Explaining the Stars: Weighted Multiple-Instance Learning for Aspect-Based Sentiment Analysis

Nikolaos Pappas
EPFL and Idiap Research Institute
Rue Marconi 19
CH-1920 Martigny, Switzerland
nikolaos.pappas@idiap.ch

Andrei Popescu-Belis
Idiap Research Institute
Rue Marconi 19
CH-1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch

Abstract

This paper introduces a model of multiple-instance learning applied to the prediction of aspect ratings or judgments of specific properties of an item from user-contributed texts such as product reviews. Each variable-length text is represented by several independent feature vectors; one word vector per sentence or paragraph. For learning from texts with known aspect ratings, the model performs multiple-instance regression (MIR) and assigns importance weights to each of the sentences or paragraphs of a text, uncovering their contribution to the aspect ratings. Next, the model is used to predict aspect ratings in previously unseen texts, demonstrating interpretability and explanatory power for its predictions. We evaluate the model on seven multi-aspect sentiment analysis data sets, improving over four MIR baselines and two strong bag-of-words linear models, namely SVR and Lasso, by more than 10% relative in terms of MSE.

1 Introduction

Sentiment analysis of texts provides a coarse-grained view of their overall attitude towards an item, either positive or negative. The recent abundance of user texts accompanied by real-valued labels e.g. on a 5-star scale has contributed to the development of automatic sentiment analysis of reviews of items such as movies, books, music or other products, with applications in social computing, user modeling, and recommender systems. The overall sentiment of a text towards an item often results from the ratings of several specific aspects of the item. For instance, the author of a review might have a rather positive sentiment about a movie, having particularly liked the plot and the music, but not too much the actors. Determining the ratings of each aspect automatically is a challenging task, which may seem to require the engineering of a large number of features designed to capture each aspect. Our goal is to put forward a new feature-agnostic solution for analyzing aspect-related ratings expressed in a text, thus aiming for a finer-grained, deeper analysis of text meaning than overall sentiment analysis.

Current state-of-the-art approaches to sentiment analysis and aspect-based sentiment analysis, attempt to go beyond word-level features either by using higher-level linguistic features such as POS tagging, parsing, and knowledge infusion, or by learning features that capture syntactic and semantic dependencies between words. Once an appropriate feature space is found, the ratings are typically modeled using a linear model, such as Support Vector Regression (SVR) with ℓ_2 norm for regularization or Lasso Regression with ℓ_1 norm. By treating a text globally, these models ignore the fact that the sentences of a text have diverse contributions to the overall sentiment or to the attitude towards a specific aspect of an item.

In this paper, we propose a new learning model which answers the following question: "To what extent does each part of a text contribute to the prediction of its overall sentiment or the rating of a particular aspect?" The model uses multiple-instance regression (MIR), based on the assumption that not all the parts of a text have the same contribution to the prediction of the rating. Specifically, a text is seen as a bag of sentences (instances), each of them modeled as a word vector. The overall challenge is to learn which sentences refer to a given aspect, and how they contribute to the text's attitude towards it, but the model applies to overall sentiment analysis as well. For instance, Figure 1 displays a positive global comment on a TED talk and the weights assigned to two of its sentences by MIR.

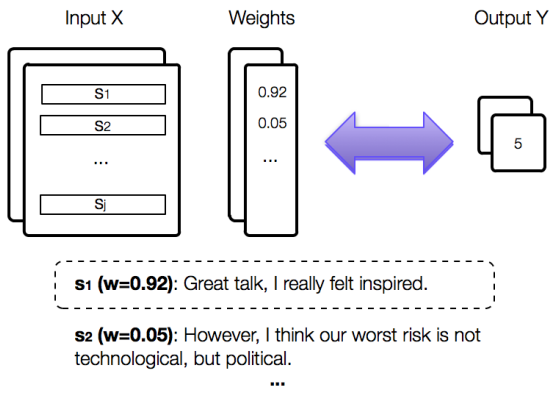


Figure 1: Analysis of a comment (bag of sentences $\{s_1, ..., s_j\}$) annotated by humans with the maximal positive sentiment score (5 stars). The weights assigned by MIR reveal that s_1 has the greatest relevance to the overall sentiment.

Using regularized least squares, we formulate an optimization objective to jointly assign instance weights and regression hyperplane weights. Then, an instance relevance estimation method is used to predict aspect ratings, or global ones, in previously unseen texts. The parameters of the model are learned using an alternating optimization procedure inspired by Wagstaff and Lane (2007). Our model requires only text with ratings for training, with no particular assumption on the word features to be extracted, and provides interpretable explanations of the predicted ratings through the relevance weights assigned to sentences. We also show that the model has reasonable computational demands. The model is evaluated on aspect and sentiment rating prediction over seven datasets: five of them contain reviews with aspect labels about beers, audiobooks and toys (McAuley et al., 2012), and two contain TED talks with emotion labels, and comments on them with sentiment labels (Pappas and Popescu-Belis, 2013). Our model outperforms previous MIR models and two strong linear models for rating prediction, namely SVR and Lasso by more than 10% relative in terms of MSE. The improvement is observed even when the sophistication of the feature space increases.

The paper is organized as follows. Section 2 shows how our model innovates with respect to previous work on MIR and rating prediction. Section 3 formulates the problem while Section 4 describes previous MIR models. Section 5 presents our MIR model and learning procedure. Section 6 presents the datasets and evaluation methods. Section 7 reports our results on rating prediction tasks, and provides examples of rating explanation.

