Language-Agnostic Model for Aspect-Based Sentiment Analysis

Md Shad Akhtar†, Abhishek Kumar†, Asif Ekbal†, Chris Biemann⋆ and Pushpak Bhattacharyya†
†*Department of CSE, Indian Institute of Technology Patna*
{shad.pcs15,abhishek.ee14,asif,pb}@iitp.ac.in
⋆*Universität Hamburg, Germany*
biemann@informatik.uni-hamburg.de

Abstract

In this paper, we propose a language-agnostic deep neural network architecture for aspect-based sentiment analysis. The proposed approach is based on Bidirectional Long Short-Term Memory (Bi-LSTM) network, which is further assisted with extra hand-crafted features. We define three different architectures for the successful combination of word embeddings and hand-crafted features. We evaluate the proposed approach for six languages (i.e. *English*, *Spanish*, *French*, *Dutch*, *German* and *Hindi*) and two problems (i.e. *aspect term extraction* and *aspect sentiment classification*). Experiments show that the proposed model attains state-of-the-art performance in most of the settings.

1 Introduction

Sentiment analysis (Pang and Lee, 2008) is often target-centric. In aspect-based sentiment analysis (ABSA), we aim to identify the polarity of expressed sentiments towards a feature or aspect. These features or aspects are usually explicitly mentioned in the text. Also, a sentence may contain more than one aspect terms, and the task is to assign separate sentiments to each of them, e.g. in *"The **food** was great! But **service** was below par."* there are two aspects ('*food*' and '*service*'), and the expressed sentiment towards *food* and *service* are *positive* and *negative*, respectively. Such analysis offers fine-grained information to a user or an organization who seeks users opinion towards any specific entity. For example, based on the users' feedback, an individual can draw a general perception about the specific attribute or aspect of a product or service, and he/she can make an informed decision about the product or service under observation. Similarly, an organization can utilize the feedback to refine its product/service or to take a decision in the business model.

Aspect-based sentiment analysis (Pontiki et al., 2014, 2016) has two subproblems at its core, i.e., aspect term identification (or opinion target extraction) and aspect sentiment classification. Given a text, aspect term identification task aims to find the boundaries of all the aspect terms present in the text, whereas aspect sentiment classification task classifies each of these identified aspect terms into one of the predefined sentiment classes (e.g., *positive*, *negative*, *neutral* etc.). A sentence may contain any number of aspect terms or no aspect term at all. The terms 'aspect term' and 'opinion target' are often used interchangeably and refer to the same span of text.

Motivation and Contribution

A survey of the literature for ABSA suggests a number of works for different languages (Kumar et al., 2016; Brun et al., 2016; Çetin et al., 2016). Although the reported performance for these works are good, they usually suffer in handling the language diversity, i.e., the systems that reported state-of-the-art performance for one language typically do not work well for the other languages. The unavailability of such a generic system motivates us to build a language-agnostic model for aspect based sentiment analysis. We propose a generic deep neural network architecture that handles the language divergence to a great extent. Our model is based on Bidirectional Long Short-Term Memory (Bi-LSTM) network

(Graves et al., 2005) that also utilizes extra hand-crafted features. We evaluate our proposed approach for four European (i.e., *Spanish*, *French*, *Dutch* & *German*), one Indian (i.e., *Hindi*) and English languages. The contributions of our work are *three-fold*: a) we propose an efficient and generic neural network architecture that works across multiple languages; b) we utilize a small set of handcrafted features (one each for aspect extraction and aspect classification) for the training and evaluation; and c) we provide the new state-of-the-art performance for two problems of ABSA across six different languages.

Rest of the paper is organized as follows: In Section 2, we present the literature survey. The proposed methodology has been discussed in detail in Section 3. In Section 4, we furnished experimental results and provided the necessary analysis. Finally, we conclude in Section 5.

2 Related Works

Sentiment analysis is a well-studied problem of natural language processing for English language (Turney, 2002; Pang et al., 2002, 2005; Pang and Lee, 2008; Jagtap and Pawar, 2013; Kim and Hovy, 2006). However, in recent times, researchers have focused on various extensions of sentiment analysis, e.g., aspect based sentiment analysis (Pontiki et al., 2014; Kiritchenko et al., 2014; Akhtar et al., 2016), multi-lingual sentiment analysis (Balamurali et al., 2012; Mishra et al., 2017; Brun et al., 2016; Kumar et al., 2016), multi-modal sentiment analysis (Poria et al., 2017; Zadeh et al., 2018; Ghosal et al., 2018), sentiment analysis in Twitter (Ghosh et al., 2015; Mohammad et al., 2013) etc.

