An Unsupervised Approach for Aspect-Based Sentiment Classification Using Attentional Neural Models

Luca Zampierin
Erasmus University Rotterdam
Rotterdam, Netherlands
luca.zampierin@hotmail.com

Flavius Frasincar
Erasmus University Rotterdam
Rotterdam, Netherlands
frasincar@ese.eur.nl

ABSTRACT

With the vast amount of reviews available on the Web, in the past decades, a growing share of literature has focused on sentiment analysis. Aspect-based sentiment classification is the subtask that seeks to detect the sentiment expressed by the content creators towards a defined target typically within a sentence. This paper introduces two novel unsupervised attentional neural network models for aspect-based sentiment classification, and tests them on English restaurant reviews. The first model employs an autoencoder-like structure to learn a sentiment embedding matrix where each row of the matrix represents the embedding for one sentiment. To improve the model, a target-based attention mechanism is included that de-emphasizes irrelevant words. Last, a redundancy and a seed regularization term constrain the sentiment embedding matrix. The second model extends the first by including a Bi-LSTM layer in the attention mechanism to exploit contextual information. Although both models construct meaningful sentiment embeddings, experimental results indicate that the inclusion of the Bi-LSTM in the attention mechanism leads to a more precise attention mechanisms and, thus, better predictions. The best model, i.e., the second, outperforms all investigated unsupervised and weakly supervised algorithms for aspect-based sentiment classification from the literature.

CCS CONCEPTS

• **Information systems** → **Sentiment analysis**; *Information extraction*; Web mining.

KEYWORDS

Attentional neural model; unsupervised learning; aspect-based sentiment classification

ACM Reference Format:
Luca Zampierin and Flavius Frasincar. 2025. An Unsupervised Approach for Aspect-Based Sentiment Classification Using Attentional Neural Models. In *The 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25), March 31-April 4, 2025, Catania, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3672608.3707884

SAC '25, March 31-April 4, 2025, Catania, Italy

ACM ISBN 979-8-4007-0629-5/25/03…$15.00
https://doi.org/10.1145/3672608.3707884

1 INTRODUCTION

Word-of-Mouth (WOM) and its impact on consumers' decision-making process, have long been important topics in academic literature. WOM has been shown to strongly affect individuals' choices and product-satisfaction [12], possibly even more than traditional advertising strategies [8]. With the advent of the Web, consumer-to-consumer interactions are not constrained by geographical or temporal barriers. This phenomenon is referred to as Electronic WOM. Consumers engage with Web reviews to gather product information that can facilitate the decision-making process. Besides being valuable for customers, online reviews are useful for sellers too. In fact, they provide information about customer satisfaction with respect to existing products [21], and they allow for a more targeted market segmentation and product development [32].

Sentiment analysis is the subfield of Natural Language Processing (NLP) concerned with automatically detecting the sentiment expressed by content creators. In general, however, Web reviews can discuss different aspects of a certain product and can thus present opposite sentiment polarities. For instance, in the fictitious restaurant review "*The food was great, but the service was horrible*", the reviewer was satisfied with the "food" but he/she was clearly unhappy with the "service". Aspect-Based Sentiment Analysis (ABSA) is a subtask of sentiment analysis that aims to identify the aspects of a product or service that are discussed within a review and compute the sentiment score specifically for each aspect. More precisely, [23] defines three subtasks of ABSA: Opinion Target Extraction (OTE), Aspect Category Detection (ACD), and Aspect-Based Sentiment Classification (ABSC). OTE aims to identify aspect terms, ACD categorizes the aspects into predefined aspect categories, and ABSC detects the sentiments expressed towards specific targets. In this research, the main focus is the ABSC task.

The models for ABSC can be grouped into three sets: dictionary-based, supervised machine learning, and unsupervised machine learning [27]. Dictionary-based approaches utilize pre-developed dictionaries, where words are assigned sentiment scores, to compute an aggregate polarity score. Supervised machine learning approaches extract meaningful patterns using labeled example data. While these models have shown promising performances, especially when combined with dictionary-based approaches in the creation of hybrid approaches [28] [31] [34], they require large labeled training datasets. Since the process of annotating data is time-consuming and costly, weakly supervised and unsupervised approaches have recently gained traction.

Lately, Deep Neural Networks (DNN) have shown the most promising results for ABSA [6]. If the use of DNN in a supervised manner has been thoroughly studied within ABSA, and more specifically for ABSC, the unsupervised application of these models is

still relatively unexplored. [11] introduces an unsupervised neural attention model called Attention-Based Aspect Extraction (ABAE) which is shown to outperform baseline unsupervised methods in the aspect extraction task.

In this work, inspired by the seminal work from [11], we propose two unsupervised methods for ABSC. The first one, called Attention-Based Aspect-Based Sentiment Extraction 1 (ABABSE1), adapts ABAE to the ABSC task. For the second, called Attention-Based Aspect-Based Sentiment Extraction 2 (ABABSE2), we propose an unprecedented usage of a Bi-directional Long Short-Term Memory (Bi-LSTM) layer within the attention mechanism of an unsupervised attentional neural network. We evaluate the performance of these models using the SemEval-2015 task 12 [24] and the SemEval-2016 task 5 subtask 1 [23] datasets. The Python scripts can be found at https://github.com/LucaZampierin/ABABSE.
The contribution of this research can be summarized as follows:

- We propose ABABSE1: an unsupervised attention-based neural networks that adapts ABAE [11] to the ABSC task. The model employs an autoencoder-like structure to learn a sentiment embedding matrix and uses an attention mechanism to emphasize relevant words. We extend ABAE by including a seed regularization term to exploit common knowledge [37] and we include the aspect category information by means of a one-hot-encoded vector.
- We propose ABABSE2: an extension of ABABSE1 where we advance an unprecedented unsupervised usage of a Bi-LSTM within the attention mechanisms in order to capture contextual information and better detect relevant words.
- We analyze the performance of the two proposed models on restaurant reviews. Experimental results show that ABABSE2 outperforms all investigated unsupervised and weakly supervised methods achieving state-of-the-art results.

The remainder of the paper is structured as follows. Section 2 provides a review of the related work. The datasets used for the empirical analysis are described in Section 3. Next, Section 4 presents the algorithms proposed in this research. In Section 5, the performances of the novel models are compared against other unsupervised approaches as well as state-of-the-art algorithms for ABSC from the literature. Last, in Section 6 we present some concluding remarks and suggestions for future research.