2 Related Work

2.1 Multiple-Instance Regression

Multiple-instance regression (MIR) belongs to the class of multiple-instance learning (MIL) problems for real-valued output, and it is a variant of multiple regression where each data point may be described by more than one vectors of values. Many MIL studies focused on classification (Andrews et al., 2003; Bunescu and Mooney, 2007; Settles et al., 2008; Foulds and Frank, 2010; Wang et al., 2011) while fewer focused on regression (Ray and Page, 2001; Davis and others, 2007; Wagstaff et al., 2008; Wagstaff and Lane, 2007). Related to document analysis, several MIR studies have focused on news categorization (Zhang and Zhou, 2008; Zhou et al., 2009) or web-index recommendation (Zhou et al., 2005) but, to our knowledge, no study has attempted to use MIR for aspect rating prediction or sentiment analysis with real-valued labels.

MIR was firstly introduced by Ray et al. (2001), proposing 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 estimate instance weights per bag for crop yield modeling (not applicable to prediction). A similar method which learns the internal structure of bags using clustering was proposed by Wagstaff et al. (2008) for crop yield prediction, and we will use it for comparison in the present study. Later, the method was adapted to map bags into a single-instance feature space by Zhang et al. (2009). Wang et al. (2008) assumed that each bag is generated by random noise around a primary instance, while Wang et al. (2012) represented bag labels with a probabilistic mixture model. Foulds et al. (2010) concluded that various assumptions are differently suited to different tasks, and should be stated clearly when describing an MIR model.

2.2 Rating Prediction from Text

Sentiment analysis aims at analyzing the polarity of a given text, either with classification (for discrete labels) or regression (for real-valued labels). Early studies introduced machine learning techniques for sentiment classification, e.g. Pang et al. (2002), including unsupervised techniques based on the notion of semantic orientation of phrases, e.g. Turney et al. (2002). Other studies focused on subjectivity detection, i.e. whether a

text span expresses opinions or not (Wiebe et al., 2004). Rating inference was defined by Pang et al. (2005) as multi-class classification or regression with respect to rating scales. Pang and Lee (2008) discusses the large range of features engineered for this task, though several recent studies focus on feature learning (Maas et al., 2011; Socher et al., 2011), including the use of a deep neural network (Socher et al., 2013). In contrast, we do not make any assumption about the nature or dimensionality of the feature space.

The fine-grained analysis of opinions regarding 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 and Liu, 2004; Zhuang et al., 2006). Most attempts to perform this task have engineered various feature sets, augmenting words with topic or content models (Mei et al., 2007; Titov and McDonald, 2008; Sauper et al., 2010; Lu et al., 2011), or with linguistic features (Pang and Lee, 2005; Baccianella et al., 2009; Qu et al., 2010; Zhu et al., 2012). Other studies have advocated joint modeling of multiple aspects (Snyder and Barzilay, 2007) or multiple reviews for the same product (Li et al., 2011). McAuley et al. (2012) introduced new corpora of multi-aspect reviews, which we also partly use here, and proposed models for aspect detection, sentiment summarization and rating prediction. Lastly, joint aspect identification and sentiment classification have been used for aggregating product review snippets by Sauper at al. (2013). None of the above studies considers the multiple-instance property of text in their modeling.

3 MIR Definition

Let us consider a set B of m bags with numerical labels Y as input data $D = \{(\{b_{1j}\}_{n_1}^d, y_1), ..., (\{b_{mj}\}_{n_m}^d, y_m)\}$, where $b_{ij} \in \mathbb{R}^d$ (for $1 \leq j \leq n_i$) and $y_i \in \mathbb{R}$. Each bag B_i consists of n_i data points (called 'instances'), hence it is a matrix of n_i d-dimensional vectors, e.g. word vectors. The challenge is to infer the label of the bag given a variable number of instances n_i. This requires finding a set of bag representations $X = \{x_1, \ldots, x_m\}$ of size m where $x_i \in \mathbb{R}^d$, from which the class labels can be computed. The goal is then to find a mapping from this representation, noted $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$, which is able to predict the label of a given bag. Ideally, assuming that X is the best bag representation for our task, we look for the optimal regression hyperplane Φ which minimizes a loss function $\mathcal{L}$ plus a regularization term Ω as follows:

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

Since the best set of representations X for a task is generally unknown, one has to make assumptions to define it or compute it jointly with the regression hyperplane Φ. Thus, the main difficulty lies in finding a good assumption for X, as we will now discuss.

4 Previous MIR Assumptions

We describe here three assumptions frequently made in past MIR studies, to which we will later compare our model: aggregating all instances, keeping them as separate examples, or choosing the most representative one (Wang et al., 2012). For each assumption, we will experiment with two state-of-the-art regression models (noted abstractly as f), namely SVR (Drucker et al., 1996) and Lasso (Tibshirani, 1996) with respectively the ℓ_2 and ℓ_1 norms for regularization.

The *Aggregated* algorithm assumes that each bag is represented as a single d-dimensional vector, which is the average of its instances (hence $x_i \in \mathbb{R}^d$). 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:

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

In fact, a simple sum can also be used instead of the mean, and we observed in practice that with an appropriate regularization there is no difference on the prediction performance between these options. 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. BOW with counts or TF-IDF weights).

The *Instance* algorithm considers each of the instances in a bag as separate examples, by labeling each of them with the bag's label. A regression model f is learned over the training set made of all vectors of all bags, $D_{ins} = \{(b_{ij}, y_i) \mid j = 1, \ldots, n_i; i = 1, \ldots, m\}$, assuming that there are m labeled bags. To label a new bag B_i, given that

there is no representation x_i, the method simply averages the predicted labels of its instances:

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

Instead of the average, the median value can also be used, which is more appropriate when the bags contain outlying instances.