For ABSA, System GTI (Alvarez-López et al., 2016) used a Support Vector Machine (SVM) and Conditional Random Field (CRF) based approach for aspect extraction and sentiment classification, respectively. They used language-dependent features like lemmas and Part-of-Speech (PoS) tags to achieve the state-of-the-art score for aspect extraction in Spanish. IIT-TUDA (Kumar et al., 2016) also used a number of hand-crafted features like character n-grams, dependency relations, prefix and suffix for SVM and CRF. They achieved comparable performance for *Spanish*, *French* & *Dutch*. System XRCE (Brun et al., 2016) used a feedback ensemble network that obtained the best performance for aspect classification on the French dataset. System TGB (Çetin et al., 2016) used a Logistic Regression based model to address the aspect sentiment classification and reported to achieve the best score on Dutch dataset. Mishra et al. (2017) used a Bi-LSTM based model, whereas Naderalvojoud et al. (2017) adopted a deep recurrent neural network model for the German dataset. Akhtar et al. (2016) developed an aspect based sentiment analysis datasets for Hindi. They employed CRF and SVM for aspect term extraction and aspect sentiment classification, respectively. For aspect based sentiment analysis in English, Kiritchenko et al. (2014) reported the best performance in SemEval-2014 shared task on ABSA (Pontiki et al., 2014).

There have been few attempts at injecting handcrafted features into the neural network architecture for enhancing the overall performance (Akhtar et al., 2016; Araque et al., 2017) of sentiment analysis. Akhtar et al. (2016) combined CNN representation and optimized features for learning a Support Vector Machine. Authors in (Araque et al., 2017) proposed a classifier ensemble model that combines surface-level features and generic word vectors for the sentiment classification. However, our work differs from these systems in the following ways: **a)** we perform aspect level sentiment analysis for six different languages (belong to different language family); **b)** we propose four different architectures to successfully combine the neural network learned representations and the handcrafted features; **c)** the proposed architectures handle both aspect extraction (a sequence labelling task) and aspect sentiment classification (a classification task); and **d)** we achieve better performance for most of the problem/language pairs.

3 Proposed Method

Overall, aspect based sentiment analysis can be thought of as a two-step process, i.e. aspect term extraction and aspect sentiment classification. Aspect term extraction is a sequence labelling task where each token of a sentence needs to be classified as either inside the boundary of an aspect term or outside. We adopted ***BIO*** notation to mark each token as either ***B****egin*, ***I****ntermediate* or ***O****utside* of an aspect term. A '***B***' signifies the beginning of an aspect term and successive '***I***s' signify a multi-token aspect

term (e.g. *spicy tuna rolls*). A single-token aspect term will be tagged as '*B*'. For the second problem, i.e. aspect sentiment classification, we define a context window of size ± 5 around each aspect term and consider all the tokens within the window for an instance. The intuition behind such an approach is that the sentiment-bearing clue words often occur close to the aspect terms. An example scenario is depicting in Table 1.

Review:	***Rice***	was	good	but	the	main	attraction	was	***spicy***	***tuna***	***rolls***	.
BIO Notation:	**B**	O	O	O	O	O	O	O	**B**	**I**	**I**	O
Aspect Terms:	**Rice** and **Spicy tuna rolls**											

Context window (± 5)	$Prev_5$	$Prev_4$	$Prev_3$	$Prev_2$	$Prev_1$	$Aspect_{term}$	$Next_1$	$Next_2$	$Next_3$	$Next_4$	$Next_5$
Rice	null	null	null	null	null	***Rice***	was	good	but	the	main
Spicy tune roll	but	the	main	attraction	was	***spicy tuna roll***	.	null	null	nulll	null
Aspect Sentiment:	***Positive*** for Rice and ***Positive*** for Spicy tuna rolls.										

Table 1: An example review from restaurant domain and its respective processing for aspect term extraction (i.e. BIO notations) and aspect sentiment classification (i.e. contextual processing).

Our proposed neural network architecture employs a Bi-LSTM network for learning sentence embeddings, which are then fed to a fully-connected dense layer for classification. Given a sentence, we first compute the word embeddings of each word and feed them into the Bi-LSTM network at different time steps for the prediction. We refer to this architecture as A1. In addition, we inject extra hand-crafted manual features to assist the neural architecture. We design three architectures (i.e. A2, A3 & A4 in Figure 1) for the successful combination of word embeddings and the hand-crafted features. The basic difference among these three architectures are the way features are injected into the model. A high-level architecture of our proposed method is depicted in Figure 1.