2 RELATED WORK

The focus of sentiment analysis is identifying the sentiment that is expressed in written contents. In particular, this entails detecting the sentiment (s), the target (g), the individual expressing the sentiment (h), and the moment when the sentiment was revealed (t) [17]. The quadruple (s,g,h,t) forms an opinion, that is, the part of a sentence that expresses or implies a sentiment. In the opinion, both target and sentiment can be explicitly stated (e.g., "the pasta is delicious"), or implicitly stated (e.g., "I was blown away!"). Since implicit targets in reviews are relatively infrequent [7] and require particular attention, in this paper, we assume the presence of an explicit target. If the target is a specific aspect of an entity, then the task is defined as ABSA. [27] presents a detailed outline of the models that have been proposed for the three tasks of ABSA, and here we outline the most relevant ones for ABSC.

The first category are the dictionary-based approaches. These algorithms utilize lexicons to compute the sentiment score of the words in the sentence and then combine these scores to develop an aggregate sentiment score for the aspect studied. WordNet [20] and SenticNet [5], in particular the latest SenticNet 6 [4], are domain-independent sentiment lexicons often employed for aspect-based sentiment analysis. In order to exploit common domain knowledge, domain-specific ontologies are common in the literature [29].

In the past decades, a large share of literature has focused on the application of machine learning models in ABSC, especially supervised approaches. Machine learning has proven to provide more flexibility and has achieved very promising performances in sentiment analysis tasks. Traditionally, the use of supervised classifiers was combined with sentiment lexicons that were used to engineer relevant features. For example, [15] creates a restaurant-specific sentiment lexicon that is then employed to obtain informative features for training a linear Support Vector Machine (SVM) model. Similarly, [33] employs a linear SVM classifier trained utilizing features obtained from sentiment lexicons. Here, however, the authors use publicly available lexicons which are then combined and adjusted using domain-specific knowledge.

Supervised classifiers suffer from the drawback that they require labeled data to be trained. On the other hand, weakly supervised machine learning approaches only require language-related tools, such as seed words, to pre-process the input and do not require predetermined labels [10]. Many weakly supervised models present in the literature build upon the Latent Dirichlet Allocation (LDA) model [3]. This is done by utilizing some a priori knowledge, often in the form of a sentiment lexicon, to bias the LDA model. For example, [16] proposes the weakly supervised Joint Sentiment-Topic (JST) model that aims to identify the topic and the sentiment expressed simultaneously. The authors modify the original LDA model by including an additional sentiment layer and by using a domain-independent lexicon for supervision. Similarly, the W2VLDA model advanced by [10] builds upon a topic modeling approach and performs concurrently aspect-term and opinion-words detection as well as ABSC. The first task is performed by training a Maximum Entropy classifier based on example aspect/sentiment words modeled using Brown clusters. The second task, instead, is performed by biasing the hyperparameters of the generative Dirichlet distribution that characterizes the LDA model by means of seed words. This bias is determined by the semantic similarity between domain and seed words, computed using the word2vec embeddings [19].

Unsupervised machine learning approaches have not been explored as much in detail as supervised approaches, especially for the sentiment classification task. However, the fact that they do not need any labels makes them very interesting models to study. [25] proposes an unsupervised information extraction algorithm, called OPINE. Besides extracting important aspects discussed in a review, this algorithm employs relaxation labeling for determining the sentiment polarity of words. Similarly, [9] introduces a simple unsupervised approach, called V3, for ABSC. There the authors exploit word2vec embeddings [19] in order to construct an in-domain lexicon. Given a seed word for either positive or negative polarity, the other words are assigned the polarity of the seed word they are closer to in the embedding space. Each sentence is then assigned the

most frequent polarity. Note that, in this case, two aspects reviewed in the same sentence are always assigned to the same polarity class.

Lately, a growing attention has been dedicated to the deep learning realm. DNNs are particularly attractive because of their ability to extract information from sentences without needing feature engineering [36]. Given the sequential property of language, Recurrent Neural Networks (RNN) have gained importance in the sentiment analysis field as they are able to capture contextual relations. In fact, Long-Short Term Memory (LSTM) [13] and their variants are often used in a supervised manner in ABSA [6]. Important advancements were achieved with the employment of neural attention models. These models utilize attention mechanisms that nudge the models into focusing on the most relevant words [18] [35] [36]. Last, the state-of-the-art performances have been achieved by combining the flexibility of deep learning models with the domain-specific knowledge of lexicons in the creation of hybrid models [31] [34].

Recently, researchers have explored the possibility of employing neural networks also in an unsupervised manner. [11] proposes an unsupervised model for the aspect extraction task, called Attention-Based Aspect Extraction (ABAE). First, the word2vec word embeddings are used to map words onto an embedding space. Then, a representation of the sentence is created by using a weighted average of the word embeddings, where the weights are computed by an attention layer that filters out irrelevant words. Last, an autoencoder-like structure conveys the information carried by the sentence representation through a lower-dimensional space and reconstructs it using an aspect embedding matrix. To enhance the training procedure, the authors use negative sampling and a redundancy regularization term. The model is proven capable of extracting coherent aspects and outperforms LDA-based and RNN-based models.

Given the promising performance of ABAE, other researchers have tried to apply the same reasoning to the ABSC task. The Joint Aspect Sentiment Autoencoder (JASA) model [37] adapts the structure of ABAE to simultaneously extract aspects and classify the sentiment as either positive or negative with minimal user intervention. The final goal is to construct an aspect and a sentiment embedding matrices at the same time. To ensure that meaningful dictionaries are constructed, the authors propose the usage of a seed regularization term, in addition to the redundancy one advanced by [11], which leverages user's information in the form of seed words.

3 DATA

To evaluate the performance of the proposed models, we employ two English datasets often used in the ABSA literature, namely SemEval-2015 task 12 [24] and SemEval-2016 task 5 subtask 1 [23][1]. This choice allows us to extract the performance of the baseline approaches directly from the original publications. Moreover, since these datasets comprise actual reviews, they include grammatical mistakes, spelling errors, and slang words, allowing us to test the real-life applicability of the models studied.

3.1 SemEval-2015 Task 12

The SemEval-2015 task 12 dataset [24] comprises reviews for three different domains: laptops, restaurants, and hotels. Following [31], in this research, only the restaurant reviews are utilized. The dataset consists of 254 training reviews and 96 test reviews. Each sentence is stored independently and the final dataset results in 1315 training sentences and 685 test sentences. Figure 1 presents a sample sentence from the training set. Each sentence can express a sentiment towards one or more aspects, defined as `target`. Moreover, each target is categorized within a set of predefined aspect categories, defined as `category`. Last, the `polarity` of the sentiment expressed towards the given aspect can be either positive, neutral, or negative.