The *Prime* algorithm assumes that a single instance in each bag is responsible for its label (Ray and Page, 2001). This instance is called the primary or prime one. The method 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 and m is the number of bags. The prime instances are discovered through an iterative algorithm which refines the regression model f. Given an initial model f, in each iteration the algorithm selects from each bag a prime candidate which is the instance with the lowest prediction error. Then, a new model is trained over the selected prime candidates, until convergence. For a new bag, the target class is computed as in Eq. 3.

5 Proposed MIR Model

We propose a new MIR model which assigns individual relevance values (weights) to each instance of a bag, thus making fewer simplifying assumptions than previous models. We extend instance-relevance algorithms such as (Wagstaff and Lane, 2007) by supporting high-dimensional feature spaces, as required for text regression, and by predicting both the class label and the content structure of previously unseen (hence unlabeled) bags. The former is achieved by minimizing a regularized least squares loss (RLS) instead of solving normal equations, which is prohibitive in large spaces. The latter represents a significant improvement over *Aggregated* and *Instance* algorithms, which are unable to pinpoint the most relevant instances with respect to the label of each bag, being thus applicable only to bag label prediction. Similarly, *Prime* only identifies the prime instance when the bag is already labeled. Instead, our model learns an optimal method to aggregate instances, rather than a pre-defined one, and allows more degrees of freedom in the regression model than previous ones. Moreover, the weight of an instance is interpreted as its relevance both in training and prediction.

5.1 Instance Relevance Assumption

Each bag defines a bounded region of a hyperplane orthogonal to the y-axis (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}:

$$x_i = \sum_{j=1}^{n_i} \psi_{ij} b_{ij}, \psi_{ij} \geq 0 \text{ and } \sum_{j=1}^{n_i} \psi_{ij} = 1 \quad (4)$$

where ψ_{ij} is the weight of the j^{th} instance of the i^{th} bag. Each weight ψ_{ij} indicates the relevance of an instance j to the prediction of the class y_i of the i^{th} bag. The constraint forces x_i to fall within the bounded region of the points in bag i and guarantees that the i^{th} bag will influence the regressor.

5.2 Modeling Bag Structure and Labels

Let us consider a set of m bags, where each bag B_i is represented by its n_i d-dimensional instances, i.e. $B_i = \{b_{ij}\}_{n_i}^d$ along with the set of target class labels for each bag, $Y = \{y_i\}_N, y_i \in \mathbb{R}$. The representation set of all B_i in the feature space, $X = \{x_1, \ldots, x_m\}, x_i \in \mathbb{R}^d$, is obtained using the n_i instance weights associated to each bag B_i, $\psi_i = \{\psi_{ij}\}_{n_i}$, $\psi_{ij} \in [0, 1]$ which are initially unknown. Thus, we look for a linear regression model f that is able to model the target values using the regression coefficients $\Phi \in \mathbb{R}^d$, where X and Y are respectively the sets of training bags and their labels: $Y = f(X) = \Phi^T X$. We define a loss function according to the least squares objective dependent on X, Y, Φ and the set of weight vectors $\Psi = \{\psi_1, \ldots, \psi_m\}$ using Eq. 4 as follows:

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

Using the above loss function, accounting for the constraints of our assumption in Eq. 4 and assuming ℓ_2-norm for regularization with ϵ_1 and ϵ_2 terms for each $\psi_i \in \Psi$ and Φ respectively, we obtain the

following least squares objective from Eq. 1:

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

$$\text{where } \Delta_i^2 = \Big(y_i - \Phi^T(B_i\psi_i)\Big)^2, \qquad (6)$$

subject to $\psi_{ij} \geq 0\ \forall i,j$ and $\sum_{j=1}^{n_i}\psi_{ij} = 1\ \forall i$. The selection of the ℓ_2-norm was based on preliminary results showing that it outperforms ℓ_1-norm. Other combinations of p-norm regularization can be explored for f_1 and f_2, e.g. to learn sparser instance weights and denser regression coefficients or vice versa.

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. As indicated in Eq. 6 above, we will note f_1 a model that is learned from the minimization only with respect to $\psi_1,\ldots,\psi_m$ and f_2 a model obtained from the minimization with respect to Φ only. In Eq. 6, we can observe that if one of the two is known or held fixed, then the other one is convex and can be learned with the well-known least squares solving techniques. In Section 5.3, we will describe an algorithm that is able to exploit this observation.

Having computed $\psi_1,\ldots,\psi_m$ and Φ, we could predict a label for an unlabeled bag using Eq. 3, but would not be able to compute the weights of the instances. Moreover, information that has been learned about the instances during the training phase would not be used during prediction. For these reasons, we introduce a third regression model f_3 with regression coefficients $O \in \mathbb{R}^d$ assuming a ℓ_2-norm for the regularization with ϵ_3 term, which is trained on the relevance weights obtained from the Eq. 6, $D_w = \{(b_{ij},\psi_{ij}) \mid i = 1,...,m; j = 1,...,n_i\}$. The optimization objective for the f_3 model is the following:

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

This minimization can be easily performed with the well-known least squares solving techniques. The learned model is able to estimate the weights of the instances of an unlabeled bag during prediction time as: $\hat{\psi}_i = f_3(B_i) = \Omega^T B_i$. The $\hat{\psi}_i$ weights are estimations which are influenced by the relevance weights learned in our minimization objective of Eq. 6 but they are not constrained at prediction time. To obtain interpretable weights, we can convert the estimated scores to the $[0,1]$ interval as follows: $\hat{\psi}_i = \hat{\psi}_i / sum(\hat{\psi}_i)$. Finally, the prediction of the label for the i^{th} bag using the estimated instance weights $\hat{\psi}_i$ is done as follows:

$$\hat{y} = f_2(B_i) = \Phi^T B_i \hat{\psi}_i \qquad (8)$$

5.3 Learning with Alternating Projections

Algorithm 1 solves the non-convex optimization problem of Eq. 6 by using a powerful class of methods for finding the intersection of convex sets, namely alternating projections (AP). The problem is firstly divided into two convex problems, namely f_1 loss function and f_2 loss function, which are then solved in an alternating fashion. Like EM algorithms, AP algorithms do not have general guarantees on their convergence rate, although, in practice, we found it acceptable at generally fewer than 20 iterations.

