Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis

Stefanos Angelidis and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
s.angelidis@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

We consider the task of fine-grained sentiment analysis from the perspective of multiple instance learning (MIL). Our neural model is trained on document sentiment labels, and learns to predict the sentiment of text segments, i.e. sentences or elementary discourse units (EDUs), without segment-level supervision. We introduce an attention-based polarity scoring method for identifying positive and negative text snippets and a new dataset which we call SPOT (as shorthand for **S**egment-level **PO**lari**T**y annotations) for evaluating MIL-style sentiment models like ours. Experimental results demonstrate superior performance against multiple baselines, whereas a judgement elicitation study shows that EDU-level opinion extraction produces more informative summaries than sentence-based alternatives.

1 Introduction

Sentiment analysis has become a fundamental area of research in Natural Language Processing thanks to the proliferation of user-generated content in the form of online reviews, blogs, internet forums, and social media. A plethora of methods have been proposed in the literature that attempt to distill sentiment information from text, allowing users and service providers to make opinion-driven decisions.

The success of neural networks in a variety of applications (Bahdanau et al., 2015; Le and Mikolov, 2014; Socher et al., 2013) and the availability of large amounts of labeled data have led to an increased focus on sentiment classification. Supervised models are typically trained on documents (Johnson and Zhang, 2015a; Johnson and Zhang, 2015b; Tang et al., 2015; Yang et al., 2016), sentences (Kim, 2014), or phrases (Socher et al., 2011;

[Rating: ★★] I had a very mixed experience at The Stand. The burger and fries were good. The chocolate shake was divine: rich and creamy. The drive-thru was horrible. It took us at least 30 minutes to order when there were only four cars in front of us. We complained about the wait and got a half–hearted apology. I would go back because the food is good, but my only hesitation is the wait.

Summary
+ The burger and fries were good
+ The chocolate shake was divine
+ I would go back because the food is good
– The drive-thru was horrible
– It took us at least 30 minutes to order

Figure 1: An EDU-based summary of a 2-out-of-5 star review with positive and negative snippets.

Socher et al., 2013) annotated with sentiment labels and used to predict sentiment in unseen texts. Coarse-grained document-level annotations are relatively easy to obtain due to the widespread use of opinion grading interfaces (e.g., star ratings accompanying reviews). In contrast, the acquisition of sentence- or phrase-level sentiment labels remains a laborious and expensive endeavor despite its relevance to various opinion mining applications, e.g., detecting or summarizing consumer opinions in online product reviews. The usefulness of finer-grained sentiment analysis is illustrated in the example of Figure 1, where snippets of opposing polarities are extracted from a 2-star restaurant review. Although, as a whole, the review conveys negative sentiment, aspects of the reviewer's experience were clearly positive. This goes largely unnoticed when focusing solely on the review's overall rating.

In this work, we consider the problem of segment-level sentiment analysis from the perspective of *Multiple Instance Learning* (MIL; Keeler, 1991).

17

Instead of learning from individually labeled segments, our model only requires document-level supervision and learns to introspectively judge the sentiment of constituent segments. Beyond showing how to utilize document collections of rated reviews to train fine-grained sentiment predictors, we also investigate the granularity of the extracted segments. Previous research (Tang et al., 2015; Yang et al., 2016; Cheng and Lapata, 2016; Nallapati et al., 2017) has predominantly viewed documents as sequences of sentences. Inspired by recent work in summarization (Li et al., 2016) and sentiment classification (Bhatia et al., 2015), we also represent documents via Rhetorical Structure Theory's (Mann and Thompson, 1988) *Elementary Discourse Units* (EDUs). Although definitions for EDUs vary in the literature, we follow standard practice and take the elementary units of discourse to be clauses (Carlson et al., 2003). We employ a state-of-the-art discourse parser (Feng and Hirst, 2012) to identify them.

Our contributions in this work are three-fold: a novel multiple instance learning neural model which utilizes document-level sentiment supervision to judge the polarity of its constituent segments; the creation of SPOT, a publicly available dataset which contains **S**egment-level **PO**lari**T**y annotations (for sentences and EDUs) and can be used for the evaluation of MIL-style models like ours; and the empirical finding (through automatic and human-based evaluation) that neural multiple instance learning is superior to more conventional neural architectures and other baselines on detecting segment sentiment and extracting informative opinions in reviews.[1]

2 Background

Our work lies at the intersection of multiple research areas, including sentiment classification, opinion mining and multiple instance learning. We review related work in these areas below.

Sentiment Classification Sentiment classification is one of the most popular tasks in sentiment analysis. Early work focused on unsupervised methods and the creation of sentiment lexicons (Turney, 2002; Hu and Liu, 2004; Wiebe et al., 2005; Baccianella et al., 2010) based on which the overall po-

larity of a text can be computed (e,g., by aggregating the sentiment scores of constituent words). More recently, Taboada et al. (2011) introduced SO-CAL, a state-of-the-art method that combines a rich sentiment lexicon with carefully defined rules over syntax trees to predict sentence sentiment.

Supervised learning techniques have subsequently dominated the literature (Pang et al., 2002; Pang and Lee, 2005; Qu et al., 2010; Xia and Zong, 2010; Wang and Manning, 2012; Le and Mikolov, 2014) thanks to user-generated sentiment labels or large-scale crowd-sourcing efforts (Socher et al., 2013). Neural network models in particular have achieved state-of-the-art performance on various sentiment classification tasks due to their ability to alleviate feature engineering. Kim (2014) introduced a very successful CNN architecture for sentence-level classification, whereas other work (Socher et al., 2011; Socher et al., 2013) uses recursive neural networks to learn sentiment for segments of varying granularity (i.e., words, phrases, and sentences). We describe Kim's (2014) approach in more detail as it is also used as part of our model.

Let x_i denote a k-dimensional word embedding of the i-th word in text segment s of length n. The segment's input representation is the concatenation of word embeddings $x_1, \ldots, x_n$, resulting in word matrix X. Let $X_{i:i+j}$ refer to the concatenation of embeddings $x_i, \ldots, x_{i+j}$. A convolution filter $W \in \mathbb{R}^{lk}$, applied to a window of l words, produces a new feature $c_i = \mathrm{ReLU}(W \circ X_{i:i+l} + b)$, where ReLU is the *Rectified Linear Unit* non-linearity, '$\circ$' denotes the entrywise product followed by a sum over all elements and $b \in \mathbb{R}$ is a bias term. Applying the same filter to every possible window of word vectors in the segment, produces a feature map $c = [c_1, c_2, \ldots, c_{n-l+1}]$. Multiple feature maps for varied window sizes are applied, resulting in a fixed-size segment representation v via max-over-time pooling. We will refer to the application of convolution to an input word matrix X, as $\mathrm{CNN}(X)$. A final sentiment prediction is produced using a softmax classifier and the model is trained via back-propagation using sentence-level sentiment labels.