[1]The XML files can be downloaded from http://metashare.ilsp.gr:8080/

Table 1: Polarity frequencies in the SemEval-2015 task 12 and SemEval-2016 task 5 subtask 1 datasets

	Negative	Neutral	Positive
SemEval-2015 train	21.9%	2.8%	75.3%
SemEval-2015 test	34.7%	6.2%	59.1%
SemEval-2016 train	26.0%	3.8%	70.2%
SemEval-2016 test	20.9%	4.9%	74.2%

```
<sentence id="1153034:0">
    <text>Great pizza and fantastic service.</text>
    <Opinions>
        <Opinion target="pizza" category="FOOD#QUALITY"
                 polarity="positive" from="6" to="11"/>
        <Opinion target="service" category="SERVICE#GENERAL"
                 polarity="positive" from="26" to="33"/>
    </Opinions>
```

Figure 1: Sample sentence from the training set of SemEval-2015 task 12.

Tables 1 and 2 present the polarity and aspect category distributions, respectively, for the pre-processed training and test sets. Regarding the polarity, the class distribution differs between training and test sets, which could represent a challenge for machine learning models. However, we expect unsupervised models to be less affected by this property than supervised models as the polarity labels are used only by the latter. Conversely, the training and test aspect category distributions have a strong resemblance, with the majority of the data points belonging to the `FOOD#QUALITY`, `RESTAURANT#GENERAL`, or `SERVICE#GENERAL` categories.

3.2 SemEval-2016 Task 5 Subtask 1

The SemEval-2016 task 5 subtask 1 dataset [23] comprises reviews for seven domains available in eight languages. As before, in this research, only the English reviews for the restaurant domain are used. The dataset consists of 350 reviews for training and 90 for testing. In total, the number of training and testing sentences corresponds to 2000 and 676, respectively, with the same structure as for the SemEval-2015 dataset (see Figure 1 for an example).

The polarity and aspect category distributions can be found in Tables 1 and 2, respectively, where the statistics for the pre-processed train and test sets are presented. As opposed to SemEval-2015, here both train and test sets present a strong unbalancedness between the polarity classes with more than 70% of the observations belonging to the positive class for both collections. With regards to the aspect category, more than 65% of the observations belong to the top three categories: `FOOD#QUALITY`, `SERVICE#GENERAL`, and `AMBIENCE#GENERAL`. For both polarity and aspect category, the distributions are similar between the train and test set.

Table 2: Absolute and percentage frequency of aspect category in the SemEval-2015 task 12 and SemEval-2016 task 5 subtask 1 datasets

Categories	SemEval-2015		SemEval-2016	
	Train	Test	Train	Test
AMBIENCE#GENERAL	12.82%	11.39%	12.13%	9.08%
DRINKS#PRICES	1.17%	0.84%	1.06%	0.62%
DRINKS#QUALITY	2.50%	1.84%	2.34%	3.38%
DRINKS#STYLE_OPTIONS	2.03%	1.00%	1.70%	1.69%
FOOD#PRICES	3.21%	4.36%	3.78%	3.38%
FOOD#QUALITY	41.05%	40.54%	40.69%	43.54%
FOOD#STYLE_OPTIONS	6.33%	5.53%	6.17%	7.85%
LOCATION#GENERAL	1.09%	1.34%	1.17%	1.54%
RESTAURANT#GENERAL	9.70%	9.88%	9.73%	8.92%
RESTAURANT#MISCELLANEOUS	2.35%	3.18%	2.61%	2.77%
RESTAURANT#PRICES	0.78%	2.68%	1.38%	0.77%
SERVICE#GENERAL	16.97%	17.42%	17.24%	16.46%

3.3 Pre-processing

Since we assume explicit targets, all the sentences for which the `target` is `NULL`, i.e., implicit, are removed. After this step, the number of training and testing sentences for the SemEval-2015 dataset is 1279 and 597, respectively. For the SemEval-2016 dataset, the training and testing samples become 1880 and 650. Then, we use the NLTK library [2] to remove punctuation and tokenize words.

In this research, we use the 300-dimensional GloVe word representations that were pre-trained on 42 billion tokens from Common Crawl [22]. This word embedding matrix contains representations for 1.9 million words. The reason for this choice is twofold. Firstly, since contextual word embeddings result in different context-dependent word representations, the employment of the seed regularization presented in Section 4.5 is facilitated by the usage of non-contextual word embeddings. Secondly, [31] shows that the hybrid model HAABSA++ developed using GloVe representations outperforms all other non-contextual word embeddings and achieves results comparable to those obtained using context-dependent embeddings. The possibility of utilizing context-dependent embedding is left as a suggestion for future research. The embeddings of the words that are not present in the GloVe vocabulary are initialized randomly according to a uniform distribution U(-0.01, 0.01).

4 METHODOLOGY

The main goal of this paper is to develop an unsupervised approach that is able to categorize the sentiment expressed towards a pre-specified target as either positive, negative, or neutral. The models proposed are inspired by the work done by [11] in unsupervised aspect extraction. While in the latter the final goal is to develop a set of aspect embeddings, here the objective is to construct a set of sentiment embeddings for the positive/negative/neutral polarity classification. Both proposed models consist of two steps, an attention-based sentence representation and a sentence reconstruction. After presenting common preliminary steps in Section 4.1, Sections 4.2 and 4.3 introduce the first step for ABABSE1 and ABABSE2, respectively. The sentence reconstruction step, instead, is identical for both models and is explained in Section 4.4.

4.1 Preliminaries

Let $s = [w_1, w_2, \ldots, w_N]$ be an input sentence formed by N words. The actual input that is fed to the model is the vector representation of the input words. To create these, each word w present in the vocabulary is linked to a feature vector $\mathbf{e}_w \in \mathbb{R}^d$, where d represents the dimension of each word vector. In this paper, we use the GloVe pre-trained word embeddings [22]. The word embedding matrix is defined as $\mathbf{E} \in \mathbb{R}^{V \times d}$, where V represents the total number of words in the vocabulary. The goal of both models is to develop a sentiment embedding matrix defined as $\mathbf{P} \in \mathbb{R}^{C \times d}$, where C is the number of sentiments for which we want to find a representation in the embedding space. In this research the sentiments can be either positive, neutral, or negative, and thus C equals three.

4.2 Attention-Based Aspect-Based Sentiment Extraction 1

A visual representation of ABABSE1 is presented in Figure 2. In the equations specific to this model we use the subscript "ABABSE1".

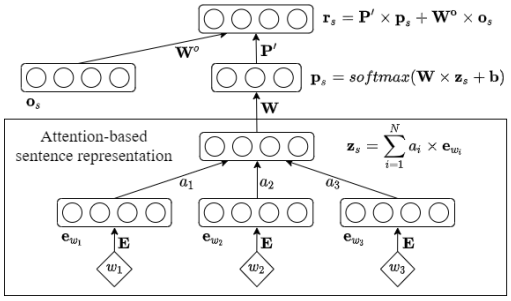

Figure 2: An example of the ABABSE1 structure.