Algorithm 1 APWeights(B, Y, ϵ_1, ϵ_2, ϵ_3)

```
Initialize(ψ_1, ..., ψ_N, Φ, X)
while not converged do
    for B_i in B do
        ψ_i = cRLS(Φ^T Bi, Y_i, ε_1) # f_1 model
        x_i = B_i ψ_i^T
    end for
    Φ = RLS(X, Y, ε_2) # f_2 model
end while
Ω = RLS({b_ij ∀i,j}, {ψ_ij ∀i,j}, ε_3) # f_3 model
```

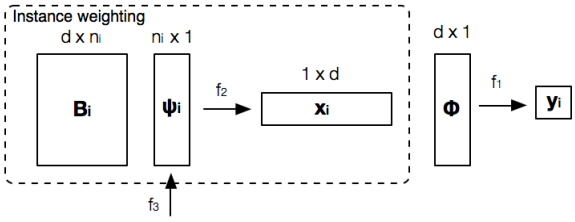


Figure 2: Visual representation for the training and testing procedure of Algorithm 1.

The algorithm takes as input the bags B_i, their target class labels Y and the regularization terms $\epsilon_1, \epsilon_2, \epsilon_3$ and proceeds as follows. First, under a fixed regression model (f_2), it proceeds with f_1 to the optimal assignment of weights to the instances of each bag (projection of Φ vectors on the ψ_i space which is a n_i-simplex) and computes its new representation set X. Second, given the fixed instance weights, it trains a new regression model (f_2) using X (projection back to the Φ

	Bags		Instances		Dimension	Aspect ratings
Dataset	**Type**	**Count**	**Type**	**Count**	**Count**	**Classes**
BeerAdvocate	review	1,200	sentence	12,189	19,418	feel, look, smell, taste, overall
RateBeer (ES)		1,200		3,269	2,120	appearance, aroma, overall, palate, taste
RateBeer (FR)		1,200		4,472	903	appearance, aroma, overall, palate, taste
Audiobooks		1,200		4,886	3,971	performance, story, overall
Toys & Games		1,200		6,463	31,984	educational, durability, fun, overall
TED comments	comment	1,200	sentence	3,814	957	sentiment (polarity)
TED talks	comments per talk	1,200	comment	11,993	5,000	unconvincing, fascinating, persuasive, ingenious, longwinded, funny, inspiring, jaw-dropping, courageous, beautiful, confusing, obnoxious

Table 1: Description of the seven datasets used for aspect, sentiment and emotion rating prediction.

space). This procedure repeats until convergence, i.e. when there is no more decrease on the training error, or until a maximum number of iterations has been reached. The regression model f_3 is trained on the weights learned from the previous steps.

5.4 Complexity Analysis

The overall time complexity T of Algorithm 1 in terms of the input variables, noted $h = \{m, \hat{n}, d\}$, with m being the number of bags, $\hat{n}$ the average size of the bags, and d the dimensionality of the feature space (here, the size of word vectors), is derived as follows:

$$\begin{aligned} T(h) &= T_{ap}(h) + T_{f_3}(h) \\ &= \mathrm{O}\big(m(\hat{n}^2 + d^2)\big) + \mathrm{O}\big(m\hat{n}d^2\big) \\ &= \mathrm{O}\big(m(\hat{n}^2 + d^2 + \hat{n}d^2)\big), \end{aligned} \tag{9}$$

where T_{ap} and T_{f_3} are respectively the time complexity of the AP procedure and of training the f_3 model. Eq. 9 shows that when $\hat{n} \ll m$, the model complexity is linear with the input bags m and always quadratic with the number of features d.

Previous works on relevance assignment for MIR have prohibitive complexity for high-dimensional feature spaces or numerous bags and hence they are not most appropriate for text regression tasks. Wagstaff and Lane (2007) have cubic time complexity with the average bag size $\hat{n}$ and number of features d; Zhou et al. (2009) use kernels, thus their complexity is quadratic with the number of bags m; and Wang et al. (2011) have cubic time wrt. d. Our formulation is thus competitive in terms of complexity.

6 Data, Protocol and Metrics

6.1 Aspect Rating Datasets

We use seven datasets summarized in Table 1. Five publicly available datasets were built for aspect prediction by McAuley et al. (2012) – BeerAdvocate, Ratebeer (ES), RateBeer (FR), Audiobooks and Toys & Games – and have aspect ratings assigned by their creators on the respective websites. On the set of comments on TED talks from Pappas and Popescu-Belis (2013), we aim to predict two things: talk-level emotion dimensions assigned by viewers through voting, and comment polarity scores assigned by crowdsourcing. The distributions of aspect ratings per dataset are shown in Figure 3. Five datasets are in English, one in Spanish (Ratebeer) and one in French (RateBeer), so our results will also demonstrate the language-independence of our method.

