This task is typically cast as a supervised learning problem and has been previously addressed by engineering or learning a large number of features to represent each review, which are then fed to a linear classifier (McAuley, Leskovec, & Jurafsky, 2012; Zhu, Zhang, & Ma, 2012; Tang, 2015). However, such models ignore that sentences have diverse contributions to a document's overall or aspect-specific sentiments, and that document-level labels are coarse, in the sense that it is uncertain which parts of the text motivate them.

One way to ameliorate this issue is to select only sentences that discuss the targeted aspect (Zhu et al., 2012) but this requires a preliminary segmentation of texts, which is costly, and ignores sentences that have only a partial relation to an aspect. Another solution

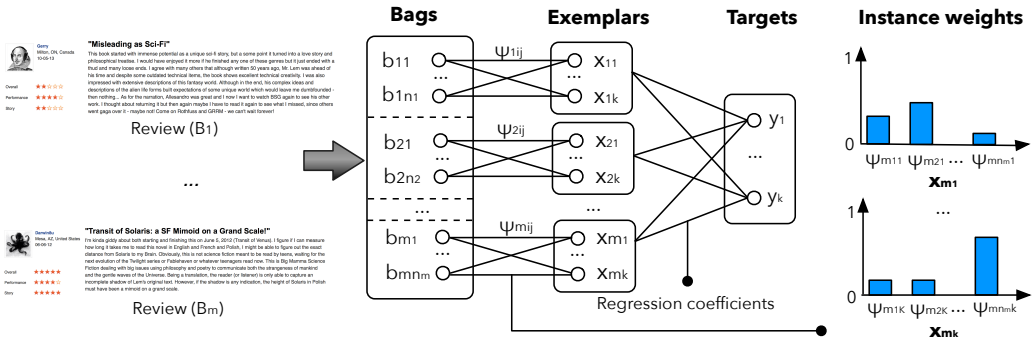


Figure 1: Explicit document modeling using weighted multiple-instance regression. The model takes as input a document B_i (bag), which consists of multiple input vectors b_{ij} (instances), possibly from a neural network. The model learns to compute a weighted average of these vectors by estimating the weights ψ_{ij} for each document B_i and its target categories $y_i \in R^k$.

is to model sentence-level labels as latent variables to be learned jointly with the document classifier (Titov & McDonald, 2008a; McAuley et al., 2012; Lei, Barzilay, & Jaakkola, 2016). Such binary latent assignments, however, are prone to errors on partially relevant sentences, and lead to latent document representations which are typically rigid, implicit, and difficult to interpret.

In this paper, we propose a weakly supervised approach based on weighted multiple-instance regression (MIR) represented in Figure 1. We aim to answer the question: "To what extent does each part of a document contribute to the prediction of a document-level label?" Given a set of input vectors or intermediate hidden states of a neural network for each document, the model learns to explicitly assign relevance weights to parts of the document according to the context, as well as a classifier for the target labels. The model thus learns decomposable and flexible document vectors for each individual category and overcomes the representational bottleneck in previous methods that use one fixed document vector for all categories. For training, our model only requires document-level labels (e.g. aspect ratings), makes no particular assumption on the word features, and has reasonable computational demands. Lastly, the learned instance weights have explanatory power, and can be combined with sequence models for aspect-based review segmentation and summarization. Specifically, we make the following contributions:

1. We propose a joint model comprised of an instance relevance mechanism, which explicitly summarizes the important contents of a document, through a decomposable representation and a document classifier for the target labels. The model is directly based on a weighted multiple-instance learning framework (Pappas & Popescu-Belis, 2014) and is mathematically equivalent to recent attention mechanisms in NLP (Bahdanau, Cho, & Bengio, 2015).

2. We demonstrate that for document-level aspect rating prediction, our model consistently outperforms several standard MIR and non-MIR baselines, as well as state-

of-the-art neural baselines. The benefit is observed across several datasets, feature spaces, and types of input vectors or intermediate hidden states of a network.

3. We evaluate the explanatory power of our model by comparing its predicted aspect saliency values with those assigned by humans, using a novel dataset for evaluating attention-based methods in document classification (Pappas & Popescu-Belis, 2016).

4. We show that the learned aspect saliency values are beneficial to review segmentation and summarization, as they augment word or topic feature spaces with structural information about the input.

The paper is organized as follows. We compare our proposal with previous studies of multi-aspect sentiment analysis and of multiple instance learning in Section 2. In Section 3, we formulate the problem of aspect rating prediction as weakly supervised text regression. In Section 4, we propose an efficient weighted MIR model to solve it, and in Section 5 we show how the learned features can be used to segment and summarize reviews. In Section 6, we describe the data used in the experiments. In Section 7, we evaluate the model on aspect rating prediction. Then, we evaluate the aspect saliency values, first intrinsically, by comparing them to human annotations (Section 8), and then extrinsically, on review segmentation and summarization (Section 9).

2. Related Work

We review related work along three main directions: sentiment analysis, with a focus on aspect-based studies; interpretable models in machine learning, comparing to neural networks with attention; and multiple-instance learning, the main framework of our study.

2.1 Multi-aspect Sentiment Analysis

Sentiment analysis typically aims to detect the polarity of a given text, and is commonly formulated as a classification problem for discrete labels such as 'positive' and 'negative', or a regression problem for real-valued labels (Pang & Lee, 2005, 2008). Pang and Lee (2008) survey the large range of features that have been engineered, either for rule-based sentiment analysis methods (Hatzivassiloglou & Wiebe, 2000; Hu & Liu, 2004; Wilson, Wiebe, & Hoffmann, 2005), or for corpus-based ones (Pang, Lee, & Vaithyanathan, 2002; Thomas, Pang, & Lee, 2006). Machine learning techniques for sentiment classification have been introduced quite early by Pang et al. (2002) among others.

Following initial studies on *feature engineering* (e.g., Pang & Lee, 2008), more recent studies have focused on feature learning (Maas et al., 2011; Socher et al., 2011; Tang et al., 2014), including the use of deep neural networks (Socher et al., 2013; Mikolov et al., 2013; Tang, 2015). These methods do not require costly engineering of features, but the learned features are typically difficult to interpret in terms of semantic properties (Li, Chen, Hovy, & Jurafsky, 2015). The above methods use one fixed-length vector to represent a document for each label, whereas our proposal accounts for varying contributions of sentences. Our study innovates with respect to feature engineering and learning approaches at the same time, because our multiple-instance formulation focuses on the vector composition objective and not on the learning of the feature space itself.

The fine-grained analysis of opinions on specific aspects or features of items is known as *multi-aspect sentiment analysis*. This task usually requires aspect-related text segmentation, followed by prediction or summarization (Hu & Liu, 2004; Zhuang, Jing, & Zhu, 2006). Most attempts to perform this task have engineered various feature sets, augmenting words with topic or content models (Mei, Ling, Wondra, Su, & Zhai, 2007; Titov & McDonald, 2008b; Sauper, Haghighi, & Barzilay, 2010; Lu, Ott, Cardie, & Tsou, 2011), or with linguistic features (Pang & Lee, 2005; Baccianella, Esuli, & Sebastiani, 2009; Qu, Ifrim, & Weikum, 2010; Zhu et al., 2012). Other studies have advocated the joint modeling of multiple aspects (Snyder & Barzilay, 2007) or of multiple reviews for the same item (Li et al., 2011). McAuley et al. (2012) introduced new corpora of multi-aspect reviews, and proposed an interpretable probabilistic model for modeling aspect reviews called PALE LAGER, which we use for comparison in this paper. Kim, Zhang, Chen, Oh, and Liu (2013) proposed an hierarchical model to discover the structure of aspect-related sentiment from unlabeled corpora. Joint aspect identification and sentiment classification have been used by Sauper and Barzilay (2013) to aggregate product reviews, while Lakkaraju, Socher, and Manning (2014) proposed a hierarchical deep learning framework to achieve the same goal. Another related task is entity-based sentiment analysis, where the goal is to model the sentiment regarding specific entities (Mitchell, Aguilar, Wilson, & Van Durme, 2013; Deng & Wiebe, 2015; Choi, Rashkin, Zettlemoyer, & Choi, 2016).

Previous studies of aspect rating prediction from text have used automatically segmented texts for training (Zhu et al., 2012; McAuley et al., 2012), or have modeled the relationships between different aspect ratings (Lin & He, 2009; Gupta, Di Fabbrizio, & Haffner, 2010; McAuley et al., 2012) to go beyond standard supervised models such as SVM with bags-of-words. Recently, several studies combined collaborative filtering and topic modeling to perform aspect rating prediction (McAuley & Leskovec, 2013; Bao, Fang, & Zhang, 2014; Wu, Beutel, Ahmed, & Smola, 2015), but as they are not only based on text, they are not comparable to our work. To our knowledge, none of the previous studies considered in their modeling the weak relationship between text labels and the parts of texts (e.g. sentences) as we propose here. Furthermore, we trained our model over the entire unsegmented text, reducing the computational cost and human intervention that is required to obtain segmented text. In addition, we capture meaningful structural information of the input text, instead of the output labels only, and thus provide interpretable sentence weights, which can be used for segmenting and summarizing reviews.

Most previous studies of *review segmentation and summarization* are unsupervised (Titov & McDonald, 2008a; Zhu, Wang, Tsou, & Zhu, 2009; Wang, Lu, & Zhai, 2010; Brody & Elhadad, 2010; Lu et al., 2011), while fewer studies explored supervised learning (Li et al., 2010). Recently, the availability of annotated data (McAuley et al., 2012; Pontiki et al., 2014) has increased the interest in supervised methods, e.g. with constrained structured models (McAuley et al., 2012), or with linear chain CRF models (Patra, Mandal, Das, & Bandyopadhyay, 2014; Hamdan, Bellot, & Bechet, 2015). While sentence sentiment has been shown to be useful for inferring sentence aspects (Brody & Elhadad, 2010; Ganu, Elhadad, & Marian, 2009), the aspect saliency and sentiment of sentences from in-domain corpora have not been considered before, while here, they will be used to augment word or topic spaces.

2.2 Interpretable Models in Machine Learning

Our model can be seen as a parametrized pooling layer in a neural network, for instance substituting intermediate pooling layers, which typically use average or summation with equal weights. Our initial proposal for an instance relevance mechanism (Pappas & Popescu-Belis, 2014) has a close resemblance to the "attention" mechanisms proposed later for machine translation, which selectively focus on parts of the input (Bahdanau et al., 2015; Hermann et al., 2015; Luong et al., 2015; Zhao et al., 2015), and we argue in Section 4.2 that they are in fact equivalent. Attention mechanisms have been shown lately to be quite useful for a variety of NLP tasks, including machine translation (Bahdanau et al., 2015; Luong et al., 2015; Sennrich, Haddow, & Birch, 2016), question answering (Sukhbaatar, Szlam, Weston, & Fergus, 2015; Xiong, Merity, & Socher, 2016), summarization (Rush, Chopra, & Weston, 2015; Chopra, Auli, & Rush, 2016), image captioning (Xu et al., 2015), and document classification (Yang et al., 2016). The last study is perhaps the closest to ours: it applied an attention mechanism to each level of an hierarchical neural network over documents, with large improvements over the state-of-the-art. However, it did not evaluate quantitatively the explanatory potential of the attention mechanism. To the best of our knowledge, we were the first to introduce an attention-based method for document classification (Pappas & Popescu-Belis, 2014), albeit outside the neural framework.

Recently, there has been an interest in algorithms which are interpretable and capable of explaining classification predictions (Lou, Caruana, & Gehrke, 2012; Lou, Caruana, Gehrke, & Hooker, 2013; Kim, 2015; Lipton, 2016; Das, Agrawal, Zitnick, Parikh, & Batra, 2016). For instance, Ribeiro, Singh, and Guestrin (2016) proposed a model-agnostic technique that explains the predictions of any type of classifier. This method is unaware of the internal mechanisms of the classifier, and hence the explanation may not directly align with the actual prediction process, while our mechanism is learned jointly with the classifier. Lei et al. (2016) proposed a method for rationalizing neural predictions using an encoder which defines a distribution over text fragments, called rationales, which are passed on to the decoder for prediction. This method was evaluated on the BeerAdvocate dataset from McAuley et al. (2012), and we include it in our comparisons below, where applicable. The method was shown to outperform previous attention models at the word level, but it is unclear whether binary selections also work well on larger parts of an input text, such as sentences, or other intermediate hidden states of a deeper neural network, which may be important for prediction. In contrast, this is not an issue for the attention-based methods like ours. Another open question is how to extend Lei et al.'s method to hierarchical neural networks, which can learn deep compositional text representations (Yang et al., 2016).