The first step entails creating a representation of the sentence in the embedding space $\mathbf{z}_{s,ABABSE1} \in \mathbb{R}^d$. To do so, $\mathbf{z}_{s.ABABSE1}$ is defined as the weighted average of the word embeddings as follows:

$$\mathbf{z}_{s_{ABABSE1}} = \sum_{i=1}^{N} a_{i_{ABABSE1}} \times \mathbf{e}_{w_i}. \tag{1}$$

In order to construct a sentence representation that contains the most important information, an attention mechanism is employed to focus on the words that are most relevant for the given target. A weight $a_{i_{ABABSE1}} \in [0, 1]$ is determined for each vocable w_i in the given sentence. This weight can be interpreted as the probability that the corresponding word is relevant for detecting the target-specific sentiment. First, an average representation of the target ($\mathbf{y}_{t_{ABABSE1}}$) is computed by means of an average pooling layer applied to the M word embeddings forming the target as follows:

$$\mathbf{y}_{t_{ABABSE1}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{e}_{w_m}. \tag{2}$$

Then, the attention mechanism is constructed where each vocable w_i with $i \in 1, 2, .., N$ in the sentence is assigned an attention score d_i according to Equation 3. There, the entries of the weight matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ and the bias b are learned during the training process. The *tanh* activation function maps all the values between -1 and 1 and it allows for non-linearity.

$$d_{i_{ABABSE1}} = tanh(\mathbf{e}'_{w_i} \times \mathbf{M} \times \mathbf{y}_{t_{ABABSE1}} + b) \tag{3}$$

Given the attention scores, the attention weights can be computed by employing a softmax function which scales all the sentence scores in the [0, 1] range as follows:

$$a_{i_{ABABSE1}} = \frac{\exp(d_{i_{ABABSE1}})}{\sum_{j=1}^{N} \exp(d_{j_{ABABSE1}})}. \tag{4}$$

4.3 Attention-Based Aspect-Based Sentiment Extraction 2

The ABABSE2 model builds upon ABABSE1. Since the word embeddings used are context-independent, ABABSE1 might suffer from a lack of contextual information. To overcome this weakness, ABABSE2 includes a Bi-LSTM layer that is able to model contextual information in both left-to-right and right-to-left directions. ABABSE2 exploits the contextual information in the attention mechanism in order to detect with more precision the most important sentiment-carrying words. Given the structural similarity between ABABSE1 and ABABSE2, the visual representation of the latter is reported in the Appendix A in Figure 5.

In the notation that follows, we use the subscript "ABABSE2". The first step in ABABSE2 is to feed the word embeddings to a Bi-LSTM layer. Then, for each word embedding $\mathbf{e}_{w_i}$, with $i \in 1, 2, .., N$, the Bi-LSTM returns the hidden outputs $\mathbf{h}_{i,f} \in \mathbb{R}^d$ and $\mathbf{h}_{i,b} \in \mathbb{R}^d$ for the forward (left-to-right) and backward (right-to-left) propagation, respectively. By averaging the forward and backward representations, we construct a hidden output $\mathbf{h}_i \in \mathbb{R}^d$ that captures both the right-to-left and left-to-right contextual information. The hidden outputs are then used to construct a contextual target-based attention mechanism. First, we build a representation of the target ($\mathbf{y}_{t_{ABABSE2}}$) by averaging its hidden outputs as in Equation 5.

$$\mathbf{y}_{t_{ABABSE2}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{h}_m \tag{5}$$

Then, the attention mechanism is constructed where each word w_i with $i \in 1, 2, .., N$ is assigned a target-based attention score $d_{i_{ABABSE2}}$ (Equation 6), which is then used in Equation 7 to compute the attention weights $a_{i_{ABABSE2}}$. As before, the matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ and the bias b are learned during training. Note, while ABABSE1 employs the word embeddings to compute the attention score, ABABSE2 uses the hidden outputs computed by the Bi-LSTM ($\mathbf{h}_i$).

$$d_{i_{ABABSE2}} = tanh(\mathbf{h}'_i \times \mathbf{M} \times \mathbf{y}_{t_{ABABSE2}} + b) \tag{6}$$

$$a_{i_{ABABSE2}} = \frac{\exp(d_{i_{ABABSE2}})}{\sum_{j=1}^{N} \exp(d_{j_{ABABSE2}})} \tag{7}$$

The sentence representation $\mathbf{z}_{s_{ABABSE2}} \in \mathbb{R}^d$ is computed as the weighted average of the word embeddings as in Equation 8. The choice to use the word embeddings instead of the hidden outputs of the Bi-LSTM, as often done in the supervised realm [31], is to constrain the Bi-LSTM and prevent it from creating a sentence representation that is loosely related to the original sentence but that facilitates the sentence reconstruction task presented in Section 4.4. Other restrictions could be employed to use the hidden outputs to construct the sentence representation. For example, penalizing sentence representations that differ significantly from the average of the word embeddings would constrain the behavior of the Bi-LSTM. However, this analysis is left for future research.

$$\mathbf{z}_{s_{ABABSE2}} = \sum_{i=1}^{N} a_{i_{ABABSE2}} \times \mathbf{e}_{w_i} \tag{8}$$

4.4 Sentence Reconstruction Using Sentiment Embeddings

The second phase of the procedure works essentially as an autoencoder. The objective is to convey the information carried by the attention-based sentence representation through a lower dimension and then reconstruct the sentence representation utilizing the sentiment embedding matrix $\mathbf{P} \in \mathbb{R}^{C \times d}$. Learning the optimal embedding matrix essentially means discovering the most relevant semantic areas in the word embedding space [37].

First, the sentence representation is reduced to the C-dimensional $\mathbf{p}_s \in \mathbb{R}^C$ vector as presented in Equation 9. The softmax activation function introduces non-linearity and allows us to interpret each node activation as the probability that the target-specific sentiment expressed in the sentence is either positive, negative, or neutral. Both $\mathbf{W} \in \mathbb{R}^{C \times d}$ and $\mathbf{b} \in \mathbb{R}^{C \times 1}$ are learned during training.

$$\mathbf{p}_s = softmax\,(\mathbf{W} \times \mathbf{z}_s + \mathbf{b}) \tag{9}$$

Second, the model reconstructs the original sentence representation by means of the sentiment embedding matrix $\mathbf{P}$. [37] performs both aspect and sentiment extraction simultaneously showing that the two tasks are correlated. Consequently, since the aspect category of the target is assumed to be known in this paper, we use one-hot-encoding to include an additional binary vector $\mathbf{o}_s \in \mathbb{R}^K$ that represents a distribution over the K aspect categories available. Then, the reconstruction $\mathbf{r}_s \in \mathbb{R}^d$ is computed as follows, where both $\mathbf{P} \in \mathbb{R}^{C \times d}$ and $\mathbf{W^o} \in \mathbb{R}^{d \times k}$ are learned during training:

$$\mathbf{r}_s = \mathbf{P'} \times \mathbf{p}_s + \mathbf{W^o} \times \mathbf{o}_s. \tag{10}$$