From every dataset we kept 1,200 texts as bags of sentences, but we also used three *full-size* datasets, namely Ratebeer ES (1,259 labeled reviews), Ratebeer FR (17,998) and Audiobooks (10,989). The features for each of them are word vectors with binary attributes signaling word presence or absence, in a traditional bag-of-words model (BOW). The word vectors are provided with the first five datasets and we generated them for the latter two, after lowercasing and stopword removal. Moreover, for TED comments, we computed TF-IDF scores using the same dimensionality as with BOW to experiment with a different feature space. The target class labels were normalized by the maximum rating in their scale, except for TED talks where the votes were normalized by the maximum number of votes over all the emotion classes for each talk, and two emotions, 'informative' and 'ok', were excluded as they are neutral ones.

6.2 Evaluation Protocol

We compare the proposed model, noted *AP-Weights*, with four baseline ones – *Aggregated*, *Instance*, *Prime* (Section 4) and *Clus-*

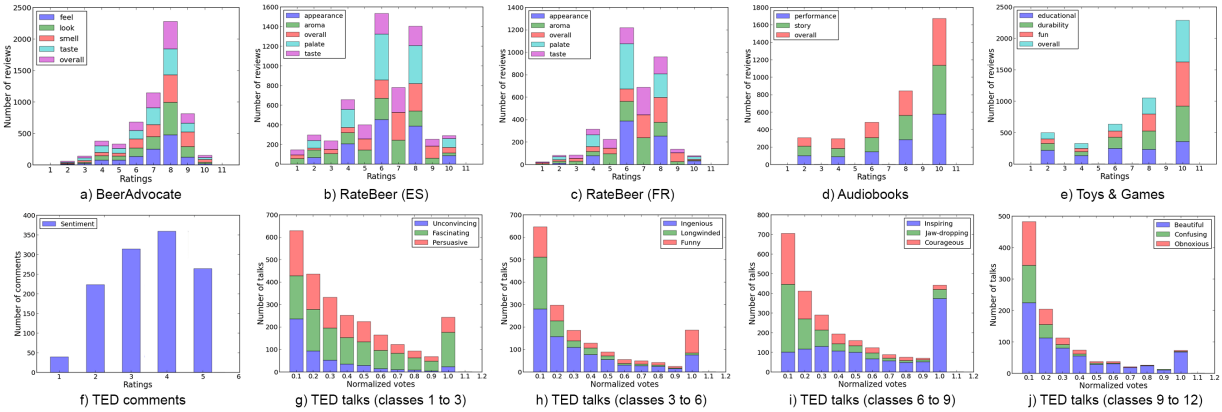


Figure 3: Distributions of rating values per aspect rating class for the seven datasets.

tering (from `github.com/garydoranjr/mcr`), which is an instance relevance method proposed by Wagstaff et al. (2008) for aspect rating prediction. First, for each aspect class, we optimize all methods on a development set of 25% of the data (300 randomly selected bags). Then, we perform 5-fold cross-validation for every aspect on each entire data set and report the average error scores using the optimal hyper-parameters per method. In addition, we report for comparison the scores of *AverageRating*, which always predicts the average rating over the training set.

We report standard error metrics for regression, namely the Mean Absolute Error (MAE) and the Mean Squared Error (MSE). The former measures the average magnitude of errors in a set of predictions while the latter measures the average of their squares, which are defined over the test set of bags B_i respectively as $\text{MAE} = (\sum_{i=1}^{k} |f(B_i) - y_i|)/k$ and $\text{MSE} = (\sum_{i=1}^{k} (f(B_i) - y_i)^2)/k$. The cross-validation scores are obtained by averaging the MAE and MSE scores on each fold.

To find the optimal hyper-parameters for each model, we perform 3-fold cross-validation on the development set using exhaustive grid-search over a fine-grained range of possible values and select the ones that perform best in terms of MAE. The hyper-parameters to be optimized for the baselines (except AverageRating) are the regularization terms λ_2, λ_1 of their possible regression model f, namely SVR which uses the ℓ_2 norm and Lasso which uses the ℓ_1 norm. As for APWeights, it relies on three regularization terms, namely $\epsilon_1, \epsilon_2, \epsilon_3$ of the ℓ_2-norm for f_1, f_2 and f_3 regression models. Lastly, for the Clustering baseline, 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, here k-Means. 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. For the regression models and evaluation protocol, we use the *scikit-learn* machine learning library (Pedregosa et al., 2012). Our code and data are available in the first author's website.

7 Experimental Results

7.1 Aspect Rating Prediction

The results for aspect rating prediction are given in Table 2. The proposed APWeights method outperforms Aggregated (ℓ_2) and Aggregated (ℓ_1) i.e. SVR and Lasso along with all other baselines on each case. The SVR baseline has on average 11% lower performance than APWeights in terms of MSE and about 6% in terms of MAE. Similarly, the Lasso baseline has on average 13% lower MSE and 8% MAE than APWeights. As shown in Figure 4, APWeights also outperforms them for each aspect in the five review datasets. The Instance method with ℓ_1 performed well on BeerAdvocate and Toys & Games (for MSE), and with ℓ_2 performed well on Ratebeer (ES), RateBeer (FR) and Toys & Games (for MAE). Therefore, the instance-as-example assumption is quite appropriate for this task, however both options score below APWeights – by about 5% MAE, and 8%/9% MSE, respectively. The Prime method with ℓ_1 performed well only on the BeerAdvocate dataset and Prime with ℓ_2 only on the Toys & Games dataset, always with lower scores than APWeights, namely about 9% MAE for both and 15%/18% MSE respectively. This suggests that the primary-instance