2.3 Multiple Instance Learning

Multiple-instance learning (MIL) is a machine-learning approach originally proposed by Dietterich, Lathrop, and Lozano-Prez (1997) to deal effectively with coarse-grained input labels. The MIL algorithms receive as input a set of labeled bags, each of which contains a variable number of instances; however, these instances are not individually labeled, as in traditional supervised learning. The goal of MIL is either to learn a classifier which assigns correct labels to individual instances, or to predict the labels of the bags without necessarily inducing the labels of each individual instance. Comprehensive surveys and comparisons of

MIL methods are available (Foulds & Frank, 2010; Amores, 2013). MIL has been successfully applied to a variety of domains such as image classification, molecule classification for drug discovery, drug activity prediction, remote sensing and text or document categorization. The majority of the MIL studies focused on classification (Andrews, Tsochantaridis, & Hofmann, 2003; Bunescu & Mooney, 2007; Settles, Craven, & Ray, 2008; Wang, Nie, & Huang, 2011; Doran & Ray, 2014; Cheplygina, Tax, & Loog, 2015; Zhu, Wu, Xu, Chang, & Tu, 2015). In particular, MIL was applied to information extraction (Hoffmann, Zhang, Ling, Zettlemoyer, & Weld, 2011) and relation extraction (Surdeanu, Tibshirani, Nallapati, & Manning, 2012; Xu, Hoffmann, Zhao, & Grishman, 2013). However, fewer studies focused on regression, which we now discuss.

Multiple-instance regression (MIR) belongs to the class of MIL problems with real-valued output, and is a variant of multiple regression where each data point may be described by more than one vector of values. MIR was firstly introduced by Ray and Page (2001), who proposed an EM algorithm which assumes that one primary instance per bag is responsible for its label. Wagstaff and Lane (2007) proposed to simultaneously learn a regression model and to estimate instance weights per bag for crop yield modeling, but their method is not applicable to prediction. A similar method which learns the internal structure of bags using clustering was later proposed by Wagstaff, Lane, and Roper (2008) for crop yield prediction, and we will compare to it in Section 7. Later, the method was adapted to map bags into a single-instance feature space by Zhang and Zhou (2009). Different assumptions were made in other studies: Wang, Radosavljevic, Han, Obradovic, and Vucetic (2008) assumed that each bag is generated by random noise around a primary instance, while Wang, Lan, and Vucetic (2012) represented bag labels with a probabilistic mixture model. The main disadvantage of the above methods for text regression tasks is that they do not scale well to high-dimensional feature spaces, and that some of them are not applicable to prediction.

Attempts to apply MIR to document analysis have concerned news categorization (Zhang & Zhou, 2008; Zhou, Sun, & Li, 2009) or web-index recommendation (Zhou, Jiang, & Li, 2005). Kotzias, Denil, de Freitas, and Smyth (2015) combined MIL with deep learning features and applied it to sentiment prediction with the goal of transferring label information from group labels (review) to instance labels (sentences). However, their study focused solely on binary sentiments rather than gradual ones, and did not take into account the sentiments towards aspects as we do here. To the best of our knowledge, no previous study has attempted to use MIR for text regression tasks with real-valued labels such as aspect rating prediction, sentiment and emotion prediction, as we do here.

3. Problem Definition: Weakly Supervised Text Regression

We consider as input data D a set B of m reviews accompanied by numerical labels Y (represented in Fig. 1). Each review B_i is a bag of n_i sentences (i.e. instances), where each sentence is a d-dimensional vector, typically a vector of words. Hence, a review is noted as $B_i = \{b_{ij}\}_{n_i}^{d}$ with $b_{ij} \in \mathbb{R}^d$ for $1 \leq j \leq n_i$. If there are k sentiment aspects rated for each review, the labels of all reviews are noted as $Y = \{y_i\}_m^k$ with $y_i \in \mathbb{R}^k$. Therefore, the dataset D is more precisely noted as $D = \{(\{b_{1j}\}_{n_1}^{d}, y_1), ..., (\{b_{mj}\}_{n_m}^{d}, y_m)\}$.

The challenge is to infer the label of new bags, which do not have a constant number of instances. This requires finding a set of bag representations $X = \{x_1, \ldots, x_m\}$ with $x_i \in \mathbb{R}^d$

to train a regression model, from which the class labels of new bags can then be inferred. In other words, we propose to represent each review in the feature space as one exemplar vector per aspect class, obtained from the convex combination of its instances. This requires computing the sentence relevance to an aspect rating k, called aspect saliency and noted $\Psi = \{\psi_{ijk}\}_{n_i}^1$, with unit constraints per review. The goal is then to find a mapping, noted $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$, from this type of representation to numeric values, which is able to predict the label of a given bag. To obtain this mapping, if we assume that X is the optimal bag representation for our task, then we need to look for the optimal regression hyperplane Φ which minimizes a loss function $\mathcal{L}$ plus a regularization term Ω, as follows:

$$\Phi = \underset{\Phi}{arg\,min} \left(\underbrace{\mathcal{L}(Y, X, \Phi)}_{\text{loss}} + \underbrace{\Omega(\Phi)}_{\text{reg.}} \right) \tag{1}$$

However, the best set of representations X for a task is generally unknown. Therefore, one has to make assumptions about how to compose them through a fixed function of B, or a parametrized function of B learned jointly with the regression hyperplane Φ. Such assumptions have a strong impact on the learning performance. Three main assumptions have been made in the past: aggregating all instances, keeping them as separate examples, or choosing the most representative one, as defined below. As for the regression, noted as f, several state-of-the-art models can be used, and we will present below results using Support Vector Regression (Drucker, Burges, Kaufman, Smola, & Vapnik, 1996) and Lasso (Tibshirani, 1996) with respectively the ℓ_2 and ℓ_1 norms for regularization.

Aggregated Instances. Each bag is represented as a single d-dimensional vector, which is the average of its instances. Then, a regression model f is trained on pairs of vectors and class labels, $D_{agg} = \{(x_i, y_i) \mid i = 1, \ldots, m\}$, and the predicted class of an unlabeled bag $B_i = \{b_{ij} \mid j = 1, \ldots, n_i\}$ is computed as follows:

$$\hat{y}(B_i) = f(mean(\{b_{ij} \mid j = 1, \ldots, n_i\})) \tag{2}$$

A sum can be used instead of the mean, and we observed that with appropriate regularization the two have similar prediction performance. This baseline corresponds to the typical approach for text regression tasks, where each text sample is represented by a single vector in the feature space, e.g. bag-of-words (BOW) with counts or with TF-IDF weights.

Instance as Example. Each of the instances in a bag is considered as a separate example, with its bag's label. A regression model f is learned over the training set made of all vectors of all bags, $D_{ins} = \{(b_{ij}, y_i) \mid i = 1, \ldots, m; j = 1, \ldots, n_i\}$. To label a new bag B_i, the predicted labels of the instances are averaged:

$$\hat{y}(B_i) = mean(\{f(b_{ij}) \mid j = 1, \ldots, n_i\}) \tag{3}$$

The median can be used instead of the average, especially when the bags contain outliers.

Primary Instance. This assumption considers that a single instance in each bag, called primary or prime, is responsible for the bag's label (Ray & Page, 2001). The approach is similar to the previous one, except that only one instance per bag is used as training data: $D_{pri} = \{(b_i^p, y_i) \mid i = 1, \ldots, m\}$, where b_i^p is the prime instance of the i^th bag B_i. The prime instances are discovered through an iterative algorithm which refines the regression model f. The class of a new bag is computed as in Eq. 3 above.

The main drawbacks of the previous instance relevance methods that we address below are the following ones: (1) prohibitive complexity for high dimensional feature spaces, which are common in text regression tasks; (2) inefficiency or inability to estimate the importance of the instances of unseen bags; and (3) lack of modeling of weights with sparsity, which is needed for different types of data.

4. Proposed Model: Weighted Multiple Instance Regression

The new MIR model that we propose assigns individual relevance values (weights) for each instance of a bag to model the input structure, considering that it is the weighted combination of instances that is responsible for the label of the bag. Building upon the existing 'Instance Relevance' assumption (Wagstaff & Lane, 2007; Wagstaff et al., 2008; Wang et al., 2011), we address its shortcomings presented in Section 2.3 and those above, and make it applicable to document modeling as follows.

To jointly learn instance weights and target labels efficiently, we minimize a Regularized Least Squares loss (RLS), which enables us to support high-dimensional feature spaces, required for text regression tasks. In addition, our model is able to predict for previously unseen bags both their class label and the contribution of each instance to the bag's predicted label. Our model learns an optimal method to aggregate instances directly from the training data and allows more degrees of freedom in the regression model than previous proposals. Essentially, the weight of an instance can be interpreted as its relevance both in training and prediction, thus enabling an explicit modeling of documents.

4.1 Modeling Instance Relevance and Bag Labels

Each bag of instances defines a bounded region of a hyperplane orthogonal to the y-axis, namely the envelope of all its points. The goal is to find a regression hyperplane that passes through each bag B_i and to predict its label by using at least one data point x_i within that bounded region. Thus, the point x_i is a convex combination of the points in the bag, in other words B_i is represented by the weighted average of its instances b_{ij}:

$$x_i = \sum_{j=1}^{n_i} \psi_{ij} b_{ij} \text{ with } 1 \geq \psi_{ij} \geq 0 \text{ and } \sum_{j=1}^{n_i} \psi_{ij} = 1 \tag{4}$$

Here, ψ_{ij} is the weight of the j^{th} instance of the i^{th} bag, and indicates the saliency or relevance of an instance j to the prediction of the class y_i of the i^{th} bag. The first constraints force x_i to fall within the bounded region of the points in bag i and guarantee that the i^{th} bag will influence the regressor.

We propose the following method for learning to predict the instance weights and at the same time the target labels of bags. Initially, the n_i instance weights associated to each bag B_i, noted $\psi_i = \{\psi_{ij}\}_{n_i}^1$ with $\psi_{ij} \in [0, 1]$ are unknown. We aim for a linear regression model f that is able to model each target value y_i for each bag using regression coefficients $\Phi \in \mathbb{R}^d$, that is, $Y = f(X) = \Phi^T X$, where X and Y are respectively the sets of training bag representations and their labels. Note that the coefficients Φ, the weights Ψ and representations X are different for each aspect. For brevity, we formulate the model for one aspect only ($k = 1$).

We define a loss function according to a least squares objective dependent on B, Y, Φ and the set of weight vectors $\Psi = \{\psi_1, \ldots, \psi_m\}$ using Eq. 4 as follows:

$$\mathcal{L}(Y, B, \Psi, \Phi) = ||Y - \Phi^T X||_2^2 \overset{(4)}{=} \sum_{i=1}^{N} \Big(y_i - \Phi^T \big(\sum_{j=1}^{n_i} \psi_{ij} b_{ij}\big)\Big)^2 = \sum_{i=1}^{N} \Big(y_i - \Phi^T (B_i \psi_i)\Big)^2 \quad (5)$$

Using $\mathcal{L}$ and assuming ℓ_2-norm for regularization with ϵ_1 and ϵ_2 terms for $\psi_i \in \Psi$ and Φ respectively, we transform the objective from Eq. 1 into the least squares objective shown in the following equation, subject to $\psi_{ij} \geq 0 \ \forall i, j$ and $\sum_{j=1}^{n_i} \psi_{ij} = 1 \ \forall i$.

$$\psi_1, \ldots, \psi_m, \Phi = \underset{\psi_1, \ldots, \psi_m, \Phi}{arg\,min} \underbrace{\sum_{i=1}^{m} \Bigg(\underbrace{\Big(y_i - \Phi^T (B_i \psi_i)\Big)^2}_{f_1 \text{ loss}} + \underbrace{\epsilon_1 ||\psi_i||}_{f_1 \text{ reg.}} \Bigg)}_{f_2 \text{ loss}} + \underbrace{\epsilon_2 ||\Phi||^2}_{f_2 \text{ reg.}} \quad (6)$$

The selection of the ℓ_2-norm was based on preliminary results which demonstrated that it led to a more accurate function than the ℓ_1-norm, but other combinations of p-norm regularizations can be explored for f_1 and f_2, e.g. to control the sparsity of instance weights and regression coefficients.