4.5 Unsupervised Training Procedure

Each model is trained to minimize the difference between the sentence representation $\mathbf{z}_{s,i}$ and its reconstruction $\mathbf{r}_s$, with $i \in \{ABABSE1, ABABSE2\}$. Following the methodology suggested by [11], the model is also trained to maximize the difference between the reconstruction $\mathbf{r}_s$ and the sentence representation $\mathbf{x}_q \in \mathbb{R}^d$ of other Q distinct and randomly selected sentences, called negative samples, where $\mathbf{x}_q$ is the average of the vector embeddings of the words in the sentence. In this paper, the difference between two vectors is measured using the cosine similarity (see Equation 11) which outputs a value in the [-1,1] range, where -1 and 1 indicate

dissimilar and similar vectors, respectively. The model is therefore trained in an unsupervised manner minimizing a hinge loss (Equation 12) over the training data set D [11].

$$sim(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \times \|\mathbf{b}\|} \tag{11}$$

$$J(\Theta) = \sum_{s \in D} \sum_{q=1}^{Q} \max(0, 1 - sim(\mathbf{r}_s, \mathbf{z}_s) + sim(\mathbf{r}_s, \mathbf{x}_q)) \tag{12}$$

In addition, following the methodology proposed by [37], both a redundancy and a seed regularization are employed. The redundancy regularization ensures that the sentiment embeddings are sufficiently different between each other. This regularization was employed also by [11] and is computed as in Equation 13. $\mathbf{P}_n \in \mathbb{R}^{C \times d}$ is the sentiment embedding matrix, where each row is normalized to length 1 (hence the subscript n), and $\mathbf{I} \in \mathbb{R}^{C \times C}$ is the identity matrix. Clearly, $U(\Theta)$ is minimized when the dot product between different rows of $\mathbf{P}_n$ is close to zero, indicating that this regularization term enforces orthogonality between the sentiment embeddings.

$$U(\Theta) = \|\mathbf{P_n} \times \mathbf{P_n}' - \mathbf{I}\| \tag{13}$$

The seed regularization is employed to exploit common-sense knowledge. To do so, the user defines a set of seed words for each of the polarity classes which are then used to constrain the sentiment embedding matrix $\mathbf{P}$. For example, if the seed words "good", "bad", and "indifferent" are defined, respectively, as positive, negative, and neutral, then the sentiment embeddings are penalized when deviating significantly from the word embedding of the given seed. In practice, if G seed words are provided for each sentiment, then the c-th row of the "prior" matrix $\mathbf{R} \in \mathbb{R}^{C \times d}$ [37] is constructed by averaging the embedding of the seed words as follows:

$$\mathbf{R}_c = \frac{1}{G} \sum_{g=1}^{G} \mathbf{e}'_{w_g}. \tag{14}$$

Given the "prior" matrix, the seed regularization is developed to maximize the similarity between each row of $\mathbf{R}$ and $\mathbf{P}$. Thus, with $\mathbf{R}_c$ and $\mathbf{P}_c$ being the c-th row of the two matrices, the regularization is computed as follows:

$$V(\Theta) = \sum_{c=1}^{C} [1 - sim(\mathbf{R}_c, \mathbf{P}'_c)]. \tag{15}$$

The final objective of the model is to minimize the loss function $J(\theta)$ subject to the regularization terms whose influence is determined by the hyperparameters λ_1 and λ_2. Moreover, in order to reduce the likelihood of overfitting, the L_2 regularization term is included in the loss function and its influence is controlled by the hyperparameter λ_3. The overall loss function $L(\Theta)$ is presented in Equation 16, where Θ represents the set of parameters of the model.

$$L(\Theta) = J(\Theta) + \lambda_1 U(\Theta) + \lambda_2 V(\Theta) + \lambda_3 \|\Theta\|^2 \tag{16}$$

The loss function is minimized using backpropagation [26]. All the parameters are initialized randomly following a uniform distribution U(-0.1, 0.1) and are updated using the Adam optimizer [14]. The parameters of the Adam algorithm (β_1 and β_2) are treated as hyperparameters and are re-tuned for each model and dataset. Moreover, in order to reduce the negative effect of overfitting, the dropout technique [30] is employed. The dropout rate is also a hyperparameter that is tuned specifically for each dataset and method.

Besides the aforementioned parameters of the Adam optimizer and dropout rate, also the learning rate, the batch size, the number of negative samples, and the regularization weights employed in the loss function (λ_1, λ_2, and λ_3) must be tuned simultaneously. To do so, 20% of the training observations are employed as a validation set. The optimal hyperparameters are those resulting in the lowest reconstruction loss $J(\Theta)$ (see Equation 12) for the validation set. In this research, we use the Tree-structured Parzen Estimator (TPE) [1] algorithm. Once the optimal hyperparameters are found, the full training set is used to fine-tune the model.

5 RESULTS

In this section, we test the performance of the proposed models. Section 5.1 introduces the baseline algorithms used in this paper. In Section 5.2 we provide an extended comparison of ABABSE1 and ABABSE2. Last, the best model is compared against the baseline models in Section 5.3.

5.1 Baseline Models

The considered baseline models are given below.

Majority heuristic: This heuristic always predicts the majority class, i.e, the positive sentiment. This heuristic can achieve high accuracy in case the dataset is strongly unbalanced.

W2VLDA [10]: A weakly supervised model that builds upon a topic modeling approach and performs sentiment classification by adjusting the parameters of the generative Dirichlet distribution based on seed words that are sampled and used for training. The results are available only for the SemEval-2016 dataset.

V3 [9]: An unsupervised approach that uses the known polarity of seed words and the distance in the word2vec embedding space to assign a polarity to all words. Majority voting determines the sentence polarity. The results are available only for SemEval-2015.

Ontology [28]: A dictionary-based approach that uses in-domain knowledge to construct a lexicon. The sentiment, only positive or negative, is obtained by ontology reasoning.

LCR-Rot [36]: A supervised approach that employs a Left-Center-Right separated neural network with Rotatory attention.

HAABSA++ [31]: A hybrid approach that combines an ontology with the LCR-Rot-hop supervised model [34]. The latter was improved by the addition of a hierarchical attention mechanism and use of deep contextual word embeddings.

The baseline models' results are taken from the original articles.

5.2 Comparison ABABSE1 and ABABSE2

The comparison between ABABSE1 and ABABSE2 is based on their predictive performance, the quality of the attention mechanism, and the quality of the sentiment embedding dictionary learned. The results presented hereafter correspond to those obtained using the tuned hyperparameters (see Table 6 in Appendix B) and the following five seed words per class: [amazing, great, nice, impeccable, excellent] for positive, [rude, bad, terrible, awful, horrible] for negative, and [mediocre, ordinary, decent, average, ok] for neutral.