	REVIEW LABELS									
	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
APWeights (ℓ_2)	**12.24**	**2.66**	**14.18**	**3.28**	**11.37**	**2.27**	**18.89**	**5.71**	**18.50**	**5.57**
APW vs. SVR (%)	*+16.0*	*+27.7*	*+2.0*	*+3.8*	*+7.6*	*+15.6*	+1.0	*+4.5*	*+2.6*	*+6.0*
APW vs. Lasso (%)	*+10.1*	*+15.1*	*+11.0*	*+18.4*	*+6.8*	*+11.8*	*+6.0*	*+6.9*	*+8.1*	*+11.9*
APW 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 2: Performance of aspect rating prediction (the lower the better) in terms of MAE and MSE ($\times$ 100) with 5-fold cross-validation. All scores are averaged over all aspects in each dataset. The scores of the best method are in **bold** and the second best ones are underlined. Significant improvements (paired t-test, $p < 0.05$) are in *italics*. Fig. 4 shows MSE scores per aspect for three methods on five datasets.

assumption is not the most appropriate for this task. Lastly, even though Clustering is an instance relevance method, it has similar scores to Prime, presumably because the relevances are assigned according to the computed clusters and they are not directly influenced by the task's objective.

To compare with the state-of-the-art results obtained by McAuley et al. (2012), we experimented with three of their *full-size* datasets. Splitting each dataset in half for training vs. testing, and using the optimal settings from our experiments above, we measured the average MSE over all aspects. APWeights improved over Lasso by 10%, 26% and 17% MSE respectively on each dataset – the absolute MSE scores are .038 for Lasso vs. .034 for APWeights on Ratebeer SP; .023 vs. .017 on Ratebeer FR; .063 vs. .052 on Audiobooks. Similarly, when compared to the best SVM baseline provided by the McAuley et al., our method improved by 32%, 43% and 35% respectively on each dataset, though it did not use their rating model. Moreover, the best model proposed by McAuley et al., which uses a joint rating model and an aspect-specific text segmenter trained on hand-labeled data, reaches MSE scores of .03, .02 and .03, which is comparable to our model that does not use these features (.034, .017, .052), though it could benefit from them in the future. Lastly, as mentioned by the same authors, predictors which use segmented text, for example with topic models as in (Lu et al., 2011), do not necessarly outperform SVR baselines; instead they have marginal or even no improvements, therefore, we did not further experiment with them. Interestingly, multiple-instance learning algorithms under several assumptions go beyond SVR baselines with BOW and even more sophisticated features such as TF-IDF (see below).

	SENT. LABELS		EMO. LABELS	
	TED comm.		**TED talks**	
Model \ Error	**MAE**	**MSE**	**MAE**	**MSE**
AverageRating	19.47	5.05	17.86	6.06
Aggregated (ℓ_1)	17.08	4.17	15.98	5.03
Aggregated (ℓ_2)	16.88	4.47	15.24	4.97
Instance (ℓ_1)	17.69	4.37	16.48	5.30
Instance (ℓ_2)	16.93	4.24	16.10	5.57
Prime (ℓ_1)	17.39	4.37	15.98	5.78
Prime (ℓ_2)	18.03	4.91	16.74	5.94
Clustering (ℓ_2)	17.64	4.34	17.71	6.02
APWeights (ℓ_2)	**15.91**	**3.95**	**15.02**	**4.89**
APW vs SVR (%)	*+5.7*	*+11.5*	+1.5	+1.6
APW vs Lasso (%)	*+6.8*	+5.3	*+6.0*	*+2.9*
APW vs 2^{nd} (%)	*+5.7*	+5.3	+1.5	+1.6

Table 3: MAE and MSE ($\times$ 100) on sentiment and emotion prediction with 5-fold c.-v. Scores on TED talks are averaged over the 12 emotions. The scores of the best method are in **bold** and the second best ones are underlined. Significant improvements (paired t-test, $p < 0.05$) are in *italics*.

7.2 Sentiment and Emotion Prediction

Our method is also competitive for sentiment prediction over comments on TED talks, as well as for talk-level emotion prediction with 12 dimensions from subsets of 10 comments on each talk (see Table 3). APWeights outperforms SVR and Lasso, as well as all other methods for each task. For sentiment prediction, SVR is outperformed by 11% MSE and Lasso by 5%. For emotion pre-