The above objective is non-convex and difficult to optimize because the minimization is with respect to all $\psi_1, \ldots, \psi_m$ and Φ at the same time. One solution is to divide it in two parts. If we note f_1 the model that is obtained from the minimization with respect to $\psi_1, \ldots, \psi_m$ only, and f_2 the model obtained from the minimization with respect to Φ only, then we observe that if one of the two functions is known or held fixed, then the other one is convex and can be learned with well-known least squares solving techniques. Alternatively, we propose an efficient to way to solve them jointly in the next section.

Having computed $\psi_1, \ldots, \psi_m$ and Φ for the training data, we could predict a label for a new bag from the test data using Eq. 3 by taking the average of instance predictions, but then we would not be able to compute the weights of its instances. Moreover, information that has been learned about the instances during the training phase would not be used during prediction. The solution for predicting instance weights, presented in the next section, is to learn a function parametrized by coefficients $O \in \mathbb{R}^d$, which is able to map a given instance $b_{ij} \in \mathbb{R}^d$ to a real-valued weight in the [0,1] interval and satisfies the constraints of Eq. 4.

4.2 Learning the Weights Jointly with Minibatch Stochastic Gradient Descent

One straightforward way to learn to predict instance weights is through a third regression model f_3 with coefficients $O \in \mathbb{R}^d$, which we presented in an earlier study (Pappas & Popescu-Belis, 2014, Section 5). The model was trained on the instance weights obtained from Eq. 6, noted $D_w = \{(b_{ij}, \psi_{ij}) \mid i = 1, ..., m; j = 1, ..., n_i\}$. The objective from Eq. 6 was minimized consecutively with the following one:

$$O = \underset{O}{arg\,min} \underbrace{\sum_{i=1}^{N} \sum_{j=1}^{n_i} \big(\psi_{ij} - O^T b_{ij}\big)^2}_{f_3 \text{ loss function}} + \underbrace{\epsilon_3 ||O||^2}_{f_3 \text{ reg.}} \quad (7)$$

The learned model was able to estimate the weights of the instances of an unlabeled bag B_i during prediction time as: $\hat{\psi}_i = f_3(B_i) = O^T B_i'$. Once normalized, these weights allowed us to compute the predicted label for an unseen bag B_i as $\hat{y}_i = f_2(B_i) = \Phi^T B_i \hat{\psi}_i$. To solve the non-convex optimization problem of Eq. 6 and 7, we proposed to use alternating projections (AP), a powerful method for finding the intersection of convex sets (Bauschke & Borwein, 1996; Lewis & Malick, 2008). The algorithm, called APWeights, first optimized the parameters of f_2 and f_1 from Eq. 6 in an alternating fashion, and then learned the parameters of f_3 based on the learned weights by f_2.

We propose here a new algorithm, called SGDWeights, which optimizes jointly the two consecutive objectives from Eq. 6 and 7 above. This algorithm addresses two limitations of APWeights: the redundancy of parameters when learning the instance-specific weights ψ_i per bag and the global coefficients O; and the fact that the instance weights function (f_3) and the functions of Eq. 6 (f_1 and f_2) are not influencing each other during training. The SGDWeights model merges the two objectives of APWeights into a compact form by using only the O and Φ parameters, as the Ψ parameters have been replaced with a function. The model is thus more general, unified, and self-contained than APWeights. The new objective consists of differentiable functions, which can be optimized with minibatch stochastic gradient descent (SGD).

Firstly, we replace the ψ parameters in Eq. 6 with a normalized exponential function, namely softmax, as follows:

$$\sigma(B_i, O) = P(\psi = y_i | B_i) = \frac{\exp^{(O^T B_i)}}{\sum_{k=1}^{n_i} \exp^{(O^T B_{ik})}} \tag{8}$$

This function satisfies the constraints of Eq. 6 and can be seen as a probability distribution of the weights to be learned ψ_i per bag. Moreover, it naturally derives from the weighted multiple-instance learning framework (Pappas & Popescu-Belis, 2014), and is essentially equivalent to the attention formula used later by Bahdanau et al. (2015), except that B_i has to be input to an additional non-linear hidden layer.

Secondly, we multiply the output of this function with the bag B_i and map it to the target label through coefficients Φ using a least squares objective, as before. Hence, we obtain an objective with two mutually-influencing functions: one function for estimating the instance weights per bag with non-negativity and unit constraints, and another function to model the target labels:

$$O, \Phi = \underset{O,\Phi}{arg\,min} \sum_{i=1}^{m} (y_i - \Phi^T (B_i \cdot \sigma(B_i, O)))^2 + \Omega(\Phi, O) \tag{9}$$

The model is regularized with Ω, which can support any kind of regularization for the coefficients O and Φ, for instance as in the objective of Eq 6.

The algorithm for SGDWeights (Algorithm 1) takes as input the set of bags B_i with known labels y_i and the regularization terms ϵ_1 and ϵ_2. The algorithm uses penalized ℓ_2 regularization for O and Φ.[1] First, the algorithm iterates through each batch of a predefined size (set to 50 here) in the training data and accumulates the gradient error of each example

1. In the examined datasets, after optimizing for the learning rate, we observed that the values of the regularization terms did not affect much the performance, hence we kept $\epsilon_1$and ϵ_2 equal to 1.

Algorithm 1: SGDWeights: jointly learning the parameters of the objective in Eq. 6.

Data: Reviews $B = \{b_{ij}\}_{n_i}^{m}$, Ratings $Y = \{Y_i\}^m$
Result: Parameters Φ, O
set(*max_iter*, *tolerance*, ϵ_1, ϵ_2, α, c) # a is the learning rate and c its decay
initialize(Φ, O) # Initialize model parameters
while *not converged* **do**
  for *batch in B* **do**
    # Accumulate gradient values Φ_g, O_g in the batch
    for B_i *in batch* **do**
      $\Phi_g = \Phi_g + \frac{\partial}{\partial \Phi_g} L(\Phi, B_i, Y_i)$
      $O_g = O_g + \frac{\partial}{\partial O_g} L(W, B_i, Y_i)$
    end
    # Perform gradient step on parameters Φ, O
    $\Phi = \Phi - \alpha(\frac{\Phi_g}{|batch|} + \epsilon_1 ||\Phi||)$; $O = O - \alpha\big(\frac{O_g}{|batch|} + \epsilon_2 ||O||\big)$
  end
  $e = \frac{1}{m}\sum_i (Y_i - \Phi^T B_i \hat{\psi}_i)$ # Mean absolute error
  if $e_{prev} - e <$ *tolerance or iter* $>$ *max_iter* **then**
    converged = True
  end
  $\alpha = \alpha - c$ # Learning rate decay
end

with respect to O and Φ. Then, it updates the parameters according to the accumulated gradient error and the regularization term. After all batches have been visited (one epoch), the training error of the learned model is computed to check for convergence. This repeats until convergence, i.e. when the decrease of the training error is lower than a predefined value (10^{-6}) or until a maximum number of iterations is reached. Algorithm 1 shows a simple version of minibatch SGD with learning rate decay; however, the performance and convergence time can be improved by using more sophisticated versions such as ADAGRAD (Duchi, Hazan, & Singer, 2011) or ADAM (Kingma & Ba, 2014). In practice, we selected the version based on time and scores on training sets.

The time complexity T_{SGD} of SGDWeights is computed in Eq. 10 depending on the input variables, noted $h = \{m, \hat{n}, d, \hat{b}\}$, where m is the number of bags, $\hat{n}$ the average size of the bags, d the vocabulary size (dimensions of feature space), and $\hat{b}$ the minibatch size:

$$T_{\text{SGD}}(h) = \mathcal{O}\big(\frac{m}{\hat{b}}(\hat{b}((\hat{n}d + d^2) + (\hat{n}d + \hat{n}^2 + d^2)))\big) = \mathcal{O}\big(m(\hat{n}d + d^2 + \hat{n}^2)\big) \leq T_{\text{AP}}(h) \quad (10)$$

The complexity of SGD is lower than that of APWeights (Pappas & Popescu-Belis, 2014, Eq. 9), thanks to the joint formulation compared to the consecutive one. Another advantage of SGDWeights is that SGD update is independent from the number of examples and can further benefit from parallelization across multiple cores (Zinkevich, Weimer, Smola, & Li, 2011). We also confirm experimentally in Section 7.3.3 that SGDWeights scales better than APWeights in terms of execution time and achieves a similar or smaller prediction error.

The complexity of T_{SGD} which performs segmentation and rating prediction jointly is only $\mathcal{O}(d^2)$ more expensive (assuming that $\hat{n} << m$) than that of previous methods for aspect-based sentiment analysis, namely $\mathcal{O}(md)$, as they rely on Support Vector Machines as in (McAuley et al., 2012). In practice, their efficiency is lower than T_{SGD} if we take into account their segmentation model which requires an extra dataset pass, $\mathcal{O}(m\hat{n}d)$, and their structured objective for the SVM.

5. Learning to Segment and Summarize Reviews

The proposed weighted MIR model is able to predict aspect ratings, but can also estimate the saliency of sentences with respect to each aspect. The second capability is learned in an weakly-supervised manner, i.e. without knowledge of which aspect is being discussed in each sentence of the training bags. The aspect saliency of each sentence can be used to improve aspect-based text segmentation and summarization in several ways.

These two tasks are defined here along the same lines, in terms of data annotations and metrics, as in previous work (McAuley et al., 2012). *Aspect-based segmentation* of a review simply means to assign to each of its sentences a single aspect label from the list of aspects. *Aspect-based summarization* differs from traditional extractive summarization, as its goal is to select for each review exactly one sentence per aspect, no matter how representative it is. What is then evaluated is the accuracy of labels over the selected subset of sentences, i.e. how many of them receive receive a correct aspect label.

To solve the two tasks, the aspect label of a sentence could be directly estimated by its maximum aspect saliency score found by the MIR model, that is, $\arg\max_a \hat{\psi}_i^{(a)}$, where $\hat{\psi}_i^{(a)}$ is the saliency of sentence i for aspect a. While this addresses the unsupervised version of the tasks, in a supervised setting we can combine the saliency values with Conditional Random Fields (CRFs), trained using n-gram bag-of-words (BOW) of simple features (binary counts, TF-IDF) or of more sophisticated ones (POS tags, Wordnet synsets, and others). In addition to BOW of counts or TF-IDF values, we define and use here new contextual features based on sentence-level aspect saliency and sentiment predicted by our MIR model.

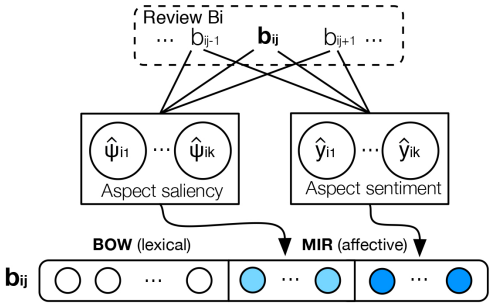
5.1 Linear Chain CRF

Given an input document $b \in B$ composed of T consecutive sentences b_t, and a set of output variables $\bar{y} \in \mathcal{Y}$, with $\bar{y}$ taking categorical values from $A = \{1, \ldots, k\}$ (the aspects to which sentences can refer), a first-order Linear Chain CRF models the conditional distribution $p(\bar{y}|b)$ as a globally normalized log-linear distribution (Lafferty, McCallum, & Pereira, 2001):

$$p(\bar{y}|b) \propto \prod_{t=1}^{T} h(\bar{y}_t, b_t) \prod_{t=1}^{T-1} h_{\curvearrowright}(\bar{y}_t, \bar{y}_{t+1}) \tag{11}$$

where $h(\bar{y}_t, b_t) = exp(w \cdot f(\bar{y}_t, b_t))$ models the aspect label of the sentence at position t by means of a parameter vector w and a feature vector $f(\bar{y}_t, b_t)$, and $h_{\curvearrowright}(\bar{y}_t, \bar{y}_{t+1})$ models the transition between the aspect variables at positions t and $t+1$ in the linear chain. Hence, this model accounts for sequential dependencies of aspects in the context of each review.