The availability of large-scale datasets (Diao et al., 2014; Tang et al., 2015) has also led to the development of document-level sentiment classifiers which exploit hierarchical neural representations.

---

[1] Our code and SPOT dataset are publicly available at: https://github.com/stangelid/milnet-sent

18

These are obtained by first building representations of sentences and aggregating those into a document feature vector (Tang et al., 2015). Yang et al. (2016) further acknowledge that words and sentences are deferentially important in different contexts. They present a model which learns to attend (Bahdanau et al., 2015) to individual text parts when constructing document representations. We describe such an architecture in more detail as we use it as a point of comparison with our own model.

Given document d comprising segments $(s_1, \ldots, s_m)$, a *Hierarchical Network* with attention (henceforth HIERNET; based on Yang et al., 2016) produces segment representations $(v_1, \ldots, v_m)$ which are subsequently fed into a bidirectional GRU module (Bahdanau et al., 2015), whose resulting hidden vectors $(h_1, \ldots, h_m)$ are used to produce attention weights $(a_1, \ldots, a_m)$ (see Section 3.2 for more details on the attention mechanism). A document is represented as the weighted average of the segments' hidden vectors $v_d = \sum_i a_i h_i$. A final sentiment prediction is obtained using a softmax classifier and the model is trained via back-propagation using document-level sentiment labels. The architecture is illustrated in Figure 2(a). In their proposed model, Yang et al. (2016) use bidirectional GRU modules to represent segments as well as documents, whereas we use a more efficient CNN encoder to compose words into segment vectors[2] (i.e., $v_i = \mathrm{CNN}(X_i)$). Note that models like HIERNET do not naturally predict sentiment for individual segments; we discuss how they can be used for segment-level opinion extraction in Section 5.2.

Our own work draws inspiration from representation learning (Tang et al., 2015; Kim, 2014), especially the idea that not all parts of a document convey sentiment-worthy clues (Yang et al., 2016). Our model departs from previous approaches in that it provides a natural way of predicting the polarity of *individual* text segments without requiring segment-level annotations. Moreover, our attention mechanism directly facilitates opinion detection rather than simply aggregating sentence representations into a single document vector.

---

[2]When applied to the YELP'13 and IMDB document classification datasets, the use of CNNs results in a relative performance decrease of $< 2\%$ compared Yang et al's model (2016).

Opinion Mining A standard setting for opinion mining and summarization (Lerman et al., 2009; Carenini et al., 2006; Ganesan et al., 2010; Di Fabbrizio et al., 2014; Gerani et al., 2014) assumes a set of documents that contain opinions about some entity of interest (e.g., camera). The goal of the system is to generate a summary that is representative of the average opinion and speaks to its important aspects (e.g., picture quality, battery life, value). Output summaries can be extractive (Lerman et al., 2009) or abstractive (Gerani et al., 2014; Di Fabbrizio et al., 2014) and the underlying systems exhibit varying degrees of linguistic sophistication from identifying aspects (Lerman et al., 2009) to using RST-style discourse analysis, and manually defined templates (Gerani et al., 2014; Di Fabbrizio et al., 2014).

Our proposed method departs from previous work in that it focuses on detecting opinions in individual documents. Given a review, we predict the polarity of every segment, allowing for the extraction of sentiment-heavy opinions. We explore the usefulness of EDU segmentation inspired by Li et al. (2016), who show that EDU-based summaries align with near-extractive summaries constructed by news editors. Importantly, our model is trained in a weakly-supervised fashion on large scale document classification datasets without recourse to fine-grained labels or gold-standard opinion summaries.

Multiple Instance Learning Our models adopt a *Multiple Instance Learning* (MIL) framework. MIL deals with problems where labels are associated with groups of instances or *bags* (documents in our case), while instance labels (segment-level polarities) are unobserved. An aggregation function is used to combine instance predictions and assign labels on the bag level. The goal is either to label bags (Keeler and Rumelhart, 1992; Dietterich et al., 1997; Maron and Ratan, 1998) or to simultaneously infer bag and instance labels (Zhou et al., 2009; Wei et al., 2014; Kotzias et al., 2015). We view segment-level sentiment analysis as an instantiation of the latter variant.

Initial MIL efforts for binary classification made the strong assumption that a bag is negative only if all of its instances are negative, and positive otherwise (Dietterich et al., 1997; Maron and Ratan, 1998; Zhang et al., 2002; Andrews and Hofmann, 2004; Carbonetto et al., 2008). Subsequent work re-

laxed this assumption, allowing for prediction combinations better suited to the tasks at hand. Weidmann et al. (2003) introduced a generalized MIL framework, where a combination of instance types is required to assign a bag label. Zhou et al. (2009) used graph kernels to aggregate predictions, exploiting relations between instances in object and text categorization. Xu and Frank (2004) proposed a multiple-instance logistic regression classifier where instance predictions were simply averaged, assuming equal and independent contribution toward bag classification. More recently, Kotzias et al. (2015) used sentence vectors obtained by a pre-trained hierarchical CNN (Denil et al., 2014) as features under an unweighted average MIL objective. Prediction averaging was further extended by Pappas and Popescu-Belis (2014; 2017), who used a weighted summation of predictions, an idea which we also adopt in our work.

Applications of MIL are many and varied. MIL was first explored by Keeler and Rumelhart (1992) for recognizing handwritten post codes, where the position and value of individual digits was unknown. MIL techniques have since been applied to drug activity prediction (Dietterich et al., 1997), image retrieval (Maron and Ratan, 1998; Zhang et al., 2002), object detection (Zhang et al., 2006; Carbonetto et al., 2008; Cour et al., 2011), text classification (Andrews and Hofmann, 2004), image captioning (Wu et al., 2015), paraphrase detection (Xu et al., 2014), and information extraction (Hoffmann et al., 2011).

When applied to sentiment analysis, MIL takes advantage of supervision signals on the document level in order to train segment-level sentiment predictors. Although their work is not couched in the framework of MIL, Täckström and McDonald (2011) show how sentence sentiment labels can be learned as latent variables from document-level annotations using hidden conditional random fields. Pappas and Popescu-Belis (2014) use a multiple instance regression model to assign sentiment scores to specific aspects of products. The Group-Instance Cost Function (GICF), proposed by Kotzias et al. (2015), averages sentence sentiment predictions during trainng, while ensuring that similar sentences receive similar polarity labels. Their work uses a pre-trained hierarchical CNN to obtain sentence embeddings, but is not trainable end-to-end, in contrast

with our proposed network. Additionally, none of the aforementioned efforts explicitly evaluate opinion extraction quality.

3 Methodology

In this section we describe how multiple instance learning can be used to address some of the drawbacks seen in previous approaches, namely the need for expert knowledge in lexicon-based sentiment analysis (Taboada et al., 2011), expensive fine-grained annotation on the segment level (Kim, 2014; Socher et al., 2013) or the inability to naturally predict segment sentiment (Yang et al., 2016).