The predictive performance is evaluated using the following five measures: in-sample accuracy (in-A), out-sample accuracy (out-A), precision (P), recall (R), and $F1$ scores, where the last three are all out-of-sample. The average results are presented in Table 3. In order to reduce the effect of random fluctuations on the scores, we report the average over ten runs which implies that the F1 scores cannot be directly computed using the reported P/R values.

Table 3: Comparison using in-sample accuracy (in-A), out-of-sample accuracy (out-A), precision (P), recall (R), and $F1$ score ($F1$). Largest values are in bold

Dataset	Model	in-A	out-A	P	R	$F1$
2015	ABABSE1	72.9%	69.1%	47.3%	50.8%	47.2%
	ABABSE2	**77.8%**	**73.8%**	**49.3%**	**51.1%**	**49.4%**
2016	ABABSE1	75.5%	75.3%	44.6%	45.2%	42.8%
	ABABSE2	**77.3%**	**78.8%**	**47.7%**	**46.7%**	**46.3%**

Clearly, ABABSE2 performs better than ABABSE1 in any performance measure for both datasets. ABABSE2 outperforms ABABSE1 by 4.7% and 3.5% in the out-of-sample prediction accuracy for SemEval-2015 and SemEval-2016, respectively. These results are evidence that the inclusion of the Bi-LSTM in the attention mechanism did improve the predictive capacity of the model significantly.

It is interesting to observe that ABABSE1 and ABABSE2 show relatively low P/R/F1 scores compared to the reported accuracies and that the latter are overall higher for the SemEval-2016 dataset. This is due to the fact that both models are on average unable to detect the minority (neutral) class, and thus their accuracy improves when the neutral class corresponds to a lower percentage of the test set (6.2% for SemEval-2015 and 4.9% SemEval-2016).

In order to understand why ABABSE2 outperforms ABABSE1, we study the attention weights that are learned by each model. The attention mechanism selects the most relevant words by assigning larger weights to them. Since here the objective is target-specific sentiment classification, we expect the attention mechanism to extract the sentiment-carrying words related to the target. To facilitate this analysis, we visualize the sentences as in Figure 3, where darker colors correspond to larger attention weights. The sum of the attention weights over the sentence equals one for both models.

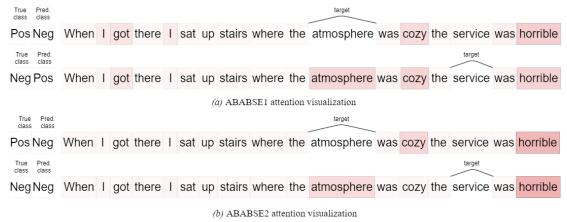

Figure 3: Attention visualizations of the ABABSE1 (*a*), and ABABSE2 (*b*) models for the sentence "When I got there I sat up stairs where the atmosphere was cozy the service was horrible".

The first example, from the SemEval-2016 dataset, is presented in Figure 3 where the sentence is: "When I got there I sat up stairs where the atmosphere was cozy the service was horrible". Note, in this sentence, there are two targets (i.e., "atmosphere" and "service") and thus two target-specific sentiments are predicted. In this case, ABABSE1 predicts the wrong sentiment for both targets, while ABABSE2 correctly classifies one of the two. The ABABSE1 model (Figure 3a) is able to detect the sentiment-carrying words, i.e., "cozy" and "horrible", but it is not able to specifically focus on either of the two words depending on whether the target studied is "atmosphere" or "service". Moreover, ABABSE1 assigns a small weight to other words that a human would judge as irrelevant for the sentiment classification task. On the other hand, ABABSE2 (Figure 3b) is able to correctly classify the negative sentiment expressed towards the "service" because it focuses on the correct word "horrible". With respect to "atmosphere", though, also ABABSE2 is unable to neglect the negative sentiment carried by the word "horrible".

Figure 4: Attention visualizations of the ABABSE1 (*a*), and ABABSE2 (*b*) models for the sentence "We stood there for 10 minutes while employees walked back and forth ignoring us".

Another relevant example from the SemEval-2016 dataset is presented in Figure 4 where the sentence studied is: "We stood there for 10 minutes while employees walked back and forth ignoring us". In this sentence the sentiment is given in an implicit manner. In this case, ABABSE1 (Figure 4a) mistakenly classifies the sentiment towards "employees" as positive. On the other hand, ABABSE2 (Figure 4b) makes the correct prediction. The difference between ABABSE1 and ABABSE2 can be explained by the fact that the latter detects the relevant word, i.e., "ignoring", while the former distributes uniformly the weights among the words in the sentence.

Table 4: List of the ten nearest neighbors for each sentiment embedding learned by ABABSE1 and ABABSE2 in SemEval-2016 using five seed words

Class	ABABSE1	ABABSE2
Pos	excellent, fantastic, nice, great, amazing, terrific, superb, wonderful, awesome, fabulous	fantastic, superb, amazing, excellent, great, wonderful, nice, terrific, awesome, incredible
Neg	horrible, awful, terrible, horrid, bad, rude, hideous disgusting, worst, horrific	horrible, terrible, awful, bad, horrid, rude, horrific worst, disgusting, nasty
Neu	average, decent, than, less, typical, normal, higher, comparable, ordinary, compared	average, decent, ordinary, mediocre, less, compared, normal, than, expect, higher

These examples provide strong evidence in favor of the hypothesis that the inclusion of the Bi-LSTM within the attention mechanism helps in focusing on the correct words. In particular, the

second example suggests that when the sentiment is expressed in a less direct manner, extracting contextual information via a Bi-LSTM helps in computing more meaningful attention weights.

Then, we explore the sentiment embedding matrix $\mathbf{P} \in \mathbb{R}^{C \times d}$ constructed by each model. We do so by searching the ten nearest neighbors in the GloVe embedding space using the cosine similarity metric. Table 4 presents the most representative words found for each sentiment class by ABABSE1 and ABABSE2 when using five seed words per class. It can be concluded that all models construct a precise and meaningful sentiment embedding. This holds true even when fewer seed words are used per class, indicating that the models require little pre-determined knowledge. Appendix C presents an analysis of how the number of seed words affects the predictive performance of the models.

5.3 Performance Results

The best-performing model, ABABSE2, is compared against the baseline models presented in Section 5.1. This comparison is based on the out-of-sample accuracies for both the SemEval-2015 and SemEval-2016 datasets, as this is the standard evaluation measure for ABSC. The results are summarized in Table 5, where unsupervised and supervised models are introduced separately. Where available, we present both in-sample and out-sample accuracy results. Note, the ABABSE2 results are the average over ten runs.