We consider three CRF models of increasing complexity: (1) a CRF without pairwise potential, which is equivalent to logistic regression, hence noted LogReg; (2) a linear chain



Sentences	Weights ($\hat{\psi}$)	Sentiment ($\hat{y}$)
current (b_{ij})	sw	sp
prev (b_{ij-1})	sw_prev	sp_prev
next (b_{ij+1})	sw_next	sp_next
all (B_i)	cw	cp

Figure 2: Combination of BOW with aspect saliency and sentiment features from MIR.

CRF with symmetric pairwise potential, noted CRF-u, using $(k^2+k)/2$ variables; and (3) a linear chain CRF with asymmetric pairwise potential, noted CRF-d, using k^2 variables.

5.2 Aspect Saliency, Sentiment, and Diversification

The typical BOW features lack affective information such as the saliency and the sentiment of the sentences towards an aspect, which could be useful when deciding which aspect a sentence refers to. We propose to use the aspect saliency (weight estimate $\hat{\psi}_a$ for an aspect $a \in A$) and the sentiment information ($\hat{y}_a$) from our weighted MIR model to augment the word or topic feature spaces for each sentence, as represented on the left side of Figure 2. The current sentence can be augmented with its own MIR features, noted respectively as sw and sp for aspect saliency and sentiment (see table in Fig. 2), but also with features from the previous or following sentences, or even with review-level features such as the average sentence saliency cw or aspect rating cp.

McAuley et al. (2012) observed that sentences from different aspects tended to receive the same label when their aspects had close meanings (e.g. palate and taste of a beer), and reduced this unwanted effect by enforcing diversity on the predicted output, i.e. enforcing that each aspect is assigned at least once in each review. We add this constraint to our model as follows. We build a bipartite graph for each review B_i, which maps its n_i sentences to n_i aspects. We define the compatibility between a sentence s and an aspect a as either its aspect saliency estimate from the MIR model for the unsupervised task (i.e. $c_{sa} = \hat{\psi}_{is} = f_3(b_{is})$) or as the CRF probability estimate for the supervised task (i.e. $c_{sa} = p(s|a)$). The graph edges are then defined as:

$$E_{s,l}^{(B_i)} = \begin{cases} c_{sl} & \text{if } 1 \leq l \leq k \\ max_a\ c_{sa} & \text{otherwise} \end{cases} \tag{12}$$

The first part enforces each of the k aspects to have a matching sentence (for summarization), while the second one allows the remaining sentences to match any aspect (for segmentation). The optimal assignment of sentences to aspects is found using the Hungarian algorithm (Kuhn, 1955; Munkres, 1957).

Dataset	Reviews	Sentences	Features	Rating Labels
BeerAdvocate	1,586,259	16,883,058	19,418	feel, look, overall, smell, taste
Toys & Games	373,974	2,105,647	31,984	educational, durability, fun, overall
Audiobooks	10,989	44,487	3,971	performance, story, overall
RateBeer (FR)	17,998	105,569	903	appearance, aroma, overall, palate, taste
RateBeer (ES)	1,259	3,511	2,120	appearance, aroma, overall, palate, taste
TED comments	1,200	3,814	957	polarity
TED talks	1,203	12,023	5,000	unconvincing, fascinating, persuasive, funny, ingenious, longwinded, inspiring, courageous, jaw-dropping, beautiful, confusing, obnoxious

Table 1: The datasets with aspect, sentiment and emotion ratings used in our experiments.

6. Datasets Used in our Evaluation

We use eight public datasets (Table 1). Five of them were built for aspect prediction by McAuley et al. (2012) (see 6.1), and subsets of them have aspect-based segmentation (see 6.3). We also use book reviews from Audible with human attention scores (see 6.2).

6.1 Multi-aspect Sentiment Analysis

The BeerAdvocate, Ratebeer (ES), RateBeer (FR), Audiobooks and Toys & Games datasets include aspect ratings assigned by the authors of the reviews, with 3 to 5 aspect dimensions. Furthermore, on the TED talks that we gathered and released earlier (Pappas & Popescu-Belis, 2013),[2] we aim to predict the 12-dimensional talk-level emotion ratings assigned by viewers through voting, as well as the polarity of comments assigned by external judges (one dimension). In all datasets, the instances are the sentences of reviews or comments, except for the TED emotion rating dataset where the instances are comments written by users below each talk, of which we keep the ten most recent ones per talk. Six of the datasets are in English, one is in Spanish (Ratebeer) and one in French (RateBeer); therefore, our results also illustrate the language-independence of our methods.

The features for each of the datasets are vectors of words with binary attributes signaling word presence or absence, in a traditional bag-of-words model (BOW). These word vectors were provided with the first six datasets, and we generated them for the latter two, after lowercasing and stopword removal. To experiment with different feature spaces, for TED comments, we also computed TF-IDF scores using the same dimensionality as BOW. Moreover, we computed sentence features based on 300-dimensional word embeddings trained on Wikipedia with word2vec (Mikolov et al., 2013).[3] Specifically, we adopted the concatenation of max, min and average pooling (900 dimensions, hence about the same size as BOW) of the embeddings of words belonging to a given sentence, an approach that has been shown to work well in practice (Tang et al., 2014).

The target class labels were normalized by the maximum rating in their scale, except for the 12 emotions of the TED talks, for which the votes were normalized by the maximum

2. Available at `https://www.idiap.ch/dataset/ted/`.
3. Available at `https://code.google.com/archive/p/word2vec/`.

number of votes over all the emotion classes for each talk. Two emotions from the original data, labeled as 'informative' and 'ok', were excluded as they are neutral ones.

6.2 Human Attention Prediction

The definition of the summarization task given in Section 5 above does not take into account whether the summary of a review actually reflects its sentiment towards the corresponding aspect. This is because sentences annotated with aspects are not necessarily opinionated and do not necessarily justify the actual aspect *ratings* of the review. To our knowledge, no existing dataset captures human attention in document classification. In other words, to evaluate the full potential of our method, and of many other attention-based models that have been recently proposed in NLP, we need an annotation that targets specifically the opinionated sentences and how they contribute to their review's aspect rating. Mere facts about an aspect not necessarily relevant to its rating.

Therefore, for the intrinsic evaluation of the MIR weights, i.e. of our instance relevance mechanism, we designed a new dataset called HATDOC (Pappas & Popescu-Belis, 2016).[4] We obtained human judgments over a set of 100 reviews of audiobooks collected from www.audible.com, by crowdsourcing the task via www.crowdflower.com. Each review is accompanied by ratings of three aspects of the audiobook: the story (plot of the book), the performance (i.e. acting of the voice(s) heard on the recording), and the overall appreciation. All three aspects are rated by the author of each review on a five-point scale (one to five stars). We collected 100 reviews with 1,662 sentences by sampling 20 reviews for each rating value of the 'overall' aspect.

The human judges were asked to provide labels which represent the explanatory value of each sentence with respect to the aspect rating of a given review, which on a five-point scale labeled as: "Not at all","A little", "Moderately", "Rather well", or "Very well". The reliability of the judges was controlled by randomly inserting questions with known answers (known as "gold questions") among the series of questions. We only kept judges with more than 70% accuracy on these questions. For each non-gold question, we collected at least 4 answers, for a total of 7,121 judgments. The ground-truth label per sentence for each aspect was the response of the majority of judges, or, in case of a draw, the response with the highest confidence score, as computed by the platform from the annotators' reliability levels. We acknowledge, however, that other aggregation strategies may provide better insights, as future work may show.

The overall confidence of the annotations, as computed by Crowdflower, was 59%. Specifically, it was 57% for the 'overall' and 'story' aspects, and 63% for 'performance'. The percentages of sentences with a confidence higher than or equal to 0.8 were, respectively, 4%, 7% and 12% for each aspect. These values suggest that the task was the most difficult for the 'overall' aspect, followed by the 'story' and 'performance' aspects. For evaluation, we will use judgments from all the confidence levels, and we will compare them with the saliency values assigned automatically by our model.

While a full analysis of the HATDOC dataset is beyond our scope, one property must be emphasized: when allowed to rate the relevance of sentences to various aspects independently, humans often consider that sentences are relevant to more than one aspect at a time.

4. We make this dataset available at `https://www.idiap.ch/paper/hatdoc/`.

Specifically, we found that only 9% of the sentences were relevant to a single aspect rating (more than "a little") and irrelevant to the two others ("not at all", on the scale presented above). Therefore, sentence saliency weights appear to represent better the opinionated content of a review than simple aspect labeling, and therefore this approach represents an innovative way to evaluate sentiment summarization.

6.3 Review Segmentation and Summarization

For the extrinsic evaluation of the MIR weights, we use six datasets annotated with segmentation labels from McAuley et al. (2012), namely: BeerAdvocate (992 reviews, 8,399 sentences), Pubs (100 reviews, 981 sentences), Toys & Games (101 reviews, 510 sentences), Audiobooks (95 reviews, 439 sentences), RateBeer (FR) (57 reviews, 279 sentences) and RateBeer (ES) (115 reviews, 319 sentences). All these datasets, except Pubs, are subsets of those presented in Table 1 of Section 6.1 above. As McAuley et al. (2012) only used the Pubs dataset for segmentation and summarization (although it contains aspect ratings), we do the same here for comparison. The sentences of the reviews are accompanied by categorical aspect labels, which were annotated by humans through crowdsourcing. In other words, the labels represent the particular aspect of an item discussed by each sentence. The baseline features are bag-of-words (BOW) provided with the datasets. For segmentation and summarization, the sentence vectors were normalized with ℓ_2, so that the MIR features obtained in Section 5.2 fall in the same range as the BOW features.

7. Evaluation on Multi-aspect Sentiment Analysis

In this section, after specifying the setup of our experiments (7.1), we present the results of the proposed model for multi-aspect sentiment analysis (7.2), including a comparison with neural networks. Next, we investigate the effects of our design choices: the MIR assumptions (7.3.1), the input feature space (7.3.2), and the learning algorithm (7.3.3).

7.1 Experimental Settings

To compare with the state of the art in aspect rating prediction from texts, we experiment with the five multi-aspect datasets provided by McAuley et al. (2012). We use the same protocol as McAuley et al., i.e. a uniform split of the data into 50% for training and 50% for testing. We compare with their structured and unstructured methods, both over segmented text, i.e. sentences that refer to the target aspect, and over unsegmented text, i.e. all the sentences of a review. Specifically, we compare our proposal to the following three state-of-the-art methods: (1) *Support Vector Machine (SVM)*, the well-known maximum margin classifier (Cortes & Vapnik, 1995); (2) *Structured-SVM*, a generalization of SVM for general structured output labels (Tsochantaridis, Joachims, Hofmann, & Altun, 2005), which learns the relationships between aspect ratings, known as a 'rating model'; and (3) *PALE LAGER*, the probabilistic model proposed by McAuley et al. (2012) which includes a structured SVM and achieves state-of-the-art performance on the tasks we consider here.