3.1 Problem Formulation

Under multiple instance learning (MIL), a dataset D is a collection of labeled *bags*, each of which is a group of unlabeled *instances*. Specifically, each document d is a sequence (bag) of segments (instances). This sequence $d = (s_1, s_2, \ldots, s_m)$ is obtained from a document segmentation policy (see Section 4 for details). A discrete sentiment label $y_d \in [1, C]$ is associated with each document, where the labelset is ordered and classes 1 and C correspond to maximally negative and maximally positive sentiment. It is assumed that y_d is an unknown function of the unobserved segment-level labels:

$$y_d = f(y_1, y_2, \ldots, y_m) \tag{1}$$

Probabilistic sentiment classifiers will produce document-level predictions $\hat{y}_d$ by selecting the most probable class according to class distribution $\mathrm{p}_d = \langle p_d^{(1)}, \ldots, p_d^{(C)} \rangle$. In a non-MIL framework a classifier would learn to predict the document's sentiment by directly conditioning on its segments' feature representations or their aggregate:

$$\mathrm{p}_d = \hat{f}_\theta(\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_m) \tag{2}$$

In contrast, a MIL classifier will produce a class distribution p_i for each segment and additionally learn to combine these into a document-level prediction:

$$\mathrm{p}_i = \hat{g}_{\theta_s}(\mathrm{v}_i), \tag{3}$$

$$\mathrm{p}_d = \hat{f}_{\theta_d}(\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_m). \tag{4}$$

In this work, $\hat{g}$ and $\hat{f}$ are defined using a single neural network, described below.
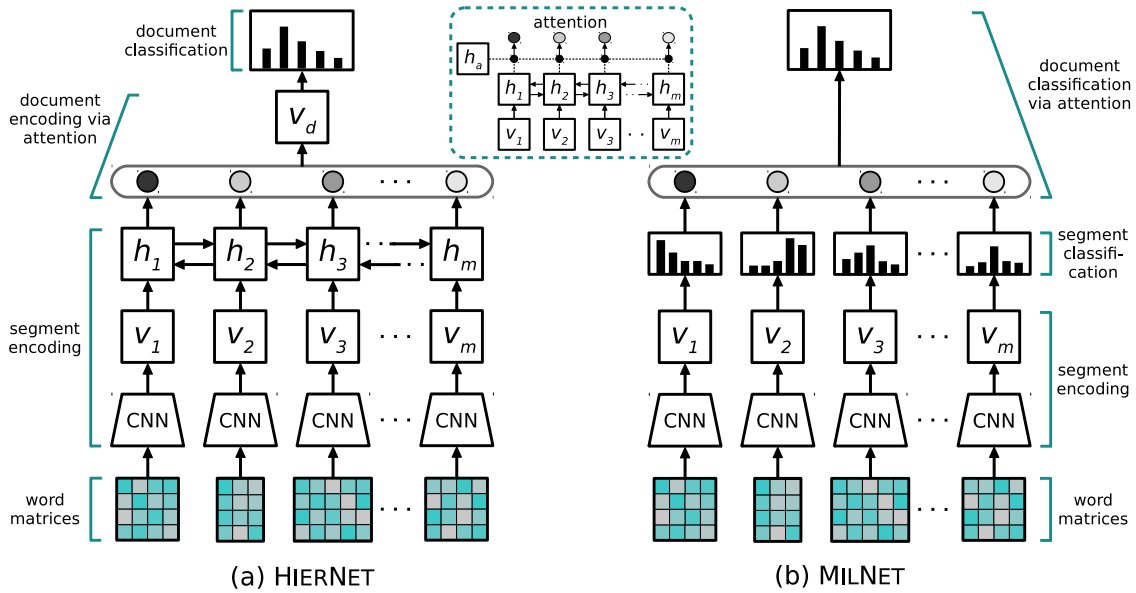
Figure 2: A *Hierarchical Network* (HIERNET) for document-level sentiment classification and our proposed *Multiple Instance Learning Network* (MILNET). The models use the same attention mechanism to combine segment vectors and predictions respectively.

3.2 Multiple Instance Learning Network

Hierarchical neural models like HIERNET have been used to predict document-level polarity by first encoding sentences and then combining these representations into a document vector. Hierarchical vector composition produces powerful sentiment predictors, but lacks the ability to introspectively judge the polarity of individual segments.

Our *Multiple Instance Learning Network* (henceforth MILNET) is based on the following intuitive assumptions about opinionated text. Each segment conveys a degree of sentiment polarity, ranging from very negative to very positive. Additionally, segments have varying degrees of importance, in relation to the overall opinion of the author. The overarching polarity of a text is an aggregation of segment polarities, weighted by their importance. Thus, our model attempts to predict the polarity of segments and decides which parts of the document are good indicators of its overall sentiment, allowing for the detection of sentiment-heavy opinions. An illustration of MILNET is shown in Figure 2(b); the model consists of three components: a CNN segment encoder, a softmax segment classifier and an attention-based prediction weighting module.

Segment Encoding   An encoding $v_i = \mathrm{CNN}(X_i)$ is produced for each segment, using the CNN architecture described in Section 2.

Segment Classification   Obtaining a separate representation v_i for every segment in a document allows us to produce individual segment sentiment predictions $p_i = \langle p_i^{(1)}, \ldots, p_i^{(C)} \rangle$. This is achieved using a softmax classifier:

$$p_i = \mathrm{softmax}(W_c v_i + b_c), \qquad (5)$$

where W_c and b_c are the classifier's parameters, shared across all segments. Individual distributions p_i are shown in Figure 2(b) as small bar-charts.

Document Classification   In the simplest case, document-level predictions can be produced by taking the average of segment class distributions: $p_d^{(c)} = 1/m \sum_i p_i^{(c)}$, $c \in [1, C]$. This is, however, a crude way of combining segment sentiment, as not all parts of a document convey important sentiment clues. We opt for a segment attention mechanism which rewards text units that are more likely to be good sentiment predictors.

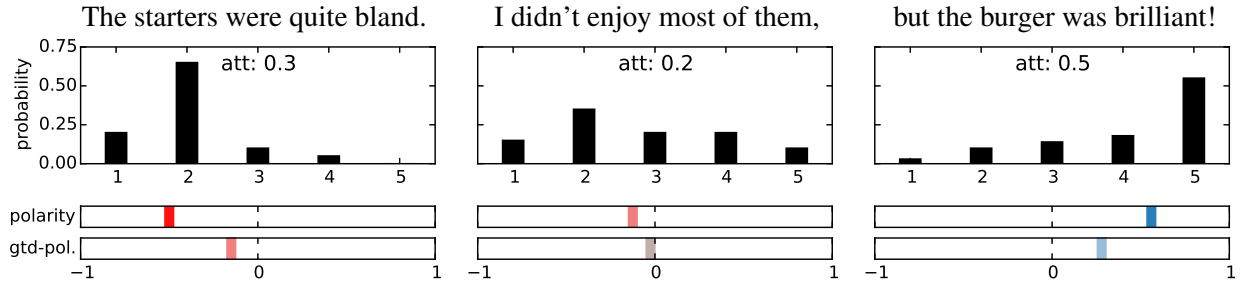Our attention mechanism is based on a bidirectional GRU component (Bahdanau et al., 2015) and

21

Figure 3: Polarity scores (bottom) obtained from class probability distributions for three EDUs (top) extracted from a restaurant review. Attention weights (top) are used to fine-tune the obtained polarities.

inspired by Yang et al. (2016). However, in contrast to their work, where attention is used to combine sentence representations into a single document vector, we utilize a similar technique to aggregate individual sentiment *predictions*.