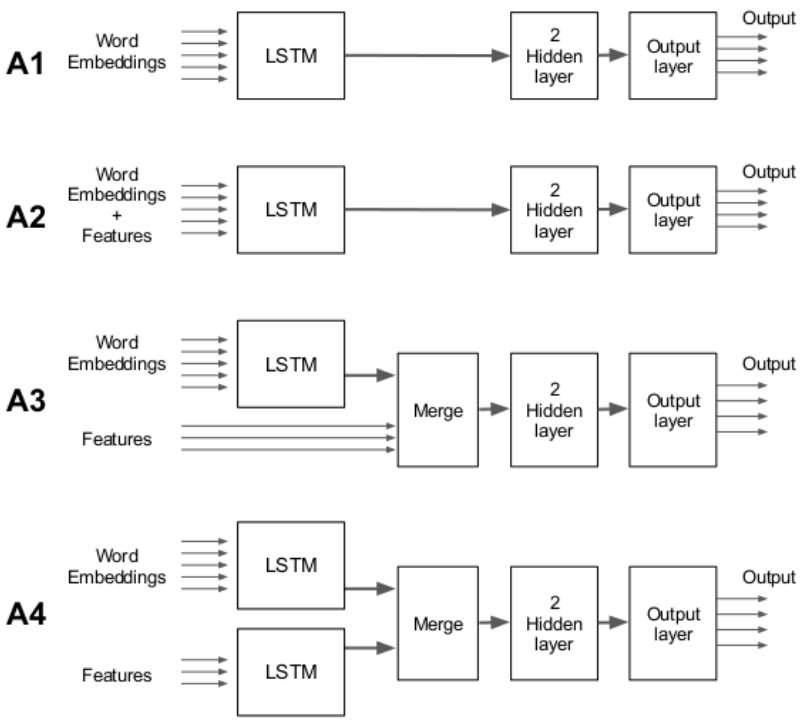

Figure 1: Proposed architectures for aspect identification and classification. **A1:** Only word embeddings are fed to Bi-LSTM network; **A2:** Word embeddings and extracted features are combined and fed into single Bi-LSTM network. **A3:** Extracted features are directly merged with Bi-LSTM output of word embedding. **A4:** One Bi-LSTM network each for word embeddings and extracted features. All the four architectures are language-agnostic in nature.

Architecture A1 makes use of word embeddings as the sole input for the network. In A2, we concatenate the word embeddings with the hand-crafted features at the input and then feed this combined input to the network for learning. In comparison, architecture A3 learns the sentence embedding through Bi-LSTM network on top of word embedding only, which is then merged with the hand-crafted features before feeding into the fully connected layers for prediction. In contrast, architecture A4 utilizes two separate Bi-LSTM networks for word embeddings and hand-crafted features, respectively. Subsequently, the learned sequences of each Bi-LSTM are concatenated and fed into the fully-connected layers for further prediction. The choice of separate Bi-LSTMs for the hand-crafted features in architecture A4 is driven by the fact that the dimension of a word embedding is usually very high as compared to its corresponding hand-crafted features. If trained together, as in architecture A2, extracted features of low dimension usually get overshadowed by the high-dimensional word embeddings. Thus making it non-trivial for the network to learn from the extracted features. Further, to exploit the sequence information of words in a sentence, we pass hand-crafted features of each word through a separate Bi-LSTM layer. E.g. in the following sentence there is one negative word (i.e. *horrible*) and one negation (i.e. *not*) but no positive words. However, in a model that takes into account only the simple polar word score, the sentence would have high relevance towards the negative sentiment. However, the sequence information of the phrase "*not any more*" dictates the positive sentiment of the sentence.

"*It was used to be a horrible place to eat but not any more.*"

In contrast to A4, architecture A3 does not rely on the sequence information of the extracted features and allows the network to learn on its own. We use 300 dimension Word2Vec (Mikolov et al., 2013) word embeddings for the experiments. Each Bi-LSTM layer contains 100 neurons while two dense layers contain 100 and 50 neurons, respectively.

Features

As additional features, we extract the following information for each token in an instance.