All the models are optimized (when applicable) on a development set, i.e. a 25% subset of the training data, through exhaustive grid-search over a fine-grained range of possible values. The hyper-parameters to optimize for the various MIR assumptions are the regularization

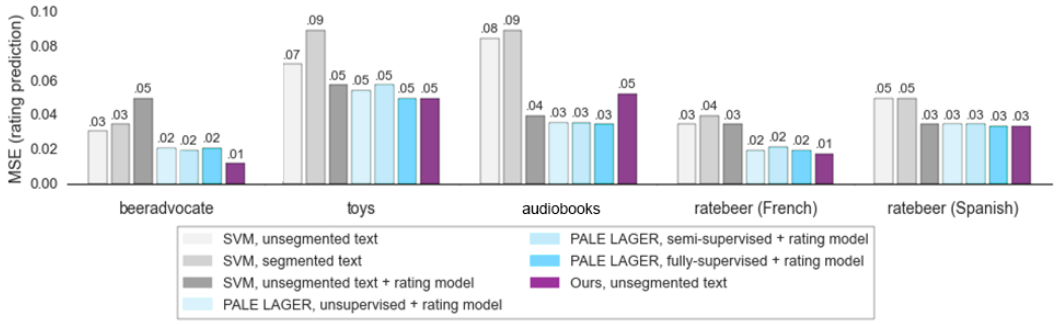


Figure 3: Mean squared error (MSE) on aspect rating prediction for our MIR model with unsegmented text, compared to structured or unstructured supervised baselines, with unsegmented or segmented text over five multi-aspect datasets. The scores of SVM and PALE LAGER are from McAuley et al. (2012).

terms λ_2 and λ_1 of their regression model f. As stated in Section 3, we experiment with two regression models: Support Vector Regression (SVR), which uses the ℓ_2 norm, and Lasso, which uses the ℓ_1 norm. The hyper-parameters to optimize for APWeights are the three regularization terms $\epsilon_1, \epsilon_2, \epsilon_3$ of the ℓ_2-norm for the f_1, f_2 and f_3 regression models. Lastly, for the Clustering MIR assumption (Wagstaff et al., 2008), we use the f_2 regression model, which relies on ϵ_2 and the number of clusters k, optimized over $\{5, ..., 50\}$ with step 5, for its clustering algorithm, which is here the k-Means one. All the regularization terms are optimized over the same range of possible values, noted $a \cdot 10^b$ with $a \in \{1, \ldots, 9\}$ and $b \in \{-4, \ldots, +4\}$, hence 81 values per term. The hyper-parameters for SGDWeights are the same ones as for APWeights, plus the learning rate or step size ϵ, the minibatch size m, and the gradient step strategy (learning rate decay, ADAGRAD, or ADAM). For the regression models and evaluation, we use the *scikit-learn* library (Pedregosa et al., 2012).[5]

We report standard error metrics for regression, namely the Mean Absolute Error (MAE) and the Mean Squared Error (MSE), which are defined over a test set of bags B_i respectively as $\text{MAE} = (\sum_{i=1}^{k} |f(B_i) - y_i|)/k$ and $\text{MSE} = (\sum_{i=1}^{k} (f(B_i) - y_i)^2)/k$.

7.2 Comparison with State-of-the-Art Systems

Figure 3 displays the performance of the proposed model for aspect rating prediction. Our MIR model outperforms all other models over all datasets except Audiobooks. When compared to the best SVM baseline (with unsegmented text) provided by McAuley et al., our method improves by 58% on BeerAdvocate, 40% on Toys, 60% on Audiobooks, 94% on Ratebeer (FR) and 47% on Ratebeer (ES), despite the fact that it does not use their rating model. McAuley et al. report MSE scores of 2%, 5%, 3%, 2% and 3% respectively on the above data sets for their best model, which uses a joint rating model and an aspect-specific text segmenter trained on hand-labeled data. These scores are comparable to those of our MIR model (1%, 5%, 5%, 1%, and 3%), which does not use these features.

5. Our code is available at `https://github.com/idiap/wmil-sgd`.

Methods	Vocabulary	d_{hidden}	Depth	$\|\theta\|$	MSE
SVM (Lei et al., 2016)	bigram (>147k)	-	-	2.5M	0.0154
MIR (this work)	unigram (19k)	-	-	38k	**0.0115**
Dense (Rumelhart et al., 1986)	unigram (19k)	200	1	41.2k	0.0101
LSTM (Hochreiter et al, 1997)	unigram (147k)	200	2	644k	0.0094
GRU (Chung et al., 2014)	unigram (19k)	200	1	241.6k	**0.0079**
RCNN (Lei et al., 2016)	unigram (147k)	200	2	323k	0.0087
Dense+MIR (this work)	unigram (19k)	200	1	41.4k	0.0091
GRU+MIR (this work)	unigram (19k)	200	1	241.8k	**0.0078**

Table 2: Comparison of our instance relevance mechanism (MIR) integrated within neural networks, with state-of-the-art neural networks, on the aspect rating prediction task in terms of mean squared error (MSE). $|\theta|$ indicates the number of parameters.

These results indicate that by modeling the saliency of sentences for aspect ratings, even with unsegmented text only, the MIR model outperforms even complex baselines, including structured models (Structured-SVM), which make use of either unsegmented or segmented text (PALE LAGER). This means that, unlike PALE LAGER which requires a separate segmentation procedure to achieve the best scores, MIR does not require human intervention to perform as well or even better. Hence, MIR is able to learn structural information without explicitly modeling the relationship between aspect ratings. The structured models are more successful than MIR only on Audiobooks, likely because this is the dataset with the fewest number of aspects. Lastly, as stated by McAuley et al. (2012), predictors which use segmented text, for example with topic models as studied by Lu et al. (2011), do not reliably outperform SVR baselines, as they have marginal or even no improvements. For this reason, we did not further experiment with them.

Recently, Lei et al. (2016) proposed a recurrent convolutional neural network (RCNN) for multi-aspect sentiment analysis and compared it to LSTM (Hochreiter & Schmidhuber, 1997). To compare our proposal with theirs, we integrated our method (MIR) as a parametrized pooling layer into two types of neural networks at the word-level, treating a document as a sequence of word embeddings: (1) a fully-connected layer noted as Dense (Rumelhart, Hinton, & Williams, 1986), and (2) Gated Recurrent Units noted as GRU (Chung, Gülcehre, Cho, & Bengio, 2014), using ReLUs for all activations (Nair & Hinton, 2010). We use the same evaluation protocol, word embeddings and dataset as Lei et al. (2016), but we use the pre-processed BeerAdvocate dataset provided by McAuley et al. (2012), with a vocabulary of 19k words, while Lei et al. used their own pre-processing, with a vocabulary of 147k unigrams.

Table 2 displays the mean squared error (MSE) on a test set with 1,000 reviews from BeerAdvocate, using 260k reviews for training. In terms of non-neural methods, the first two lines show that our MIR model with unigram features largely outperforms SVM with bigram features as evaluated by Lei et al. (2016), while using far fewer parameters ($|\theta|$). As expected, neural methods perform better than MIR alone with bag-of-word features. However, Dense+MIR outperforms Dense as well as LSTM, while using only 200 more parameters. Moreover, GRU+MIR outperforms all other methods, including GRU alone and RCNN. These results show that MIR is beneficial regardless of the feature space used and

	BeerAdvocate		**RateBeer (ES)**		**RateBeer (FR)**		**Audiobooks**		**Toys&Games**	
Model \ Error	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
AverageRating	14.20	3.32	16.59	4.31	12.67	2.69	21.07	6.75	20.96	6.75
Aggregated (ℓ_1)	13.62	3.13	15.94	4.02	12.21	2.58	20.10	6.14	20.15	6.33
Aggregated (ℓ_2)	14.58	3.68	14.47	3.41	12.32	2.70	19.08	5.99	18.99	5.93
Instance (ℓ_1)	12.67	2.89	14.91	3.54	11.89	2.48	20.13	6.17	20.33	6.34
Instance (ℓ_2)	13.74	3.28	14.40	3.39	11.82	2.40	19.26	6.04	19.70	6.59
Prime (ℓ_1)	12.90	2.97	15.78	3.97	12.70	2.76	20.65	6.46	21.09	6.79
Prime (ℓ_2)	14.60	3.64	15.05	3.68	12.92	2.98	20.12	6.59	20.11	6.92
Clustering (ℓ_2)	13.95	3.26	15.06	3.64	12.23	2.60	20.50	6.48	20.59	6.52
Weighted MIR (ℓ_2)	**12.24**	**2.66**	**14.18**	**3.28**	**11.37**	**2.27**	**18.89**	**5.71**	**18.50**	**5.57**
..vs SVR (%)	*+16.0*	*+27.7*	*+2.0*	*+3.8*	*+7.6*	*+15.6*	+1.0	*+4.5*	*+2.6*	*+6.0*
..vs Lasso (%)	*+10.1*	*+15.1*	*+11.0*	*+18.4*	*+6.8*	*+11.8*	*+6.0*	*+6.9*	*+8.1*	*+11.9*
..vs 2^{nd} best (%)	+3.3	+7.8	+1.5	*+3.3*	*+3.7*	*+4.9*	+1.0	*+4.5*	*+2.6*	*+6.0*

Table 3: Performance of aspect rating prediction in terms of Mean Absolute Error (MAE) and Mean Squared Error (MSE) with 5-fold c.-v., averaged over all aspects in each dataset. The best scores are in **bold** and the second best ones are underlined. Significant improvements (paired t-test, $p < 0.05$) are in *italics*.

the input or the intermediate hidden states it operates on. Moreover, MIR has remarkable computational efficiency in estimating explicit document representations, overcoming the bottleneck of previous methods which attempt to learn a single, rigid and fixed-length document representation. Lastly, our instance relevance mechanism is able to improve the performance of the network while retaining its explanatory potential (as we show below), in contrast to the model proposed by Lei et al., which is bounded by the performance of its encoder network. Especially, when their model reaches the performance of the encoder network, it has zero explanatory potential (Lei et al., 2016, Fig. 2).

7.3 Effects of the Model Design Choices

In this section, we experiment with 5-fold c.-v. on equal-size samples of 1,200 instances per dataset. We use APWeights for training, because it has fewer hyper-parameters to tune and is faster than SGDWeights for smaller training sets, as shown in Section 7.3.3.

7.3.1 Multiple-Instance Regression Assumptions

The proposed model relies on an original weighted instance relevance assumption to achieve state-of-the-art results, as shown in the previous section, but how does it compare with baseline MIR assumptions? We compare it here with Aggregated Instances, Instance as Example, Prime Instance, and Clustering (the instance relevance method proposed by Wagstaff et al., 2008). Moreover, we report for comparison the scores of a method that always predicts the average rating over the training set, noted as Average Rating.

Table 3 shows the performance of the models with various MIR assumptions for aspect rating prediction (columns 1 to 5). We have compared them on sentiment and emotion rating prediction in earlier work (Pappas & Popescu-Belis, 2014). The proposed model consistently outperforms the other assumptions on all datasets. In particular, it outperforms

	BOW		TF-IDF		word2vec	
Model \ Error	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
Aggregated (ℓ_1)	17.08	<u>4.17</u>	16.59	<u>3.97</u>	16.03	3.84
Aggregated (ℓ_2)	<u>16.88</u>	4.47	<u>16.25</u>	4.16	<u>14.62</u>	<u>3.30</u>
Instance (ℓ_1)	17.69	4.37	18.11	4.50	16.37	3.86
Instance (ℓ_2)	16.93	4.24	16.88	4.23	15.60	3.67
Prime (ℓ_1)	17.39	4.37	17.72	4.43	16.13	3.89
Prime (ℓ_2)	18.03	4.91	17.10	4.29	15.71	3.72
Weighted MIR (ℓ_2)	**15.91**	**3.95**	**15.36**	**3.63**	**14.25**	**3.29**
MIR vs. SVR	*+5.7%*	*+11.2%*	*+5.5%*	*+12.5%*	+2.56%	+ 0.29%
MIR vs. Lasso	*+6.8%*	+5.0%	*+7.3%*	*+8.5%*	*+12.47%*	*+16.82%*
MIR vs. 2^{nd} best	*+5.7%*	+5.0%	*+5.5%*	*+8.5%*	+2.56%	+ 0.29%

Table 4: MAE and MSE ($\times 100$) on sentiment prediction with 5-fold c.-v. over TED comments, with three features spaces: BOW with counts, TF-IDF weights, and word embeddings from word2vec. Improvements in *italics* are significant at $p < 0.05$ (pairwise t-test).

the two models with the Aggregated Instance assumption, which correspond respectively to traditional BOW with SVR (when using the ℓ_2 norm) and Lasso (when using the ℓ_1 norm). The Aggregated (ℓ_2) baseline has on average 11% lower performance than our model in terms of MSE and about 6% in terms of MAE. Similarly, the Aggregated (ℓ_1) baseline has on average 13% lower MSE and 8% MAE than the proposed model. As we have previously shown (Pappas & Popescu-Belis, 2014), the same conclusions can be drawn for each aspect of the five review datasets.

The Instance as Example assumption performs quite well on BeerAdvocate and Toys & Games (for MSE) with ℓ_1, on Ratebeer (ES), RateBeer (FR) and Toys & Games (for MAE) with ℓ_2. Therefore, this assumption is appropriate for this task, but it still scores below our model, by about 5% MAE and 8%–9% MSE. The Prime Instance assumption with ℓ_1 performs well only on BeerAdvocate, and with ℓ_2 only on Toys & Games, again with lower scores than our model, namely by about 9% MAE and 15%–18% MSE. This suggests that the Prime assumption is not the most appropriate for this task. Lastly, even though Clustering attempts to model instance relevance, as we do, it only reaches scores similar to Prime, presumably because the relevance weights are assigned based on the computed clusters and are not directly influenced by the objective of the task.