We first use separate GRU modules to produce forward and backward hidden vectors, which are then concatenated:

$$\overrightarrow{\mathrm{h}}_i = \overrightarrow{\mathrm{GRU}}(\mathrm{v}_i), \tag{6}$$

$$\overleftarrow{\mathrm{h}}_i = \overleftarrow{\mathrm{GRU}}(\mathrm{v}_i), \tag{7}$$

$$\mathrm{h}_i = [\overrightarrow{\mathrm{h}}_i, \overleftarrow{\mathrm{h}}_i], \ i \in [1, m]. \tag{8}$$

The importance of each segment is measured with the aid of a vector h_a, as follows:

$$\mathrm{h}'_i = \tanh(W_a \mathrm{h}_i + \mathrm{b}_a), \tag{9}$$

$$a_i = \frac{exp(\mathrm{h}'^{\mathsf{T}}_i \mathrm{h}_a)}{\sum_i exp(\mathrm{h}'^{\mathsf{T}}_i \mathrm{h}_a)}, \tag{10}$$

where Equation (9) defines a one-layer MLP that produces an attention vector for the i-th segment. Attention weights a_i are computed as the normalized similarity of each h'_i with h_a. Vector h_a, which is randomly initialized and learned during training, can be thought of as a trained *key*, able to recognize sentiment-heavy segments. The attention mechanism is depicted in the dashed box of Figure 2, with attention weights shown as shaded circles.

Finally, we obtain a document-level distribution over sentiment labels as the weighted sum of segment distributions (see top of Figure 2(b)):

$$p_d^{(c)} = \sum_i a_i p_i^{(c)}, \ c \in [1, C]. \tag{11}$$

Training The model is trained end-to-end on documents with user-generated sentiment labels. We use the negative log likelihood of the document-level prediction as an objective function:

$$L = -\sum_d \log p_d^{(y_d)} \tag{12}$$

4 Polarity-based Opinion Extraction

After training, our model can produce segment-level sentiment predictions for unseen texts in the form of class probability distributions. A direct application of our method is opinion extraction, where highly positive and negative snippets are selected from the original document, producing extractive sentiment summaries, as described below.

Polarity Scoring In order to extract opinion summaries, we need to rank segments according to their sentiment polarity. We introduce a method that takes our model's confidence in the prediction into account, by reducing each segment's class probability distribution p_i to a single real-valued polarity score. To achieve this, we first define a real-valued *class weight* vector $\mathrm{w} = \langle w^{(1)}, \dots, w^{(C)} \,|\, w^{(c)} \in [-1, 1] \rangle$ that assigns uniformly-spaced weights to the ordered labelset, such that $w^{(c+1)} - w^{(c)} = \frac{2}{C-1}$. For example, in a 5-class scenario, the class weight vector would be $\mathrm{w} = \langle -1, -0.5, 0, 0.5, 1 \rangle$. We compute the polarity score of a segment as the dot-product of the probability distribution p_i with vector w:

$$\mathrm{polarity}(s_i) = \sum_c p_i^{(c)} w^{(c)} \ \in [-1, 1] \tag{13}$$

Gated Polarity As a way of increasing the effectiveness of our method, we introduce a *gated* extension that uses the attention mechanism of our model to further differentiate between segments that carry

22

significant sentiment cues and those that do not:

$$\text{gated-polarity}(s_i) = a_i \cdot \text{polarity}(s_i), \quad (14)$$

where a_i is the attention weight assigned to the i-th segment. This forces the polarity scores of segments the model does not attend to closer to 0.

An illustration of our polarity scoring function is provided in Figure 3, where the class predictions (top) of three restaurant review segments are mapped to their corresponding polarity scores (bottom). We observe that our method produces the desired result; segments 1 and 2 convey negative sentiment and receive negative scores, whereas the third segment is mapped to a positive score. Although the same discrete class label is assigned to the first two, the second segment's score is closer to 0 (neutral) as its class probability mass is more evenly distributed.

Segmentation Policies As mentioned earlier, one of the hypotheses investigated in this work regards the use of subsentential units as the basis of extraction. Specifically, our model was applied to sentences and *Elementary Discourse Units* (EDUs), obtained from a Rhetorical Structure Theory (RST) parser (Feng and Hirst, 2012). According to RST, documents are first segmented into EDUs corresponding roughly to independent clauses which are then recursively combined into larger discourse spans. This results in a tree representation of the document, where connected nodes are characterized by discourse relations. We only utilize RST's segmentation, and leave the potential use of the tree structure to future work.

The example in Figure 3 illustrates why EDU-based segmentation might be beneficial for opinion extraction. The second and third EDUs correspond to the sentence: *I didn't enjoy most of them, but the burger was brilliant.* Taken as a whole, the sentence conveys mixed sentiment, whereas the EDUs clearly convey opposing sentiment.

5   Experimental Setup

In this section we describe the data used to assess the performance of our model. We also give details on model training and comparison systems.

	Yelp'13	IMDB
Documents	335,018	348,415
Average #Sentences	8.90	14.02
Average #EDUs	19.11	37.38
Average #Words	152	325
Vocabulary Size	211,245	115,831
Classes	1–5	1–10

Table 1: Document-level sentiment classification datasets used to train our models.

	Yelp'13$_{seg}$		IMDB$_{seg}$	
	Sent.	EDUs	Sent.	EDUs
#Segments	1,065	2,110	1,029	2,398
#Documents	100		97	
Classes	$\{-, 0, +\}$		$\{-, 0, +\}$	

Table 2: SPOT dataset: numbers of documents and segments with polarity annotations.

5.1   Datasets

Our models were trained on two large-scale sentiment classification collections. The Yelp'13 corpus was introduced in Tang et al. (2015) and contains customer reviews of local businesses, each associated with human ratings on a scale from 1 (negative) to 5 (positive). The IMDB corpus of movie reviews was obtained from Diao et al. (2014); each review is associated with user ratings ranging from 1 to 10. Both datasets are split into training (80%), validation (10%) and test (10%) sets. A summary of statistics for each collection is provided in Table 1.