– Aspect term extraction: Distributional thesaurus (DT)[1] (Biemann and Riedl, 2013) defines the lexicon expansion of a token based on a similar context. It is usually very effective for the handling of unseen text. If a token in the test set never appears in the training set, it becomes a non-trivial task for the classifier to make a correct prediction. By employing DT feature, the classifier can additionally utilize lexical expansion of the current token for mapping with the training set, thus minimize the chance of unseen text. For each token, we use its top 3 DT expansions as features.

Language	Train					Test				
	#sent.	#aspects	pos	neg	neu	#sent.	#aspects	pos	neg	neu
English	2,000	2,507	1,657	749	101	676	859	611	204	44
Spanish	2,070	2,720	1,925	674	120	881	1,072	750	274	48
French	1,733	2,530	1,164	1,212	154	696	954	441	434	79
Dutch	1,711	1,860	1,062	646	152	575	613	369	211	33
German	19,432	19,432	1,179	5,045	13,208	2,566	2,566	105	780	1,681
Hindi	5,417	4,469	1,986	569	1,914	10-fold cross validation				

Table 2: Dataset statistics

– Aspect sentiment classification: We employ publicly available lexicons of Chen and Skiena (2014) for extracting the polar information of each token. It contains a list of positive and negative words for 136 different languages. Additionally, we append the positive and negative words of 4 well-known

[1] http://ltmaggie.informatik.uni-hamburg.de/jobimtext/documentation/calculate-a-distributional-thesaurus-dt/

Datasets	Aspect Extraction (F1-score)				Aspect Classification (Acc)			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
English	62.0	63.1	62.4	**64.9***	82.4	82.7	82.1	**83.4**
Spanish	72.0	71.8	72.4	**73.0***	86.4	86.3	86.1	**87.1***
French	67.1	**67.8***	63.6	64.9	75.0	**75.3***	75.2	74.3
Dutch	65.2	65.6	**65.7**$^{+}$	64.2	80.9	80.7	**81.9***	81.4
German	23.1	22.0	22.4	**24.0***	86.7	**87.2***	86.6	**87.2***
Hindi	50.0	49.3	50.4	**53.5***	64.5	66.3	65.8	**66.9***

Table 3: Comparison of various models for aspect extraction and aspect classification on test dataset. *A1*, *A2*, *A3* & *A4* refers to four architectures depicted in Figure 1. *Statistically significant (*T*-test) *w.r.t.* other architectures (*p-values*< 0.05). $^{+}$Significant *w.r.t.* A4.

lexicons of English language (Bing Liu opinion lexicon, Ding et al. 2008; MPQA subjectivity lexicon, Wilson et al. 2005; SentiWordNet, Baccianella et al. 2010; and Vader sentiment, Hutto and Gilbert 2014) through the application of Google Translator. For German, we additionally use GermanPolarityClues lexical resource (Waltinger, 2010). The final list contains 2757, 2164, 3271, 1615, 17627 and 11874 positive words for *English*, *Spanish*, *Dutch*, *French*, *German* and *Hindi*, respectively. Similarly, there are 5112, 1735, 5834, 3038, 19962 and 2225 negative words in the list.

4 Experiments, Results and Analysis

4.1 Datasets

We evaluate our proposed approach on the benchmark datasets of SemEval-2016 shared task on aspect based sentiment analysis (Pontiki et al., 2016) (Task 5), which contain user reviews across multiple languages. The datasets of English, Spanish, French and Dutch are related to the reviews of consumer electronics and restaurants. We also evaluate our approach on the GermEval-2017 shared task on ABSA (Wojatzki et al., 2017), which comprises of reviews in the German language. The training datasets contain 2,070, 1,733, 1,711 & 19,432 reviews in Spanish, French, Dutch and German, respectively. Whereas, test datasets contain 881, 696, 575 & 2,566 reviews for the respective languages. For Hindi, we employed ABSA dataset developed by Akhtar et al. (Akhtar et al., 2016). There are total 4469 aspect terms in 5417 sentences across 12 domains. We perform 10-fold cross validation for the evaluation in this work. Table 2 lists the brief statistics of the various datasets for different languages.

4.2 Preprocessing

We extract each instance from the SemEval and the GermEval dataset to take into account only the relevant information and remove the XML tags. We use NLTK[2] (Shallow parser[3] for Hindi) to tokenize each sentence of the dataset. The aspect terms can span over multiple words in a sentence and hence, we use the BIO encoding scheme. In this notation, B, I and O denote the beginning, internal and outside tokens of aspect term respectively.