Our model also outperforms all other methods for sentiment prediction over comments of TED talks (see Table 4), as well as for talk-level emotion prediction with 12 dimensions. For sentiment prediction, it outperforms SVR by 11% MSE and Lasso by 5%. For emotion prediction (averaged over all 12 aspects), differences are smaller, at 1.6% and 2.9% respectively (Pappas & Popescu-Belis, 2014). The smaller differences could be explained by the fact that among the 10 comments selected per talk, several were not directly related to the emotion that the system tries to predict.

Dataset		APWeights				SGDWeights			
Name	**Size**	**MSE**	t/i	i	T **(s)**	**MSE**	t/i	i	T **(s)**
BeerAdvocate	1,586,259	1.30	7,931.3	3	23,793.9	1.26	326.1	5	1,728.3
Toys & Games	373,974	4.90	1,136.7	2	2,273.4	5.06	90.1	8	743.2
Pubs	53,492	1.92	129.8	20	2,596.0	1.65	11.0	33	370.0
Audiobooks	10,989	5.58	13.7	3	41.0	5.26	1.9	159	311.7
RateBeer (FR)	17,998	1.83	38.1	2	76.2	1.82	2.9	14	41.8
RateBeer (ES)	1,259	3.40	1.2	3	3.5	3.40	0.3	196	49.0
TED comments	1,200	4.13	0.8	5	4.2	4.05	0.2	202	38.4
TED talks	1,203	4.86	1.1	17	19.6	4.83	0.2	164	32.9
Average	-	3.49	1,321.6	-	3,600.1	**3.41**	**54.1**	-	**424.4**

Table 5: Comparison of the APWeights and SGDWeights learning algorithms for aspect rating prediction, in terms of time per iteration (t/i, in seconds), number of iterations (i) and total time (T) per aspect. SGDWeights is slower for smaller datasets, but is clearly more efficient on the larger ones, with better MSE.

7.3.2 Independence from the Feature Space

The MIR model does not make any assumption about the feature space. We examine here whether the improvements it brings remain present even with a different feature space, for instance one based on TF-IDF coefficients or word2vec features instead of BOW with counts. For sentiment prediction on TED comments, we find that by changing the feature space to TF-IDF, strong baselines such as Aggregated (ℓ_1) and (ℓ_2), i.e. SVR and Lasso, improve their performance, reaching 16.2% and 16.6% MAE respectively (with 4.2% and 4.0% MSE). However, our model still outperforms them on both MAE and MSE scores, which reach 15.3% and 3.6%, thus improving over SVR by 5.5% on MAE and 12.5% on MSE, and over Lasso by 7.4% on MAE and 8.5% on MSE.

When using word embeddings as features, all methods improve their performance, however, MIR still outperforms all other methods. These results suggest that MIR improvements hold also on more sophisticated feature spaces. Moreover, MIR also improves the performance of neural networks when it is used as a pooling layer for attending their intermediate hidden states, as we showed in Section 7.2. Of course, apart from better performance on rating prediction, MIR also provides meaningful structural information about the input, as we demonstrate in Section 9.

7.3.3 Efficiency of the Learning Algorithms

We compare here the performance and the execution time of two learning algorithms, the initial APWeights and SGDWeights (see Section 4.2 above). As in Section 7.2, we use the full-sized datasets with uniform training/testing splits. Table 5 displays the MSE score per dataset, the average duration per iteration and per rating dimension, in seconds. The datasets are listed by decreasing number of reviews.

In terms of efficiency, SGDWeights scales better than APWeights when the number of training examples increases: the average time per iteration and per rating dimension of SGDWeights increases more smoothly than for APWeights, which has a steeper increase.

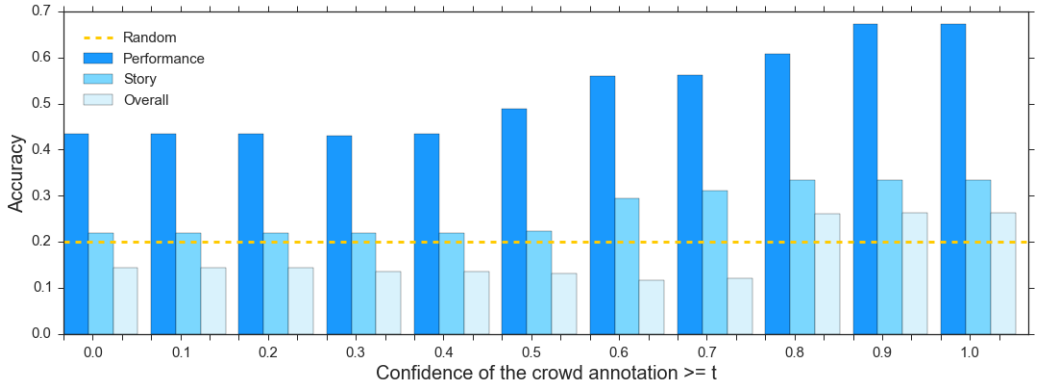


Figure 4: Accuracy of the proposed MIR model on predicting the explanatory value of sentences with respect to review-level ratings of the three aspects, for subsets of increasing crowd confidence values. Random accuracy is 1 out of 5, i.e. 20%.

However, APWeights is faster on smaller datasets such as TED talks, TED comments and Ratebeer (ES). At the same time, SGDWeights equals or outperforms APWeights on most datasets, as shown by its average MSE scores.

8. Intrinsic Evaluation of Saliency Values: Comparison to Humans

Apart from the competitive results for rating prediction, one of our central claims is that the saliency values of our model reflect the contribution of each sentence to its rating predictions, in other words, that the model provides *interpretable instance weights*. In this section, we evaluate this important property intrinsically, while in Section 9 we demonstrate the benefits of the weights in application to other tasks. Qualitative analyses on several examples can be found elsewhere (Pappas & Popescu-Belis, 2014, Sec. 7.2).

We compare the MIR weights with the explanatory values assigned by humans to review sentences, on the HATDOC dataset with 1,662 sentences (presented in Section 6.2). For training, we collected a uniform sample of audiobook reviews from Audible, with 10,000 reviews for each overall rating (1 to 5), hence 50,000 reviews not included in the test set. This ensures that our system can learn to model reviews from all the rating levels, while the Audiobook set presented in Section 6.1 does not have a uniform rating distribution. All reviews are accompanied by review-level ratings of three aspects of the respective audiobook: 'overall quality', 'performance', and 'story'. To compare our model's output with this ground truth, we convert our $[0, 1]$ saliency values into categorical labels from 1 to 5 using $round(4\hat{\psi} + 1)$. Based on tests over a development subset, our model is trained with SGDWeights and ADAGRAD (see Section 4.2 above), with a step size of 0.001.

We first examine the accuracy of the saliency values predicted by our model, i.e. the proportion of sentences with an explanatory label that is identical to the one assigned by humans. Predicting the exact label is a challenging task, given that even the human judges do not have full agreement. Figure 4 displays the accuracy of our model, for each aspect, for test subsets of increasing crowd confidence, from the entire test set to only the most reliable

labels. The model achieves the highest accuracy on the 'performance' aspect, with up to 60% accuracy for labels assigned with at least 0.8 confidence by humans. The accuracy for the 'story' aspect reaches up to 33%, while 'overall' has the lowest accuracy, at 26%. Our model thus significantly outperforms the random baseline, which has 20% accuracy (1 out of 5). The low performance on the 'overall' aspect can be attributed to its generality, as it includes elements of the other two aspects along with other evaluative and non-evaluative statements.

When relaxing the constraints of exact label matching, i.e. accepting neighboring labels as matches (distance of 1), the accuracies at the same confidence levels as above increase to 71%, 43% and 52% respectively. Interestingly, the 'overall' aspect benefits the most from this relaxation, showing that many predictions were actually close but not identical to the human label. The performance of MIR on all aspects is greater at higher crowd confidence values (≥ 0.8), which shows that both the system and the humans find similar difficulties in labeling explanatory power. To provide a more nuanced assessment of MIR weights, they can also be compared with those from human judges by placing all of them on the same scale of qualification, and estimating reliability as agreement with the average (see Pappas & Popescu-Belis, 2016).

9. Evaluation of Saliency Values for Segmentation and Summarization

We report the results on segmentation and summarization using the aspect saliency and sentiment of sentences within the sequence labeling model proposed in Section 5. Then, we investigate the discriminative potential and importance of the proposed features when varying the parameters of the sequence labeling model on these tasks.

9.1 Experimental Settings

To compare with state-of-the-art on aspect segmentation and summarization, we use a uniform split of the data (50% for training and 50% for testing) as did McAuley et al. (2012). To account for randomness effects in the splitting, we report the average scores of each method over five runs (Section 9.2).

We compare our model with the methods used by McAuley et al. (2012). As Lei et al. (2016) used a modified version of McAuley's segmentation task to evaluate their word-based selection method, this is not directly comparable with McAuley's or our method. We consider *Latent Dirichlet Allocation* (LDA; Blei, Ng, & Jordan, 2003), which was trained by McAuley et al. with various numbers of topics: 10, 50, or the number of aspects k. The abstract topics obtained from LDA on the training set were then manually aligned by McAuley at al. with the aspects, which is why they considered this method to be semi-supervised. We report in Figure 5 only the best LDA scores out of their three configurations. Further tuning of the number of topics would be impractical due to the manual alignment; moreover, our main focus is on the unsupervised or fully-supervised methods and baselines. We also consider McAuley et al.'s *PALE LAGER*, which achieves state-of-the-art performance on the tasks we consider and supports three types of learning, namely unsupervised, semi-supervised and fully-supervised.

For both tasks, we report the accuracy score, i.e. the fraction of correct predictions. For summarization, we also report the Area Under Curve (AUC) score, as we can view the task

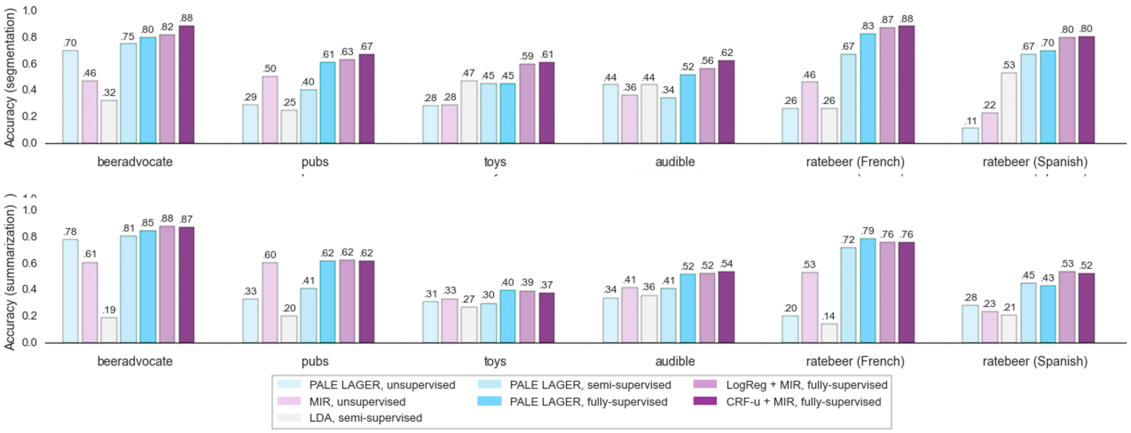


Figure 5: Accuracy on review segmentation (top) and on summarization (bottom) of the CRF models with BOW+MIR features, compared to several baselines. The scores of LDA and PALE LAGER are from McAuley et al. (2012).

as the retrieval of the most relevant sentences for each aspect, and use the probabilities for each sentence and aspect to rank sentences (correct sentences should rank higher) – the AUC can be computed using the ground-truth aspect labels.

To demonstrate the merits of our features across several sequence labeling models, we evaluate them in Section 9.3 over five random splits, 80% for training and 20% for testing. In each experiment, we select the model and features by cross-validation on the training data. All the CRF models are optimized over the same range of values for their regularization term, noted $a \cdot 10^b$, with $a \in \{1, \ldots, 9\}$ and $b \in \{-3, \ldots, +3\}$. We train our MIR model with SGDWeights (based on the findings in Section 7) on each dataset, without making available the segmentation labels, and we test all the unique combinations of MIR features, presented in Section 5, in addition to BOW features.