In order to evaluate model performance on the segment level, we constructed a new dataset named SPOT (as a shorthand for **S**egment **PO**lari**T**y) by annotating documents from the Yelp'13 and IMDB collections. Specifically, we sampled reviews from each collection such that all document-level classes are represented uniformly, and the document lengths are representative of the respective corpus. Documents were segmented into sentences and EDUs, resulting in two segment-level datasets per collection. Statistics are summarized in Table 2.

Each review was presented to three *Amazon Mechanical Turk* (AMT) annotators who were asked to judge the sentiment conveyed by each segment (i.e., sentence or EDU) as *negative*, *neutral*, or *pos-*
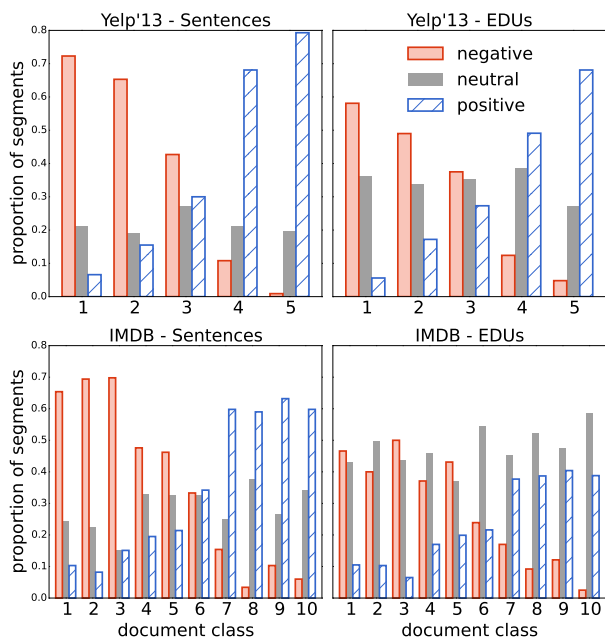
23

Figure 4: Distribution of segment-level labels per document-level class on our the SPOT datasets.

itive. We assigned labels using a majority vote or a fourth annotator in the rare cases of no agreement ($< 5\%$). Figure 4 shows the distribution of segment labels for each document-level class. As expected, documents with positive labels contain a larger number of positive segments compared to documents with negative labels and vice versa. Neutral segments are distributed in an approximately uniform manner across document classes. Interestingly, the proportion of neutral EDUs is significantly higher compared to neutral sentences. The observation reinforces our argument in favor of EDU segmentation, as it suggests that a sentence with positive or negative overall polarity may still contain neutral EDUs. Discarding neutral EDUs, could therefore lead to more concise opinion extraction compared to relying on entire sentences.

We further experimented on two collections introduced by Kotzias et al. (2015) which also originate from the YELP'13 and IMDB datasets. Each collection consists of 1,000 randomly sampled sentences annotated with binary sentiment labels.

5.2 Model Comparison

On the task of segment classification we compared MILNET, our multiple instance learning network, against the following methods:

Majority: Majority class applied to all instances.

SO-CAL: State-of-the-art lexicon-based system that classifies segments into positive, neutral, and negative classes (Taboada et al., 2011).

Seg-CNN: Fully-supervised CNN segment classifier trained on SPOT's labels (Kim, 2014).

GICF: The Group-Instance Cost Function model introduced in Kotzias et al. (2015). This is an unweighted average prediction aggregation MIL method that uses sentence features from a pretrained convolutional neural model.

HIERNET: HIERNET does not explicitly generate individual segment predictions. Segment polarity scores are obtained by assigning the document-level prediction to every segment. We can then produce finer-grained polarity distinctions via gating, using the model's attention weights.

We further illustrate the differences between HIERNET and MILNET in Figure 5, which includes short descriptions and simplified equations for each model. MILNET naturally produces distinct segment polarities, while HIERNET assigns a single polarity score to every segment. In both cases, gating is a further means of identifying neutral segments.

Finally, we differentiate between variants of HIERNET and MILNET according to:

Polarity source: Controls whether we assign polarities via *segment-specific* or *document-wide* predictions. HIERNET only allows for document-wide predictions. MILNET can use both.

Attention: We use models without gating (no subscript), with gating (*gt* subscript) as well as models trained with the attention mechanism disabled, falling back to simple averaging (*avg* subscript).

5.3 Model Training and Evaluation

We trained MILNET and HIERNET using Adadelta (Zeiler, 2012) for 25 epochs. Mini-batches of 200 documents were organized based on the reviews' segment and document lengths so the amount of padding was minimized. We used 300-dimensional pre-trained word2vec embeddings. We tuned hyperparameters on the validation sets of the document classification collections, resulting in the following configuration (unless otherwise noted). For the CNN segment encoder, we used window sizes of 3, 4
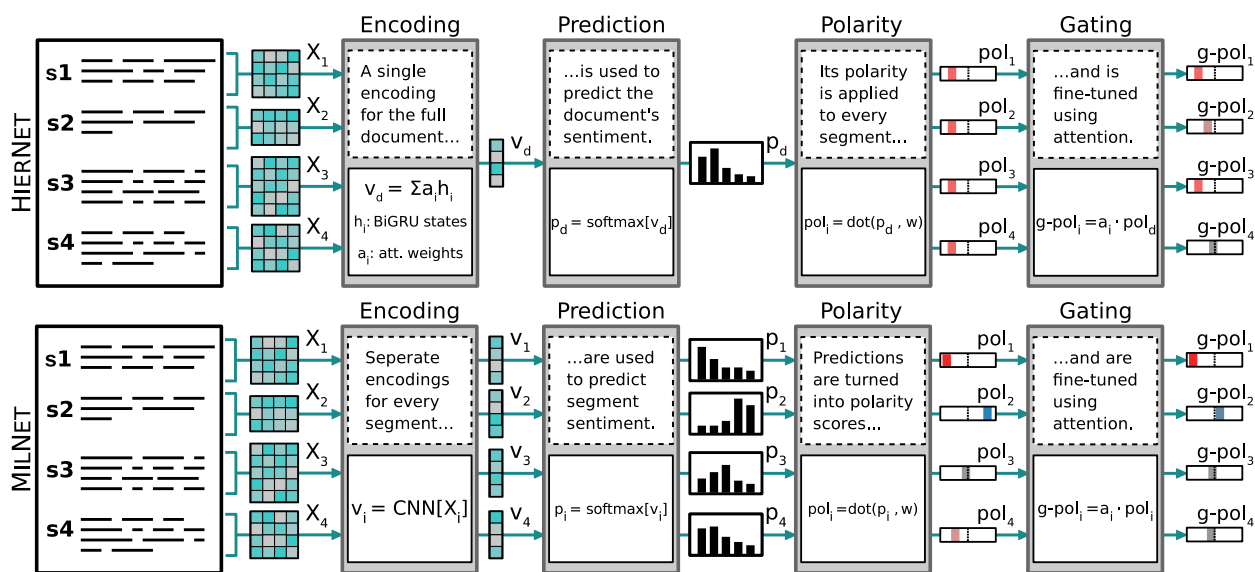
24

Figure 5: System pipelines for HIERNET and MILNET showing 4 distinct phases for sentiment analysis.

and 5 words with 100 feature maps per window size, resulting in 300-dimensional segment vectors. The GRU hidden vector dimensions for each direction were set to 50 and the attention vector dimensionality to 100. We used L2-normalization and dropout to regularize the softmax classifiers and additional dropout on the internal GRU connections.