4.3 Results

We use Python based deep learning library Keras [4] with Tensorflow[5] for implementing the systems. The weight matrices were initialized randomly using numbers from a truncated normal distribution. Model is

[2] https://www.nltk.org/

[3] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

[4] https://keras.io/

[5] https://www.tensorflow.org/

trained with 32 batch size and 0.25 *Dropout* (Srivastava et al., 2014) with *Adam* (Kingma and Ba, 2014) optimizer. We employ *Relu* (Glorot et al., 2011) as activation function for the hidden layers, whereas for the output layer we use *softmax* classifier. Following the guidelines of SemEval-2016 (Pontiki et al., 2016) and GermEval-2017 (Wojatzki et al., 2017), we employ F1-score as the evaluation metric for aspect term extraction. For classification, we compute accuracy and F1-score for SemEval-2016 and GermEval-2017, respectively. Similarly, we adopt F1-score and accuracy for the aspect term extraction and aspect sentiment classification in Hindi. In Table 3, we present the results of all the four architectures

Systems	**Aspect Extraction (F1-score)**						**Aspect Classification (Accuracy)**					
	En	Es	Fr	Du	De	Hi	En	Es	Fr	Du	De*	Hi
State-of-the-art systems at SemEval-2016 (Pontiki et al., 2016)												
Baseline (Pontiki et al., 2016)	44.0	51.9	45.4	50.6	-	-	76.4	77.7	67.4	69.3	-	-
NLANGP (Toh and Su, 2016)	**72.3**†	-	-	-	-	-	-	-	-	-	-	-
GTI (Alvarez-López et al., 2016)	66.5	68.5†	-	-	-	-	69.9	-	-	-	-	-
IIT-TUDA (Kumar et al., 2016)	42.6	64.3	66.6†	56.9†	-	-	86.7	83.5†	72.2	76.9	-	-
XRCE (Brun et al., 2016)	61.98	-	65.3	-	-	-	**88.1**†	-	**78.8**†	-	-	-
TGB (Çetin et al., 2016)	55.0	55.7	-	51.7	-	-	80.9	82.0	-	77.8†	-	-
State-of-the-art systems at GermEval-2017 (Wojatzki et al., 2017)												
Baseline (Wojatzki et al., 2017)	-	-	-	-	17.0	-	-	-	-	-	48.1*	-
System (Mishra et al., 2017)	-	-	-	-	22.0	-	-	-	-	-	42.1*	-
System (Ji-Ung Lee and Gurevych, 2017)	-	-	-	-	20.3	-	-	-	-	-	48.2*	-
State-of-the-art systems for Hindi (Akhtar et al., 2016)												
System (Akhtar et al., 2016)	-	-	-	-	-	41.0	-	-	-	-	-	54.0
System (Akhtar et al., 2016)	-	-	-	-	-	-	-	-	-	-	-	65.9
Proposed Approach	64.9	**73.0**	**67.8**	**65.7**	**24.0**	**53.5**	83.4	**87.1**	75.3	**81.9**	**87.2***	**66.9**
Architecture	A4	**A4**	**A2**	**A3**	**A4**	**A4**	A4	**A4**	A2	**A3**	**A4**	**A4**

Table 4: Comparison with the state-of-the-art systems of SemEval-2016 and GermEval-2017. *F1-score. Official evaluation metric for aspect classification at GermEval-2017 was F1-score. †Best system for respective language-problem pair.

for each language/problem pair. In aspect extraction problem, architecture A4 yields the best F1-score for *Spanish* (73.0%), *German* (24.0%), *English* (64.9%) and *Hindi* (53.5%), whereas for *French* and *Dutch* we obtain the best F1-score with architectures A2 (67.8%) and A3 (65.7%), respectively. We observe similar trends for aspect classification as well with architecture A4 performing better for *Spanish* (87.2% accuracy), *German* (87.2% F1-score), *English* (83.4% accuracy) and *Hindi* (66.9% accuracy). Similar to aspect extraction, architectures A2 and A3 report better performance for *French* (75.34%) and *Dutch* (81.9%), respectively. Among all four architectures, architecture A1 has the least performance across all six languages for both the problems. It suggests that the hand-crafted features -when fused into the network- assist the system to learn in a better way than the system learnt with only word embeddings. We also perform statistical significance test (T-test) on the obtained results and observe that the performance of the architecture A4 is significant with 95% confidence for *English*, *Spanish*, *German* and *Hindi* for both the problems.