9.2 Comparison with the State of the Art

We evaluate three of our models from Section 5: (1) the unsupervised, direct aspect assignment according to the saliency values learned by our model; (2) the fully-supervised unstructured model that uses our aspect saliency and sentiment features, noted LogReg+MIR; and (3) the structured models with MIR features, noted CRF-u + MIR and CRF-d + MIR.

Review segmentation. For segmentation, the proposed linear chain CRF with MIR features outperforms all other baselines over all datasets (see Figure 5, upper part), with about 15% relative improvement on average over the best baseline (PALE LAGER, fully-supervised). The largest differences appear on the Toys and the Audiobooks datasets with about 31% and 14% improvement respectively. Compared to the semi-supervised models, LDA and PALE LAGER, the improvements are much higher, as expected. The CRF-u model performs better than the simple CRF, demonstrating that capturing the relationship between aspects is beneficial for this task.

	BeerAdvocate		**Pubs**		**Toys & Games**		**Audiobooks**		**RateB. (FR)**		**RateB. (ES)**	
Model	BOW	+MIR	BOW	+MIR	BOW	+MIR	BOW	+MIR	BOW	+MIR	BOW	+MIR
Segmentation: accuracy scores												
Unsup.	-	**47.05**	-	**52.03**	-	**28.64**	-	**36.46**	-	**45.76**	-	**23.00**
LogReg	81.36	**82.96***	64.57	**65.97**†	58.62	**60.40**	55.83	**60.68**†	87.57	**90.30***	81.02	**82.73**
CRF-u	87.55	**88.80***	67.35	**68.35**†	59.94	**62.14**	61.85	**67.92***	88.28	**90.35***	**83.07**	81.75
CRF-d	88.96	**89.21**	**69.73**	69.66	58.30	**61.21**	64.56	**65.94**	88.88	**90.37**†	**83.42**	82.88
Summarization: accuracy scores												
Unsup.	-	**60.66**	-	**61.13**	-	**33.37**	-	**41.03**	-	**48.52**	-	**24.33**
LogReg	87.78	**88.87***	62.54	**63.16**	35.74	**42.47***	**56.20**†	53.38	74.86	**76.19**	53.33	**57.33**
CRF-u	87.89	**88.24***	61.18	**63.14**†	40.23	**41.10***	56.39	**58.01**	73.62	**74.95**	55.33	**56.00**
CRF-d	87.53	**87.55**	61.14	**62.93***	38.41	**41.41**	54.91	**55.65**	**74.29**	74.19	54.33	**56.33**
Summarization: AUC scores												
Unsup.	-	**30.19**	-	**20.50**	-	**26.92**	-	**34.73**	-	**25.29**	-	**20.61**
LogReg	87.72	**88.79***	63.01	**64.59***	45.43	**47.76***	**60.77**	60.60	**72.25**	71.42	**50.63**	48.97
CRF-u	**87.61***	87.29	63.84	**64.70***	46.81	**47.53**†	59.79	**63.64***	71.11	**72.20***	**50.44**	49.82
CRF-d	86.76	**86.83**	63.57	**65.35***	46.14	**47.48***	61.27	**62.07**	71.04	**71.81**	50.11	**51.14**

Table 6: Performance of CRF models for aspect segmentation and summarization with 5-fold c.-v. The best scores for each comparison between BOW and BOW+MIR are in bold, and the best scores for each dataset and task are underlined. Significance is noted with a '*' for the 90% level and a '†' for the 80% level.

In the unsupervised setting, PALE LAGER achieves higher scores than our model on BeerAdvocate and Audiobooks by 52% and 22%, however, our model outperforms PALE LAGER by a larger margin, namely by 72%, 77% and 100% respectively on Pubs, Ratebeer (FR) and Ratebeer (ES). On Toys, the performances of the two systems are similar. We should note that PALE LAGER has a modeling advantage in this task, because it is able to de-correlate aspect words from sentiment words. In contrast, our model captures the aspect sentiment saliency and may assign low scores on non-factual sentences regardless of whether they actually discuss a particular aspect.

Review summarization. The performance on review summarization of the CRF with MIR features is slightly higher on average (+2.3%) than the fully-supervised PALE LAGER, with the best scores found on Ratebeer (ES) and BeerAdvocate (+18.6% and +2.3%) and the worst ones on Ratebeer (FR) and Toys (−3.8% and −5%), as shown in Fig. 5, lower part. One explanation for not always improving summarization (e.g. on Toys), despite the improved segmentation model over PALE LAGER, is that here only the reviews with more sentences than the number of aspects are considered, by definition of the task. Such reviews are in fact more difficult to segment than reviews with fewer sentences.

As for the unsupervised setting, PALE LAGER outperforms our model on BeerAdvocate and Ratebeer (ES) by 28% and 21% respectively, while our model outperforms PALE LAGER on all other datasets, namely Pubs, Toys, Audiobooks, and Ratebeer (FR), respectively by 81%, 6%, 20%, and 165%. Even though both models are superior to the semi-supervised variants of LDA, our model outperforms LDA by a larger margin than PALE LAGER does. Interestingly, the performance of our model on Pubs, without task supervision, is similar to the best fully supervised model.

9.3 Discriminative Potential of MIR Features

The results of several sequence labeling models with vs. without MIR features (Table 6) show the discriminative potential of the MIR features on segmentation and summarization. The CRF models with MIR features for aspect saliency and sentiment outperform those not using them. Out of the 18 combinations of the three tasks and the six datasets, the CRF model with BOW+MIR features outperforms BOW alone for 15 of them, i.e. 83%. Furthermore, the CRF model with BOW+MIR features outperforms BOW alone in 87% of the pairwise model comparisons (47 out of 54).

Review segmentation. The BOW+MIR features segment the reviews more accurately than BOW only in 83% of the pairwise comparisons for this task, shown in the first three lines of Table 6. The structured models (CRF-u and CRF-d) outperform unstructured ones (LogReg) over all datasets, highlighting the importance of modeling the relationships between aspects. The unstructured model (LogReg) benefited the most from the MIR features, on average, followed by the structured models CRF-u and CRF-d. The MIR features were not informative on Pubs and Ratebeer (ES).

Review summarization. On the summarization task, the BOW+MIR features surpass the BOW features in 89% of the pairwise comparisons. The LogReg unstructured model outperforms CRF on all the datasets except for Audiobooks (lines 4–6 of Table 6). When the linear chain CRFs are given BOW+MIR features, they improve over LogReg for Audiobooks. Moreover, the structured models do not outperform the unstructured ones on this task; this can be attributed to the fact that the summarization labels are not sequential and thus the structured information is not as helpful as on segmentation.

On the same task evaluated with AUC metric (lines 7–9 of Table 6), the BOW+MIR features help to summarize the reviews more accurately than BOW in 78% of the pairwise comparisons. Here, the linear chain CRFs outperform the unstructured baseline (LogReg) on 3 out of 6 datasets, namely Pubs, Audioboks and Ratebeer (ES), although they benefit from MIR features on all datasets. The lowest performance is observed on Ratebeer (ES), as for the segmentation task; one reason for this might be that the small context prevents CRF+MIR to accurately learn the relationship between aspects.

Feature analysis. To identify the most informative MIR features that are the most informative for each task, from those in listed in Section 5.2, Table 2, we analyze their importance over the training data (with 10-fold c.-v.), both for summarization and segmentation. We estimate the importance of each type of features based on the likelihood of appearance in the top 5 models (out of 770 per fold, i.e. the number of feature combinations) which outperformed CRF with BOW. The results are represented as heatmap diagrams in Figure 6 for each dataset (columns).

For segmentation, the most informative features are the aspect saliency of the current, next and previous sentences (*sw*, *sw_next*, *sw_prev*) for BeerAdvocate, Ratebeer (ES) and Ratebeer (FR); and the aspect sentiment of the current sentence (*sp*) for Pubs, Toys and Audiobooks. Interestingly, the former features are more informative than the latter ones; this suggests that the context of the sentence plays an important role in finding the aspect it discusses. For summarization, the aspect saliencies of the previous and next sentences (*sw_prev*) are the most important features for BeerAdvocate and Ratebeer (ES), the aspect

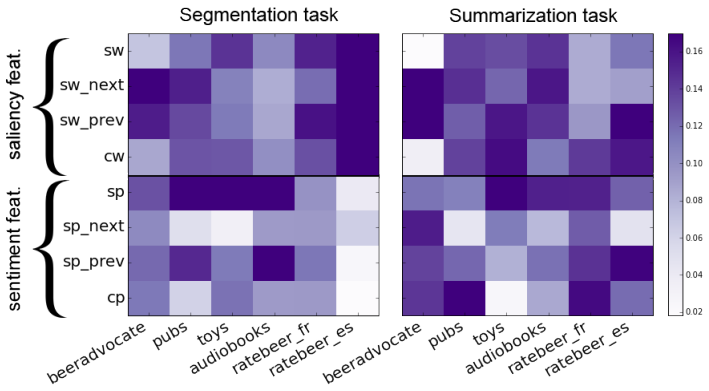


Figure 6: Importance of MIR features for segmentation and summarization as their likelihood of appearance in the top-5 models better than CRF with BOW (10-fold c.-v. on the training data).

sentiment of the current sentence (*sp*) for Toys, the global sentiment (*cp*) for Pubs and Ratebeer (FR), and the aspect saliency of next sentence (*sp_next*) for Audiobooks.

The best performing features across all datasets and models are: for segmentation, the aspect saliency of the next sentence and the sentiment of the current sentence (*sw_next* and *sp* at 0.87); and for summarization, the aspect saliency of the previous sentence (*sw_prev* at 0.90) and the sentiment of the current sentence (*sp* at 0.86). On average, the saliency features have a higher importance compared to the aspect sentiment ones for both tasks, namely 0.84 vs. 0.66 and 0.77 vs. 0.72. Overall, MIR features have good discriminative potential on both tasks, although the optimal features depend on the dataset.

The MIR weights appear to capture at least partly some of the sequential information that is captured by the CRFs. Such information is clearly beneficial for both segmentation and summarization, as seen from the differences between the scores of LogReg and CRF, without MIR, in Table 6. (And, as expected, it is more beneficial for the former than the second task, where the differences between scores are smaller.) However, when using the MIR features, the scores of LogReg are much closer to those of the CRF models, and even outperform CRF+MIR in 4 out of 6 experiments with summarization. The information about sequences is likely learned by the MIR when modeling the weights of the features from the next and previous sentence to the current one.

10. Conclusion

We proposed a weighted multiple-instance regression model for explicit document modeling, with a direct application to multi-aspect sentiment analysis. The proposed model learns instance relevance weights together with target labels. When used for rating prediction, it outperforms several MIR and non-MIR baselines on seven publicly available datasets, even when the sophistication of the feature space increases. This suggests that our contribution is to a certain extent independent from feature engineering or learning. The results vali-

date our hypothesis that the target aspect ratings are weakly connected to the individual segments of a text, and that accounting for this uncertainty is an appropriate strategy.

The document models learned with MIR are explicit in the sense that they have explanatory power, as we demonstrated in two ways. Firstly, we showed that the learned weights are comparable with sentence-level human judgments obtained by crowdsourcing. Secondly, we showed that the weights can augment word or topic feature spaces with information regarding the sentiment of sentences, which in combination with CRF models reached superior or similar performance to the state of the art. The MIR model thus captures meaningful structural information which is helpful for text understanding tasks. Such information has applications beyond multi-aspect sentiment analysis, e.g. to summarize opinionated text or to exploit user reviews within recommender systems.

Acknowledgments

We are grateful for their support to the European Union through its 7th Framework Program (inEvent project n. 287872, www.inevent-project.eu) and its Horizon 2020 program (SUMMA project n. 688139, www.summa-project.eu), and to the Swiss National Science Foundation (MODERN project n. 147653, www.idiap.ch/project/modern/). We would also like to thank the anonymous reviewers for their helpful suggestions.

References

Amores, J. (2013). Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, *201*, 81–105.

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2003). Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, pp. 561–568, Vancouver, BC, Canada.

Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Multi-facet rating of product reviews. In Boughanem, M., Berrut, C., Mothe, J., & Soule-Dupuy, C. (Eds.), *Advances in Information Retrieval*, Vol. 5478 of *Lecture Notes in Computer Science*, pp. 461–472. Springer Berlin Heidelberg.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *5th International Conference on Learning Representations*, San Diego, USA.