Real-valued polarity scores produced by the two models are mapped to discrete labels using two appropriate thresholds $t_1, t_2 \in [-1, 1]$, so that a segment s is classified as negative if $\mathrm{polarity}(s) < t_1$, positive if $\mathrm{polarity}(s) > t_2$ or neutral otherwise.[3] To evaluate performance, we use macro-averaged F1 which is unaffected by class imbalance. We select optimal thresholds using 10-fold cross-validation and report mean scores across folds.

The fully-supervised convolutional segment classifier (Seg-CNN) uses the same window size and feature map configuration as our segment encoder. Seg-CNN was trained on SPOT using segment labels directly and 10-fold cross-validation (identical folds as in our main models). Seg-CNN is not directly comparable to MILNET (or HIERNET) due to differences in supervision type (segment vs. document labels) and training size (1K-2K segment labels vs. ∼250K document labels). However, the

---

[3]The discretization of polarities is only used for evaluation purposes and is not necessary for summary extraction, where we only need a relative ranking of segments.

comparison is indicative of the utility of fine-grained sentiment predictors that do not rely on expensive segment-level annotations.

6 Results

We evaluated models in two ways. We first assessed their ability to classify segment polarity in reviews using the newly created SPOT dataset and, additionally, the sentence corpora of Kotzias et al. (2015). Our second suite of experiments focused on opinion extraction: we conducted a judgment elicitation study to determine whether extracts produced by MILNET are useful and of higher quality compared to HIERNET and other baselines. We were also interested to find out whether EDUs provide a better basis for opinion extraction than sentences.

6.1 Segment Classification

Table 3 summarizes our results. The first block in the table reports the performance of the majority class baseline. The second block considers models that do not utilize segment-level predictions, namely HIERNET which assigns polarity scores to segments using its document-level predictions, as well as the variant of MILNET which similarly uses document-level predictions only (Equation (11)). In the third block, MILNET's segment-level predictions are used. Each block further differentiates between three levels of attention integration, as previ-

	Method	**Yelp'13**$_{seg}$		**IMDB**$_{seg}$	
		Sent	EDU	Sent	EDU
	Majority	19.02†	17.03†	18.32†	21.52†
Document	HIERNET$_{avg}$	54.21†	50.90†	46.99†	49.02†
	HIERNET	55.33†	51.43†	48.47†	49.70†
	HIERNET$_{gt}$	56.64†	58.75	62.12	57.38†
	MILNET$_{avg}$	58.43†	48.63†	53.40†	51.81†
	MILNET	52.73†	53.59†	48.75†	47.18†
	MILNET$_{gt}$	59.74†	59.47	61.83†	58.24†
Segm	MILNET$_{avg}$	51.79†	46.77†	45.69†	38.37†
	MILNET	61.41	59.58	59.99†	57.71†
	MILNET$_{gt}$	**63.35**	**59.85**	**63.97**	**59.87**
	SO-CAL	56.53†	58.16†	53.21†	60.40
	Seg-CNN	56.18†	59.96	58.32†	62.95†

Table 3: Segment classification results (in macro-averaged F1). † indicates that the system in question is significantly different from MILNET$_{gt}$ (approximate randomization test (Noreen, 1989), $p < 0.05$).

ously described. The final block shows the performance of SO-CAL and the Seg-CNN classifier.

When considering models that use document-level supervision, MILNET with gated, segment-specific polarities obtains the best classification performance across all four datasets. Interestingly, it performs comparably to Seg-CNN, the fully-supervised segment classifier, which provides additional evidence that MILNET can effectively identify segment polarity without the need for segment-level annotations. Our model also outperforms the strong SO-CAL baseline in all but one datasets which is remarkable given the expert knowledge and linguistic information used to develop the latter. Document-level polarity predictions result in lower classification performance across the board. Differences between the standard hierarchical and multiple instance networks are less pronounced in this case, as MILNET loses the advantage of producing segment-specific sentiment predictions. Models without attention perform worse in most cases. The use of gated polarities benefits all model configurations, indicating the method's ability to selectively focus on segments with significant sentiment cues.

We further analyzed the polarities assigned by MILNET and HIERNET to positive, negative, and

	Neutral Segments		
		Non-Gtd	Gated
Sent	HIERNET	4.67	36.60
	MILNET	39.61	44.60
		Non-Gtd	Gated
EDU	HIERNET	2.39	55.38
	MILNET	52.10	56.60

Table 4: F1 scores for neutral segments (Yelp'13).

Method	Yelp	IMDB
GICF	86.3	86.0
GICF$_{HN}$	92.9	86.5
GICF$_{MN}$	93.2	91.0
MILNET	94.0	91.9

Table 5: Accuracy scores on the sentence classification datasets introduced in Kotzias et al. (2015).
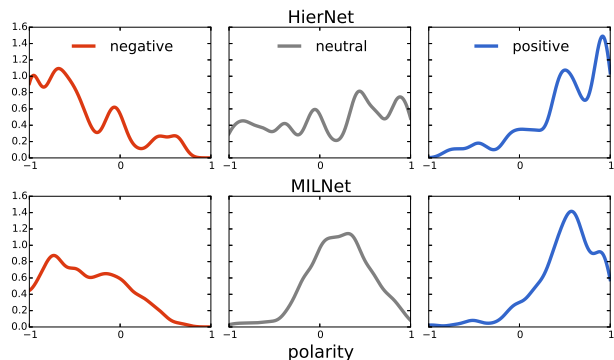

Figure 6: Distribution of predicted polarity scores across three classes (Yelp'13 sentences).

neutral segments. Figure 6 illustrates the distribution of polarity scores produced by the two models on the Yelp'13 dataset (sentence segmentation). In the case of negative and positive sentences, both models demonstrate appropriately skewed distributions. However, the neutral class appears to be particularly problematic for HIERNET, where polarity scores are scattered across a wide range of values. In contrast, MILNET is more successful at identifying neutral sentences, as its corresponding distribution has a single mode near zero. Attention gating addresses this issue by moving the polarity scores of sentiment-neutral segments towards zero. This is illustrated in Table 4 where we observe that gated variants of both models do a better job at identifying neutral segments. The effect is very significant for HIERNET, while MILNET benefits slightly and remains more effective overall. Similar trends were observed in all four SPOT datasets.