Table 5: Comparison of the models using out-of-sample accuracy (out-A) and in-sample accuracy (in-A). Largest values are in bold

	SemEval-2015		SemEval-2016	
	out-A	in-A.	out-A	in-A
Unsupervised models				
Majority heuristic	59.1%	75.3%	74.2%	70.2%
W2VLDA	-	-	77.3%	-
V3	69.4%	-	-	-
Ontology	65.8%	**79.7%**	78.3%	75.3%
ABABSE2	**73.8%**	77.8%	**78.8%**	**77.3%**
Supervised models				
LCR-Rot	78.4%	86.2%	86.9%	**92.9%**
HAABSA++	**81.7%**	**88%**	**87.0%**	88.9%

When focusing on the unsupervised (or weakly supervised) set of algorithms, ABABSE2 shows an improvement in performance compared to the baseline models. When comparing the out-of-sample accuracies for the SemEval-2015, ABABSE2 outperforms the majority class, V3, and the ontology by 14.7%, 4.4%, and 8.0%, respectively. Similarly, the out-of-sample accuracies for the SemEval-2016 indicate that ABABSE2 performs better than all three baselines exceeding the majority class by 4.6%, the W2VLDA by 1.5%, and the ontology by 0.5%. Consequently, ABABSE2 outperforms all investigated unsupervised (or weakly supervised) algorithms achieving the state-of-the-art accuracies of 73.8% and 78.8% on the SemEval-2015 and SemEval-2016 datasets, respectively. Of particular interest are the comparative results of ABABSE2 and the ontology. The ontology does not require a training set as it is simply a collection of rules and dictionaries; however, the development of such an ontology requires in-domain knowledge and can be very costly in terms of time required. Therefore, the fact that ABABSE2 achieves comparable results is remarkable.

As expected, when considering supervised algorithms, the difference in performance is quite substantial. The state-of-the-art algorithm HAABSA++ outperforms ABABSE2 by 7.9% and 8.2% on the SemEval-2015 and SemEval-2016 test sets, respectively. Similarly, the LCR-Rot model performs better than ABABSE2 by 8.1% on the SemEval-2016 dataset; however, it is surprising to see that the performance difference between the LCR-Rot model and ABABSE2 shrinks to 4.6% when evaluated on the SemEval-2015 dataset. Last, it is worth noticing that the gap between in-sample and out-of-sample accuracy is smaller for ABABSE2 compared to LCR-Rot and HAABSA++. As expected, the unsupervised nature of ABABSE2 reduces the influence of overfitting compared to supervised models.

Overall, ABABSE2 performs well on both datasets used in this research, outperforming all investigated unsupervised and weakly supervised algorithms, and achieving state-of-the-art results.

6 CONCLUSION

In this paper, we introduce two novel attentional neural networks for unsupervised ABSC. The first model (ABABSE1) adapts the ABAE model [11] for ABSC. As ABAE, ABABSE1 uses a target-based attention mechanism to detect the sentiment-carrying words and employs a redundancy regularization term. However, ABABSE1 extends the ABAE framework by adding a seed regularization to exploit common knowledge in the form of seed words. ABABSE2 further extends ABABSE1 by including a Bi-LSTM layer in the attention mechanism to capture contextual information.

The models are evaluated on their out-of-sample predictions using the SemEval-2015 task 12 [24] and the SemEval-2016 task 5 subtask 1 [23] restaurant datasets. We tune the hyperparameters using the TPE algorithm [1] and 20% of the training observations as a validation set. With an average out-of-sample accuracy of 73.8% and 78.8% on the SemEval-2015 and SemEval-2016 datasets, respectively, ABABSE2 achieves the best results. This performance difference is likely due to the attention mechanism learned by each model. In fact, both models detect the sentiment-carrying words, but ABABSE2 is better at identifying target-specific words, especially when the sentiment is expressed implicitly, evidencing the benefits of using Bi-LSTMs within the attention mechanisms. By analyzing the ten nearest neighbors of the learned sentiment embeddings, we conclude that both models construct meaningful sentiment embeddings, even when only one seed word is used.

The best model (ABABSE2) is then compared against the current state-of-the-art unsupervised and supervised algorithms for ABSC. ABABSE2 performs better than W2VLDA [10] and V3 [9], and achieves comparable results to the ontology introduced by [28]. These results evidence the benefits of using attentional neural networks in an unsupervised manner. When juxtaposed against supervised algorithms, ABABSE2 was outperformed by around 8%; however, it is concluded that it suffers less from the overfitting issue induced by the unbalancedness of the datasets.

Since this research focused on the restaurant domain, the applicability of the models to other domains has not been tested yet. This is left as a suggestion for future research. In particular, it would be interesting to evaluate the performance on the laptop

domain, as it is deemed more challenging due to its more numerous domain-specific words and aspect categories [23].

With regards to the models, ABABSE2 employs a Bi-LSTM layer to capture contextual information, but this knowledge is only used within the attention mechanism. A suggestion for future research is to employ the hidden outputs of the Bi-LSTM also in the creation of the sentence representation while constraining the Bi-LSTM by means of regularization terms. Last, one could study how the performance would change if contextual word embeddings are employed instead of the 300-dimensional GloVe [22] word embeddings.

REFERENCES

[1] James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS 2011)*. Curran Associates, 2546–2554.

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[4] Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. 2020. SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis. In *29th ACM International Conference on Information and Knowledge Management (CIKM 2020)*. ACM, 105–114.

[5] Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. SenticNet: A Publicly Available Semantic Resource for Opinion Mining.. In *2010 AAAI Fall Symposium: Commonsense Knowledge (AAAIFS 2010)*. AAAI.

[6] Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications* 118 (2019), 272–299.

[7] Nikoleta Dosoula, Roel Griep, Rick Den Ridder, Rick Slangen, Kim Schouten, and Flavius Frasincar. 2016. Detection of multiple implicit features per sentence in consumer review data. In *12th International Baltic Conference on Databases and Information Systems (DB&IS 2016) (CCIS, Vol. 615)*. Springer, 289–303.

[8] James F Engel, Robert J Kegerreis, and Roger D Blackwell. 1969. Word-of-mouth communication by the innovator. *Journal of Marketing* 33, 3 (1969), 15–19.

[9] Aitor García-Pablos, Montse Cuadros, and German Rigau. 2015. V3: Unsupervised aspect based sentiment analysis for SemEval2015 task 12. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. ACL, 714–718.

[10] Aitor García-Pablos, Montse Cuadros, and German Rigau. 2018. W2VLDA: almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications* 91 (2018), 127–137.

[11] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)* . ACL, 388–397.