Bao, Y., Fang, H., & Zhang, J. (2014). TopicMF: Simultaneously exploiting ratings and reviews for recommendation. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 2–8, Québec City, Québec, Canada.

Bauschke, H. H., & Borwein, J. M. (1996). On projection algorithms for solving convex feasibility problems. *SIAM Review*, *38*(3), 367–426.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(30), 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 ACL (HLT-NAACL)*, HLT '10, pp. 804–812, Los Angeles, CA, USA.

Bunescu, R. C., & Mooney, R. J. (2007). Multiple instance learning for sparse positive bags. In *Proceedings of the 24th Annual International Conference on Machine Learning*, ICML '07, Corvallis, OR, USA.

Cheplygina, V., Tax, D. M., & Loog, M. (2015). Multiple instance learning with bag dissimilarities. *Pattern Recognition*, *48*(1), 264–275.

Choi, E., Rashkin, H., Zettlemoyer, L., & Choi, Y. (2016). Document-level sentiment inference with social, faction, and discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 333–343, Berlin, Germany.

Chopra, S., Auli, M., & Rush, A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 93–98, San Diego, California. Association for Computational Linguistics.

Chung, J., Gülcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, *abs/1412.3555*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Das, A., Agrawal, H., Zitnick, C. L., Parikh, D., & Batra, D. (2016). Human attention in visual question answering: Do humans and deep networks look at the same regions?. *CoRR*, *abs/1606.03556*.

Deng, L., & Wiebe, J. (2015). Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 179–189, Lisbon, Portugal.

Dietterich, T. G., Lathrop, R. H., & Lozano-Prez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*(12), 31 – 71.

Doran, G., & Ray, S. (2014). A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. *Machine Learning*, *97*(1-2), 79–102.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. In *Advances in Neural Information Processing systems*, pp. 155–161, Denver, CO, USA.

Duchi, J. C., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*, 2121–2159.

Foulds, J., & Frank, E. (2010). A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, *25:1*, 1–25.

Ganu, G., Elhadad, N., & Marian, A. (2009). Beyond the stars: Improving rating predictions using review text content. In *Proceedings of the 12th International Workshop on the Web and Databases*, WebDB '09, Providence, RI, USA.

Gupta, N., Di Fabbrizio, G., & Haffner, P. (2010). Capturing the stars: Predicting ratings for service and product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, SS '10, pp. 36–43, Los Angeles, CA, USA.

Hamdan, H., Bellot, P., & Bechet, F. (2015). Lsislif: CRF and logistic regression for opinion target extraction and sentiment polarity analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pp. 753–758, Denver, Colorado.

Hatzivassiloglou, V., & Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pp. 299–305, Saarbrücken, Germany.

Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *CoRR*, *abs/1506.03340*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, *9*(8), 1735–1780.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., & Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 541–550, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge discovery and data mining*, KDD '04, pp. 168–177, Seattle, WA.

Kim, B. (2015). *Interactive and interpretable machine learning models for human machine collaboration*. Phd thesis, Massachusetts Institute of Technology.

Kim, S., Zhang, J., Chen, Z., Oh, A., & Liu, S. (2013). A hierarchical aspect-sentiment model for online reviews. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, pp. 526–533. AAAI Press.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*.

Kotzias, D., Denil, M., de Freitas, N., & Smyth, P. (2015). From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 597–606, Sydney, NSW, Australia.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1-2), 83–97.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of*

the 18th International Conference on Machine Learning, ICML '01, pp. 282–289, San Francisco, CA, USA.

Lakkaraju, H., Socher, R., & Manning, C. (2014). Aspect specific sentiment analysis using hierarchical deep learning. In *NIPS Workshop on Deep Learning and Representation Learning*, NIPS '14, Montréal, Canada.

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, Austin, Texas. Association for Computational Linguistics.

Lewis, A., & Malick, J. (2008). Alternating projections on manifolds. *Mathematics of Operations Research*, *33*(1), 216–234.

Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S., & Yu, H. (2010). Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pp. 653–661, Beijing, China.

Li, F., Liu, N., Jin, H., Zhao, K., Yang, Q., & Zhu, X. (2011). Incorporating reviewer and product information for review rating prediction. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence - Volume 3*, IJCAI '11, pp. 1820–1825, Barcelona, Spain.

Li, J., Chen, X., Hovy, E. H., & Jurafsky, D. (2015). Visualizing and understanding neural models in NLP. *CoRR*, *abs/1506.01066*.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pp. 375–384, Hong Kong, China.

Lipton, Z. C. (2016). The mythos of model interpretability. In *2016 ICML Workshop on Human Interpretability in Machine Learning*, New York, USA.

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pp. 150–158, Beijing, China.

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 623–631, New York, NY, USA. ACM.

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. In *Proceedings of the 11th IEEE International Conference on Data Mining Workshops*, ICDMW '11, pp. 81–88, Washington, DC, USA.

Luong, M., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pp. 1412–1421, Lisbon, Portugal.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 142–150, Portland, OR, USA.

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pp. 165–172, Hong Kong, China.

McAuley, J., Leskovec, J., & Jurafsky, D. (2012). Learning attitudes and attributes from multi-aspect reviews. In *Proceedings of the 12th IEEE International Conference on Data Mining*, ICDM '12, pp. 1020–1025, Brussels, Belgium.

Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on the World Wide Web*, WWW '07, pp. 171–180, Banff, AB, Canada.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., & Weinberger, K. (Eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119.

Mitchell, M., Aguilar, J., Wilson, T., & Van Durme, B. (2013). Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1643–1654, Seattle, Washington, USA.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, *5*(1), 32–38.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines.. In *International Conference in Machine Learning*, pp. 807–814, Haifa, Israel.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pp. 115–124, Ann Arbor, MI, USA.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '02, pp. 79–86, Philadelphia, PA, USA.

Pappas, N., & Popescu-Belis, A. (2013). Sentiment analysis of user comments for one-class collaborative filtering over TED talks. In *Proceedings of the 36th international ACM SIGIR Conference on Research and development in information retrieval*, SIGIR '13, pp. 773–776, Dublin, Ireland.

Pappas, N., & Popescu-Belis, A. (2014). Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pp. 455–466, Doha, Qatar.

Pappas, N., & Popescu-Belis, A. (2016). Human versus machine attention in document classification: A dataset with crowdsourced annotations. In *Proceedings of the EMNLP 4th Workshop on Natural Language Processing for Social Media*, SocialNLP 2016, pp. 94–100, Austin, TX, USA.

Patra, B. G., Mandal, S., Das, D., & Bandyopadhyay, S. (2014). JU_CSE: A conditional random field (CRF) based approach to aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pp. 370–374, Dublin, Ireland.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2012). Scikit-learn: Machine learning in python. *CoRR*, *abs/1201.0490*.

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 8th International Workshop on Semantic Evaluation*, SemEval '14, pp. 27–35, Dublin, Ireland.

Qu, L., Ifrim, G., & Weikum, G. (2010). The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pp. 913–921, Beijing, China.

Ray, S., & Page, D. (2001). Multiple instance regression. In *Proceedings of the 18th International Conference on Machine Learning*, ICML '01, pp. 425–432.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, *abs/1602.04938*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., & PDP Research Group, C. (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, pp. 318–362. MIT Press, Cambridge, MA, USA.

Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. *CoRR*, *abs/1509.00685*.

Sauper, C., & Barzilay, R. (2013). Automatic aggregation by joint modeling of aspects and values. *Journal of Artificial Intelligence Research*, *46*(1), 89–127.

Sauper, C., Haghighi, A., & Barzilay, R. (2010). Incorporating content structure into text analysis applications. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pp. 377–387, Cambridge, MA.

Sennrich, R., Haddow, B., & Birch, A. (2016). Edinburgh neural machine translation systems for WMT 16. *CoRR*, *abs/1606.02891*.

Settles, B., Craven, M., & Ray, S. (2008). Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, NIPS '08, pp. 1289–1296, Vancouver, BC, Canada.

Snyder, B., & Barzilay, R. (2007). Multiple aspect ranking using the good grief algorithm. In *In Proceedings of the Annual Conference of the North American Chapter of the ACL*, HLT-NAACL '07, pp. 300–307, Rochester, NY, USA.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 151–161, Edinburgh, UK.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pp. 1631–1642, Portland, OR, USA.

Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-to-end memory networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 28*, pp. 2440–2448. Curran Associates, Inc.

Surdeanu, M., Tibshirani, J., Nallapati, R., & Manning, C. D. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pp. 455–465, Jeju Island, Korea.

Tang, D. (2015). Sentiment-specific representation learning for document-level sentiment analysis. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, WSDM '15, pp. 447–452, Shanghai, China.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the ACL*, ACL '14, pp. 1555–1565, Baltimore, MD, USA.

Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pp. 327–335, Sydney, NSW, Australia.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society (Series B)*, *58*, 267–288.

Titov, I., & McDonald, R. (2008a). A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the Annual Meeting of the ACL*, HLT '08, pp. 308–316, Columbus, OH, USA.

Titov, I., & McDonald, R. (2008b). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pp. 111–120, Beijing, China.

Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, *6*, 1453–1484.

Wagstaff, K. L., & Lane, T. (2007). Salience assignment for multiple-instance regression. In *Proceedings of the ICML 2007 Workshop on Constrained Optimization and Structured Output Spaces*, Corvallis, OR, USA.

Wagstaff, K. L., Lane, T., & Roper, A. (2008). Multiple-instance regression with structured data. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, ICDMW '08, pp. 291–300.

Wang, H., Lu, Y., & Zhai, C. (2010). Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD Interna-*

tional Conference on Knowledge Discovery and Data Mining, KDD '10, pp. 783–792, Washington, DC, USA.

Wang, H., Nie, F., & Huang, H. (2011). Learning instance specific distance for multi-instance classification. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 507–512, San Francisco, CA, USA.

Wang, Z., Lan, L., & Vucetic, S. (2012). Mixture model for multiple instance regression and applications in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(6), 2226–2237.

Wang, Z., Radosavljevic, V., Han, B., Obradovic, Z., & Vucetic, S. (2008). Aerosol optical depth prediction from satellite observations by multiple instance regression. In *Proceedings of the SIAM International Conference on Data Mining*, SDM '08, pp. 165–176, Atlanta, GA, USA.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pp. 347–354, Vancouver, BC, Canada.

Wu, C., Beutel, A., Ahmed, A., & Smola, A. J. (2015). Explaining reviews and ratings with PACO: Poisson additive co-clustering. *CoRR*, *abs/1512.01845*.

Xiong, C., Merity, S., & Socher, R. (2016). Dynamic memory networks for visual and textual question answering. *CoRR*, *abs/1603.01417*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, *abs/1502.03044*.

Xu, W., Hoffmann, R., Zhao, L., & Grishman, R. (2013). Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 665–670, Sofia, Bulgaria. Association for Computational Linguistics.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'16, San Diego, California.

Zhang, M.-L., & Zhou, Z.-H. (2008). M3MIML: A maximum margin method for multi-instance multi-label learning. In *Proceedings of the 8th IEEE International Conference on Data Mining*, ICDM '08, pp. 688–697.

Zhang, M.-L., & Zhou, Z.-H. (2009). Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence*, *31*(1), 47–68.

Zhao, H., Lu, Z., & Poupart, P. (2015). Self-adaptive hierarchical sentence model. *CoRR*, *abs/1504.05070*.

Zhou, Z.-H., Jiang, K., & Li, M. (2005). Multi-instance learning based web mining. *Applied Intelligence*, *22*(2), 135–147.

Zhou, Z.-H., Sun, Y.-Y., & Li, Y.-F. (2009). Multi-instance learning by treating instances as non-i.i.d. samples. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1249–1256, Montreal, QC, Canada.

Zhu, J., Wang, H., Tsou, B. K., & Zhu, M. (2009). Multi-aspect opinion polling from textual reviews. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pp. 1799–1802, Hong Kong, China.

Zhu, J., Zhang, C., & Ma, M. Y. (2012). Multi-aspect rating inference with aspect-based segmentation. *IEEE Transactions on Affective Computing*, *3*(4), 469–481.

Zhu, J.-Y., Wu, J., Xu, Y., Chang, E., & Tu, Z. (2015). Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(4), 862–875.

Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pp. 43–50, Arlington, VA, USA.

Zinkevich, M., Weimer, M., Smola, A. J., & Li, L. (2011). Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems 23*, NIPS '10, pp. 2595–2603.