In order to examine the effect of training size, we trained multiple models using subsets of the original document collections. We trained on five random
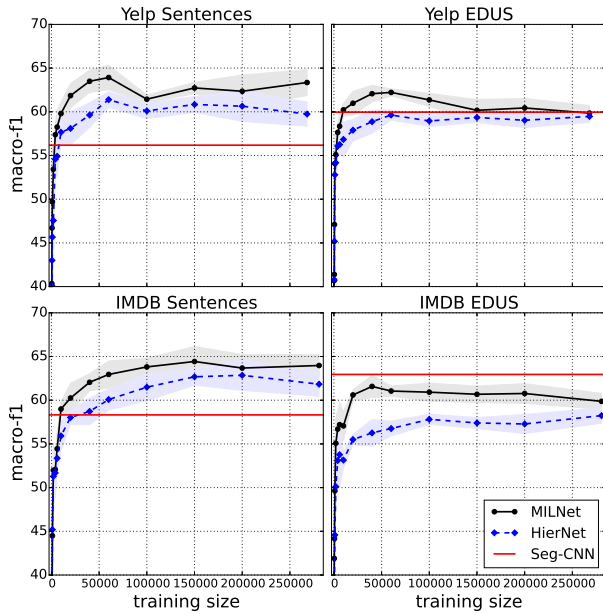
Figure 7: Performance of HIERNET$_{gt}$ and MILNET$_{gt}$ for varying training sizes.

Method	Informativeness	Polarity	Coherence
HIERNET^{sent}	43.7	33.6	43.5
MILNET^{sent}	**45.7**	**36.7**	**44.6**
Unsure	10.7	29.6	11.8
HIERNET^{edu}	34.2†	28.0†	**48.4**
MILNET^{edu}	**53.3**	**61.1**	45.0
Unsure	12.5	11.0	6.6
MILNET^{sent}	35.7†	33.4†	**70.4**†
MILNET^{edu}	**55.0**	**51.5**	23.7
Unsure	9.3	15.2	5.9
LEAD	34.0	19.0†	**40.3**
RANDOM	22.9†	19.6†	17.8†
MILNET^{edu}	**37.4**	**46.9**	33.3
Unsure	5.7	14.6	8.6

Table 6: Human evaluation results (in percentages). † indicates that the system in question is significantly different from MILNET (sign-test, $p < 0.01$).

subsets for each training size, ranging from 100 documents to the full training set, and tested segment classification performance on SPOT. The results, averaged across trials, are presented in Figure 7. With the exception of the IMDB EDU-segmented dataset, MILNET only requires a few thousand training documents to outperform the supervised Seg-CNN. HIERNET follows a similar curve, but is inferior to MILNET. A reason for MILNET's inferior performance on the IMDB corpus (EDU-split) can be low-quality EDUs, due to the noisy and informal style of language used in IMDB reviews.

Finally, we compared MILNET against the GICF model (Kotzias et al., 2015) on their Yelp and IMDB sentence sentiment datasets.[4] Their model requires sentence embeddings from a pre-trained neural model. We used the hierarchical CNN from their work (Denil et al., 2014) and, additionally, pre-trained HIERNET and MILNET sentence embeddings. The results in Table 5 show that MILNET outperforms all variants of GIFC. Our models also seem to learn better sentence embeddings, as they improve GICF's performance on both collections.

---

[4]GICF only handles binary labels, which makes it unsuitable for the full-scale comparisons in Table 3. Here, we binarize our training datasets and use same-sized sentence embeddings for all four models ($\mathbb{R}^{150}$ for Yelp, $\mathbb{R}^{72}$ for IMDB).

6.2 Opinion Extraction

In our opinion extraction experiments, AMT workers (all native English speakers) were shown an original review and a set of extractive, bullet-style summaries, produced by competing systems using a 30% compression rate. Participants were asked to decide which summary was best according to three criteria: *Informativeness* (Which summary best captures the salient points of the review?), *Polarity* (Which summary best highlights positive and negative comments?) and *Coherence* (Which summary is more coherent and easier to read?). Subjects were allowed to answer "Unsure" in cases where they could not discriminate between summaries. We used all reviews from our SPOT dataset and collected three responses per document. We ran four judgment elicitation studies: one comparing HIERNET and MILNET when summarizing reviews segmented as sentences, a second one comparing the two models with EDU segmentation, a third which compares EDU- and sentence-based summaries produced by MILNET, and a fourth where EDU-based summaries from MILNET were compared to a LEAD (the first N words from each document) and a RANDOM (random EDUs) baseline.

Table 6 summarizes our results, showing the proportion of participants that preferred each system. The first block in the table shows a slight prefer-

> [Rating: ★★★★] As with any family-run hole in the wall, service can be slow. What the staff lacked in speed, they made up for in charm. The food was good, but nothing wowed me. I had the Pierogis while my friend had swedish meatballs. Both dishes were tasty, as were the sides. One thing that was disappointing was that the food was a a little cold (lukewarm). The restaurant itself is bright and clean. I will go back again when i feel like eating outside the box.

	Extracted via HIERNET$_{gt}$	Extracted via MILNET$_{gt}$
EDU-based	(0.13) [+0.26] The food was good$^+$ (0.10) [+0.26] but nothing wowed me.$^+$ (0.09) [+0.26] The restaurant itself is bright and clean$^+$ (0.13) [+0.26] Both dishes were tasty$^+$ (0.18) [+0.26] I will go back again$^+$	(0.16) [+0.12] The food was good$^+$ (0.12) [+0.43] The restaurant itself is bright and clean$^+$ (0.19) [+0.15] I will go back again$^+$ (0.09) [–0.07] but nothing wowed me.$^-$ (0.10) [–0.10] the food was a a little cold (lukewarm)$^-$
Sent-based	(0.12) [+0.23] Both dishes were tasty, as were the sides$^+$ (0.18) [+0.23] The food was good, but nothing wowed me$^+$ (0.22) [+0.23] One thing that was disappointing was that the food was a a little cold (lukewarm)$^+$	(0.13) [+0.26] Both dishes were tasty, as were the sides$^+$ (0.20) [+0.59] I will go back again when I feel like eating outside the box$^+$ (0.18) [–0.12] The food was good, but nothing wowed me$^-$
	(number): attention weight [number]: non-gated polarity score text$^+$: extracted positive opinion text$^-$: extracted negative opinion	

Figure 8: Example EDU- and sentence-based opinion summaries produced by HIERNET$_{gt}$ and MILNET$_{gt}$.

ence for MILNET across criteria. The second block shows significant preference for MILNET against HIERNET on informativeness and polarity, whereas HIERNET was more often preferred in terms of coherence, although the difference is not statistically significant. The third block compares sentence and EDU summaries produced by MILNET. EDU summaries were perceived as significantly better in terms of informativeness and polarity, but not coherence. This is somewhat expected as EDUs tend to produce more terse and telegraphic text and may seem unnatural due to segmentation errors. In the fourth block we observe that participants find MIL-NET more informative and better at distilling polarity compared to the LEAD and RANDOM (EDUs) baselines. We should point out that the LEAD system is not a strawman; it has proved hard to outperform by more sophisticated methods (Nenkova, 2005), particularly on the newswire domain.