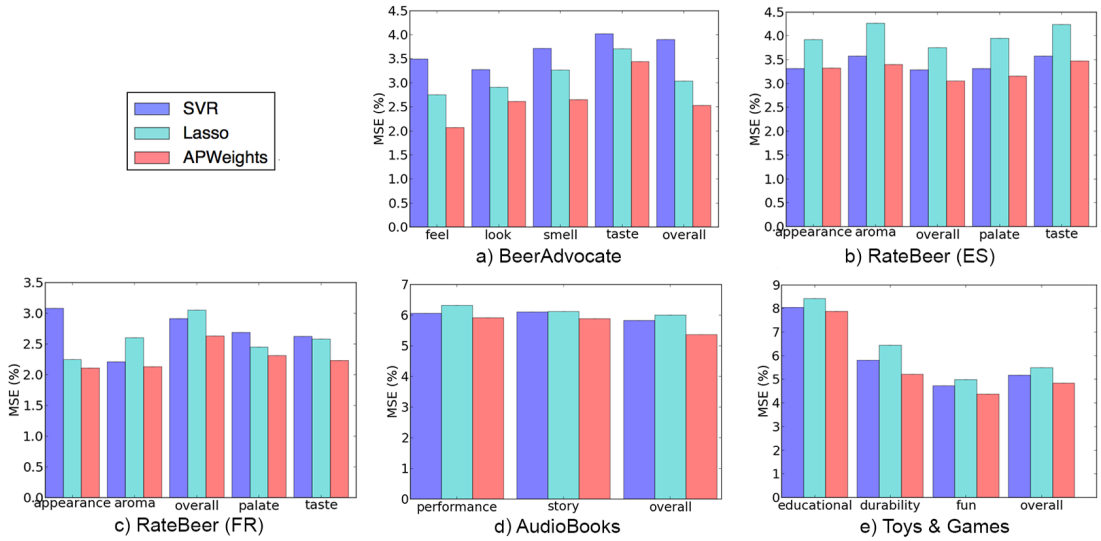


Figure 4: MSE scores of SVR, Lasso and APWeights for each aspect over the five review datasets.

diction (averaged over all 12 aspects), differences are smaller, at 1.6% and 2.9% respectively. These smaller differences could be explained by the fact that among the 10 most recent comments for each talk, many are not related to the emotion that the system tries to predict.

As mentioned earlier, the proposed model does not make any assumption about the feature space. Thus, we examined whether the improvements it brings remain present even with a different feature space, for instance based on TF-IDF instead of BOW with counts. For sentiment prediction on TED comments, we found 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 (16.25 and 16.59 MAE; 4.16 and 3.97 MSE respectively). However, APWeights still outperforms them on both MAE and MSE scores (15.35 and 3.63), 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. These promising results suggest that improvements with APWeights could be observed also on more sophisticated feature spaces.

7.3 Interpreting the Relevance Weights

Apart from predicting ratings, the MIR scores assigned by our model reflect the contribution of each sentence to these predictions.

To illustrate the explanatory power of our model (until a dataset for quantitative analysis becomes available), we provide examples of predictions on test data taken from the cross-validation folds above. Table 5 displays the most relevant comment for two correctly predicted emotions on two TED talks, based on the $\hat{\psi}_i$ relevance scores, along with the $\hat{\psi}_i$ scores of the other comments, for two emotion classes: 'beautiful' and 'courageous'. These comments appear to reflect correctly the fact that the respective emotion is the majority one in each of the comments. As noted earlier, this task is quite challenging since we use only the ten most recent comments for each talk.

Table 4 displays four TED comments selected

Sentences per comment	$\hat{\psi}_i$	$\hat{y}_i$	y_i
"Very brilliant and witty, as well as great improvisation."	0.64	5.0	5.0
"I enjoyed this one a lot."	0.36		
"That's great idea, I really like it!"	0.56	4.2	4.0
"I can't wait to try it, but first thing, I need a house with big windows, next year, maybe I can do that."	0.44		
"Unfortunately countries are not led by gifted children."	0.48	2.4	2.0
"They are either dictated by the most extreme personalities who crave nothing but power or managed by politicians who are voted in by a far from gifted population."	0.52		
"I am very disappointed by this, smug, cliched and missing so much information as to be almost (...)"'	0.43	1.8	1.0
"No mention of ship transport lets say 50% of all material transport, no mention of rail transport, (...)"	0.29		
"I am sorry to be so negative, this just sounds like a sales pitch that he has given too many times (...)."	0.28		

Table 4: Predicted sentiment for TED comments: y_i is the actual sentiment, $\hat{y}_i$ the predicted one, and $\hat{\psi}_i$ the estimated relevance of each sentence.

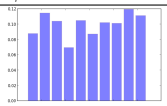
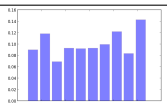
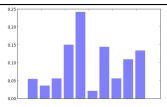
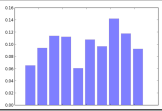
Class	Top comment per talk (according to weights ψ_i)	$\hat{\psi}_i$ distribution
inspiring	"It seems to me that the idea worth spreading of this TED Talk is inspiring and key for a full life. 'No-one else is the authority on your potential. You're the only person that decides how far you go and what you're capable of.' It seems to me that teens actually think that. As a child one is all knowing and all capable. How did we get to the (...)"	
beautiful	"The beauty of the nature. It would be more interesting just integrates his thought and idea into a mobile device, like a mobile, so we can just turn on the nature gallery in any time. The paintings don't look incidental but genuinely thought out, random perhaps, but with a clear grand design behind the randomness. Drawing is an art where it doesn't (...)"	
funny	"Funny story, but not as funny as a good 'knock, knock' joke. My favorite knock-knock joke of all time is Cheech & Chong's 'Dave's Not Here' gag from the early 1970s. I'm still waiting for someone to top it after all these years. [Knock, knock] 'Who is it?' the voice of an obviously stoned male answers from the other side of a door, (...)"	
courageous	"I was a soldier in Iraq and part of the unit represented in this documentary. I would question anyone that told you we went over there to kill Iraqi people. I spent the better part of my time in Iraq protecting the Iraqi people from insurgents who came from countries outside of Iraq to kill Iraqi people. We protected families men, women, and (...)"	

Table 5: Two examples of top comments (according to weights ψ_i) for correctly predicted emotions in four TED talks (score 1.0) and the distribution of weights over the 10 most recent comments in each talk.