[12] Paul M Herr, Frank R Kardes, and John Kim. 1991. Effects of word-of-mouth and product-attribute information on persuasion: An accessibility-diagnosticity perspective. *Journal of Consumer Research* 17, 4 (1991), 454–462.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[14] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR 2015)*.

[15] Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. ACL, 437–442.

[16] Chenghua Lin, Yulan He, Richard Everson, and Stefan Ruger. 2012. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering* 24, 6 (2012), 1134–1145.

[17] Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

[18] Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. 2018. Content attention model for aspect based sentiment analysis. In *Proceedings of the 27th International World Wide Web Conference (WWW 2018)*. ACM, 1023–1032.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations (ICLR 2013)*.

[20] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3, 4 (1990), 235–244.

[21] Bo Pang and Lillian Lee. 2007. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2, 1-2 (2007), 1–135.

[22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2014)*. ACL, 1532–1543.

[23] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. ACL, 19–30.

[24] Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. ACL, 486–495.

[25] Ana-Maria Popescu, Bao Nguyen, and Oren Etzioni. 2005. OPINE: Extracting Product Features and Opinions from Reviews. In *2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*. ACL, 32–33.

[26] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature* 323, 6088 (1986), 533–536.

[27] Kim Schouten and Flavius Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* 28, 3 (2016), 813–830.

[28] Kim Schouten and Flavius Frasincar. 2018. Ontology-driven sentiment analysis of product and service aspects. In *Proceedings of the 15th Extended Semantic Web Conference (ESWC 2018) (LNCS, Vol. 10843)*. Springer, 608–623.

[29] Kim Schouten, Flavius Frasincar, and Franciska de Jong. 2017. Ontology-enhanced aspect-based sentiment analysis. In *Proceedings of the 17th International Conference on Web Engineering (ICWE 2017) (LNCS, Vol. 10360)*. Springer, 302–320.

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[31] Maria Mihaela Trusca, Daan Wassenberg, Flavius Frasincar, and Rommert Dekker. 2020. A hybrid approach for aspect-based sentiment analysis using deep contextual word embeddings and hierarchical attention. In *Proceedings of the 20th International Conference on Web Engineering (ICWE 2020) (LNCS, Vol. 12128)*. Springer, 365–380.

[32] Ellen Van Kleef, Hans CM Van Trijp, and Pieternel Luning. 2005. Consumer research in the early stages of new product development: a critical review of methods and techniques. *Food Quality and Preference* 16, 3 (2005), 181–201.

[33] Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. 2014. DCU: Aspect-based polarity classification for SemEval task 4. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. ACL, 223–229.

[34] Olaf Wallaart and Flavius Frasincar. 2019. A hybrid approach for aspect-based sentiment analysis using a lexicalized domain ontology and attentional neural models. In *Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019) (LNCS, Vol. 11503)*. Springer, 363–378.

[35] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*. ACL, 606–615.

[36] Shiliang Zheng and Rui Xia. 2018. Left-center-right separated neural network for aspect-based sentiment analysis with rotatory attention. *arXiv preprint arXiv:1802.00892* (2018).

[37] Honglei Zhuang, Fang Guo, Chao Zhang, Liyuan Liu, and Jiawei Han. 2020. Joint Aspect-Sentiment Analysis with Minimal User Guidance. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*. ACM, 1241–1250.

A ABABSE2 VISUAL REPRESENTATION

This appendix presents the visual representation of the ABABSE2 model. The structure resembles that of ABABSE1 to which the Bi-LSTM is added in the attention mechanism.

B HYPERPARAMETERS

This appendix presents the list of optimal hyperparameters for each model and each dataset as resulted after tuning the parameters using the Tree-structured Parzen Estimator (TPE) [1].

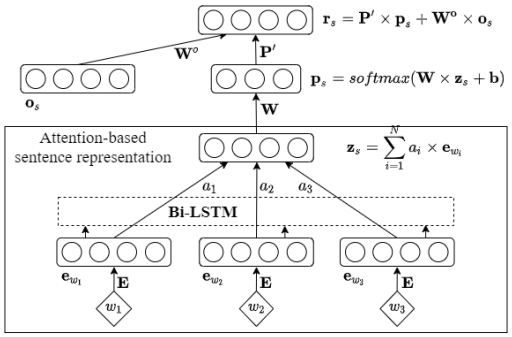

Figure 5: An example of the ABABSE2 structure.

Table 6: Optimal hyperparameters for each model and each dataset

	SemEval-2015		SemEval-2016	
	ABABSE1	ABABSE2	ABABSE1	ABABSE2
learning rate	0.01	0.01	0.005	0.005
dropout rate	5%	5%	5%	5%
β_1	0.95	0.99	0.99	0.97
β_2	0.99	0.99	0.99	0.99
λ_3	0.04	0.04	0.04	0.04
λ_2	5	5	5	5
λ_1	5	5	1	1
batch size	30	30	30	30
neg. samples	10	10	10	10
#epochs	3	2	3	2

C IMPACT OF SEED WORDS

In this appendix, we provide an analysis of the impact that the number of seed words has on the predictive performance of the models proposed in this paper. Assessing the level of human intervention needed to achieve the desired results is particularly useful for real-life applications.

In Figure 6 we plot the average out-of-sample accuracies for each model against the number of sentiment seed words for both datasets, ceteris paribus. With regards to ABABSE2, as expected, on average the accuracy increases when more human knowledge is exploited in the form of seed words. On the other hand, the behaviour of ABABSE1 is surprising. While the performance improves as the number of seed words is increased from one to two, its performance levels off or even decreases as the number of seed words exceeds three. This behavior is clearly visible in Figure 6 where the predictive performance of ABABSE1, evaluated on SemEval-2015 and SemEval-2016, reaches its maximum, on average, when three and two seed words are employed, respectively. This contrasting behaviour could be very subjective to the seed words used in this experiment and further analysis on the effect of seed words might be required in future research. It should also be noticed that both models achieve good accuracies already when only one seed word is used. We conclude that both models are able to achieve good performances with little human intervention, but that ABABSE2 does improve significantly, on average, when more seed words are used, probably due to its higher complexity.

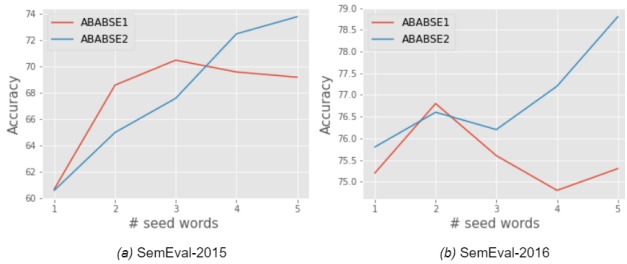

Figure 6: Out-of-sample accuracy as a function of the number of seed words per sentiment class for SemEval-2015 (*a*) and SemEval-2016 (*b*).