Example EDU- and sentence-based summaries produced by gated variants of HIERNET and MIL-NET are shown in Figure 8, with attention weights and polarity scores of the extracted segments shown in round and square brackets respectively. For both granularities, HIERNET's positive document-level prediction results in a single polarity score assigned to every segment, and further adjusted using the corresponding attention weights. The extracted segments are informative, but fail to capture the negative sentiment of some segments. In contrast, MIL-

NET is able to detect positive and negative snippets via individual segment polarities. Here, EDU segmentation produced a more concise summary with a clearer grouping of positive and negative snippets.

7 Conclusions

In this work, we presented a neural network model for fine-grained sentiment analysis within the framework of multiple instance learning. Our model can be trained on large scale sentiment classification datasets, without the need for segment-level labels. As a departure from the commonly used vector-based composition, our model first predicts sentiment at the sentence- or EDU-level and subsequently combines predictions up the document hierarchy. An attention-weighted polarity scoring technique provides a natural way to extract sentiment-heavy opinions. Experimental results demonstrate the superior performance of our model against more conventional neural architectures. Human evaluation studies also show that MILNET opinion extracts are preferred by participants and are effective at capturing informativeness and polarity, especially when using EDU segments. In the future, we would like to focus on multi-document, aspect-based extraction (Cao et al., 2017) and ways of improving the coherence of our summaries by taking into account more fine-grained discourse information (Daumé III and Marcu, 2002).

Acknowledgments

References

Stuart Andrews and Thomas Hofmann. 2004. Multiple instance learning via disjunctive programming boosting. In *Advances in Neural Information Processing Systems 16*, pages 65–72. Curran Associates, Inc.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 5th Conference on International Language Resources and Evaluation*, volume 10, pages 2200–2204, Valletta, Malta.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, California, USA.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2017. Improving multi-document summarization via text classification. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3053–3058, San Francisco, California, USA.

Peter Carbonetto, Gyuri Dorkó, Cordelia Schmid, Hendrik Kück, and Nando De Freitas. 2008. Learning to recognize objects with little supervision. *International Journal of Computer Vision*, 77(1):219–237.

Giuseppe Carenini, Rymond Ng, and Adam Pauls. 2006. Multidocument summarization of evaluative text. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 305–312, Trento, Italy.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and New Directions in Discourse and Dialogue*, pages 85–112. Springer.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany.

Timothee Cour, Ben Sapp, and Ben Taskar. 2011. Learning from partial labels. *Journal of Machine Learning Research*, 12(May):1501–1536.

Hal Daumé III and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 449–456, Philadelphia, Pennsylvania, USA.

Misha Denil, Alban Demiraj, and Nando de Freitas. 2014. Extraction of salient sentences from labelled documents. Technical report, University of Oxford.

Giuseppe Di Fabbrizio, Amanda Stent, and Robert Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 54–63, Philadelphia, Pennsylvania, USA.

Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 193–202, New York, NY, USA.

Thomas G. Dietterich, Richard H. Lathrop, and Toms Lozano-Prez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31 – 71.

Wei Vanessa Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 60–68, Jeju Island, Korea.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348, Beijing, China.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1613, Doha, Qatar.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 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*, pages 541–550, Portland, Oregon, USA.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, Washington, USA.

Rie Johnson and Tong Zhang. 2015a. Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–112, Denver, Colorado, USA.

Rie Johnson and Tong Zhang. 2015b. Semi-supervised convolutional neural networks for text categorization via region embedding. In *Advances in Neural Information Processing Systems 28*, pages 919–927. Curran Associates, Inc.

Jim Keeler and David E. Rumelhart. 1992. A self-organizing integrated segmentation and recognition neural net. In *Advances in Neural Information Processing Systems 4*, pages 496–503. Morgan-Kaufmann.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar.

Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 597–606, Sydney, Australia.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196, Beijing, China.

Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. 2009. Sentiment summarization: Evaluating and learning user preferences. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 514–522, Athens, Greece.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles, California, USA.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Oded Maron and Aparna Lakshmi Ratan. 1998. Multiple-instance learning for natural scene classification. In *Proceedings of the 15th International Conference on Machine Learning*, volume 98, pages 341–349, San Francisco, California, USA.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3075–3081, San Francisco, California.

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings of the 20th AAAI*, pages 1436–1441, Pittsburgh, Pennsylvania, USA.

Eric Noreen. 1989. *Computer-intensive Methods for Testing Hypotheses: An Introduction*. Wiley.

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*, pages 115–124. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86, Pittsburgh, Pennsylvania, USA.

Nikolaos Pappas and Andrei Popescu-Belis. 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*, pages 455–466, Doha, Qatar, October.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Explicit document modeling through weighted multiple-instance learning. *Journal of Artificial Intelligence Research*, 58:591–626.

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*, pages 913–921, Beijing, China.

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 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161, Edinburgh, Scotland, UK.

Richard Socher, Alex Perelygin, Jean 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 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Oscar Täckström and Ryan McDonald. 2011. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceedings of the 39th European Conference on Information Retrieval*, pages 368–374, Aberdeen, Scotland, UK.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, Lisbon, Portugal.

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*, pages 417–424, Pittsburgh, Pennsylvania, USA.

Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94, Jeju Island, Korea.

Xiu-Shen Wei, Jianxin Wu, and Zhi-Hua Zhou. 2014. Scalable multi-instance learning. In *Proceedings of the IEEE International Conference on Data Mining*, pages 1037–1042, Shenzhen, China.

Nils Weidmann, Eibe Frank, and Bernhard Pfahringer. 2003. A two-level learning method for generalized multi-instance problems. In *Proceedings of the 14th European Conference on Machine Learning*, pages 468–479, Dubrovnik, Croatia.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. 2015. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3469, Boston, Massachusetts, USA.

Rui Xia and Chengqing Zong. 2010. Exploring the use of word relation features for sentiment classification. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1336–1344, Beijing, China.

Xin Xu and Eibe Frank. 2004. Logistic regression and boosting for labeled bags of instances. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 272–281. Springer-Verlag.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics*, 2:435–448.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, USA.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Qi Zhang, Sally A. Goldman, Wei Yu, and Jason E. Fritts. 2002. Content-based image retrieval using multiple-instance learning. In *Proceedings of the 19th International Conference on Machine Learning*, volume 2, pages 682–689, Sydney, Australia.

Cha Zhang, John C. Platt, and Paul A. Viola. 2006. Multiple instance boosting for object detection. In *Advances in Neural Information Processing Systems 18*, pages 1417–1424. MIT Press.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. 2009. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1249–1256, Montréal, Quebec.

32