Figure 5: Top words based on Φ for predicting four emotions from comments on TED talks.

from the test set of a given fold, for the comment-level sentiment prediction task. The table also shows the $\hat{\psi}_i$ relevance scores assigned to each of the composing sentences, the predicted polarity scores $\hat{y}_i$ and the actual ones y_i. We observe that the sentences that convey the most sentiment are assigned higher scores than sentences with less sentiment, always with respect to the global polarity level. These examples suggest that, given that APWeights has more degrees of freedom for interpretation, it is able to assign relevance to parts of a text (here, sentences) and even to words, while other models can only consider words. Hence, the assigned weights might be useful for other NLP tasks mentioned below.

8 Conclusion and Future Work

This paper introduced a novel MIR model for aspect rating prediction from text, which learns instance relevance together with target labels. To the best of our knowledge, this has not been considered before. Compared to previous work on MIR, the proposed model is competitive and more efficient in terms of complexity. Moreover, it is not only able to assign instance relevances on labeled bags, but also to predict them on unseen bags.

Compared to previous work on aspect rating prediction, our model performs significantly better than BOW regression baselines (SVR, Lasso) without using additional knowledge or features. The improvements persist even when the sophistication of the features increases, suggesting that our contribution may be orthogonal to feature engineering or learning. Lastly, the qualitative evaluation on test examples demonstrates that the parameters learned by the model are not only useful for prediction, but they are also interpretable.

In the future, we intend to test our model on sentiment classification at the sentence-level, based only on document-level supervision (Täckström and McDonald, 2011). Moreover, we will experiment with other model settings, such as regularization norms other than ℓ_2 and feature spaces other than BOW or TF-IDF. In the longer term, we plan to investigate new methods to estimate instance weights at prediction time, and to evaluate the impact of assigned weights on sentence ranking, segmentation or summarization.

Acknowledgments

The work described in this article was supported by the European Union through the inEvent project FP7-ICT n. 287872 (see `http://www.inevent-project.eu`).

References

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. 2003. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, pages 561–568, Vancouver, British Columbia, Canada.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Multi-facet rating of product reviews. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 461–472. Springer Berlin Heidelberg.

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

Jesse Davis et al. 2007. Tightly integrating relational learning and multiple-instance regression for real-valued drug activity prediction. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 425–432, Corvallis, OR, USA.

Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1996. Support vector regression machines. In *Advances in Neural Information Processing systems*, pages 155–161, Denver, CO, USA.

James Foulds and Eibe Frank. 2010. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25:1:1–25.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, KDD '04, pages 168–177, Seattle, WA.

Fangtao Li, Nathan Liu, Hongwei Jin, Kai Zhao, Qiang Yang, and Xiaoyan Zhu. 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, pages 1820–1825, Barcelona, Catalonia, Spain.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *Proceedings of the 11th IEEE International Conference on Data Mining Workshops*, ICDMW '11, pages 81–88, Washington, DC.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 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, pages 142–150, Portland, OR.

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

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th Int. Conf. on the World Wide Web*, WWW '07, pages 171–180, Banff, AB.

Bo Pang and Lillian Lee. 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, pages 115–124, Ann Arbor, MI.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL Conf. on Empirical Methods in Natural Language Processing*, EMNLP '02, pages 79–86, Philadelphia, PA.

Nikolaos Pappas and Andrei Popescu-Belis. 2013. Sentiment analysis of user comments for one-class collaborative filtering over TED talks. In *36th ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, Dublin, Ireland.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2012. Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490.

Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 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, pages 913–921, Beijing, China.

Soumya Ray and David Page. 2001. Multiple instance regression. In *Proceedings of the 18th International Conference on Machine Learning*, ICML '01, pages 425–432.

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

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2010. Incorporating content structure into text analysis applications. In *Proceedings of the 2010 Conference on Empirical Methods in Natural*

Language Processing, EMNLP '10, pages 377–387, Cambridge, MA.

Burr Settles, Mark Craven, and Soumya Ray. 2008. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, NIPS '08, pages 1289–1296, Vancouver, BC.

Benjamin Snyder and Regina Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '07, pages 300–307, Rochester, NY, USA.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 151–161, Edinburgh, UK.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 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, pages 1631–1642, Portland, OR.

Oscar Täckström and Ryan McDonald. 2011. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 368–374, Berlin, Heidelberg. Springer-Verlag.

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

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 111–120, Beijing, China.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Philadelphia, PA.

Kiri L. Wagstaff and Terran Lane. 2007. Salience assignment for multiple-instance regression. In *ICML 2007 Workshop on Constrained Optimization and Structured Output Spaces*, Corvallis, Oregon, USA.

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

Zhuang Wang, Vladan Radosavljevic, Bo Han, Zoran Obradovic, and Slobodan Vucetic. 2008. Aerosol optical depth prediction from satellite observations by multiple instance regression. In *Proceedings of the SIAM Int. Conf. on Data Mining*, SDM '08, pages 165–176, Atlanta, GA.

Hua Wang, Feiping Nie, and Heng Huang. 2011. Learning instance specific distance for multi-instance classification. In *AAAI Conference on Artificial Intelligence*.

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

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308, September.

Min-Ling Zhang and Zhi-Hua Zhou. 2008. M3MIML: A maximum margin method for multi-instance multi-label learning. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pages 688–697, Dec.

Min-Ling Zhang and Zhi-Hua Zhou. 2009. Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence*, 31(1):47–68.

Zhi-Hua Zhou, Kai Jiang, and Ming Li. 2005. Multi-instance learning based web mining. *Applied Intelligence*, 22(2):135–147.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. 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, pages 1249–1256, Montreal, Quebec, Canada.

Jingbo Zhu, Chunliang Zhang, and Matthew Y. Ma. 2012. Multi-aspect rating inference with aspect-based segmentation. *IEEE Trans. on Affective Computing*, 3(4):469–481.

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 43–50, Arlington, VA.