Further, we compare our proposed system with state-of-the-art systems as listed in Table 4. Our proposed system shows an improvement over the existing state-of-the-art for 9 out of 12 language/problem pairs. For aspect extraction, the system achieves an improvement of 4.5, 1.2, 8.8, 2 and 12.5 points for *Spanish*, *French*, *Dutch*, *German* and *Hindi*, respectively. Our system manages to improve the score of sentiment classification for *Spanish*, *Dutch*, *German*, and *Hindi* by 3.56, 4.17, 12.3 and 1 points, respectively. Improvement of the system performance across the language/problem pairs suggests about the generic nature of our proposed approach. Also, significance *T-test* shows that improvement of the proposed method over the state-of-the-art systems are statistically significant with p-values< 0.05.

From Table 3, we observe that architecture A4 performs the best for four languages, i.e., *Spanish*, *German*, *English* and *Hindi* irrespective of the problems. Similarly, the performance of the architectures

A2 & A3 is best for *French* and *Dutch*, respectively. Since architecture A4 is the clear winner in 8 out of 12 language/problem pairs and also reports comparable performance in other cases - with maximum 2.9 points below the best architecture as reported in Table 3 -, we recommend it as the default choice for all the languages and problems.

4.4 Error Analysis

We perform error analysis on the predicted outputs, using automatic translations (Google) for languages we are not proficient in. Following are the few cases where our proposed system often faces challenges.

Aspect term extraction: Aspect term extraction is a quite challenging task. The BIO notation is an effective solution for tagging an aspect term; however, it is highly skewed towards the *O* class, i.e., only a small percentage of tokens in the vocabulary qualify for the aspect term. Despite this limitation, BIO notations result in decent outputs with the few exceptions. In Table 5, we list a few common error patterns along with the examples. Our system faced difficulties when one or more terms can independently qualify as an aspect term. In the first two examples, our system misclassifies the multi-token aspect terms '*customer service*' and '*atencin del personal*' (attention of the staff) as single aspect terms. It predicts the first token of the aspect term (i.e., '*customer*' (first example) and '*atencin*' (attention) (second example)) as one aspect term and the last token (i.e., '*service*' and '*personal*' (staff)) as the other aspect term. Despite both the tokens of aspect term '*customer service*' is identified as aspect terms, it results in *recall=0* and *precision=0*.

Table 5: Common error pattern for aspect term extraction.

Language	Review	Gold Aspect Terms	Predicted Aspect Terms	Possible Reason
Source (EN)	*Best restaurant in the world, great decor, great customer service, friendly manager*	*restaurant, decor,* ***customer service****, manager*	*restaurant, decor,* ***customer****,* ***service****, manager, pizza*	Individual tokens in a multi-token aspect term qualify for aspect terms .
Source (ES) Translation (EN)	*La atención del personal impecable.* *Attention of the staff was impeccable.*	***atención del personal*** ***Attention of the staff***	***atención****,* ***personal*** ***Attention****,* ***staff***	
Source (EN)	*I had yummy lamb korma, saag paneer, samosas, naan, etc.*	*lamb korma, saag paneer, samosas, naan*	*lamb korma*	Sequence of dishes (rare occurrence)
Source (FR) Translation (EN)	*Ravioles et tartiflette correctes, crłpe suzette passable.* *Ravioles and tartiflette correct, crepe suzette passable.*	*Ravioles, tartiflette, crłpe suzette* *Ravioles, tartiflette, crepe suzette*	*Ravioles* *Ravioles*	
Source (FR) Translation (EN)	*...le riz arborio aux truffes apparaissant dans le menu...* *...the arborio rice with truffles appearing in the menu...*	***riz arborio aux truffes*** ***arborio rice with truffles***	***riz arborio*** ***arborio rice***	Presence of subordinating conjunction in between an aspect term.
Source (EN)	*Great draft and bottle selection and the pizza rocks.*	***draft and bottle selection****, pizza*	***bottle selection****, pizza*	

In the third and fourth examples of Table 5, a number of dishes which are served in the restaurant are mentioned. For both examples, our system manages to identify only some dishes. A possible reason would be the rare occurrence of these dishes in the training set. The last two examples suffer from the presence of subordinating conjunctions (i.e. 'and', 'with' etc.) in the multi-token aspect terms (i.e. '*riz arborio aux truffles*' (arborio rice with truffles)). In general, 'and', 'with' or other conjunctions does not qualify for the aspect term except in the company of multi-token aspect terms. However, such occurrences are not very common, and the underlying system misclassifies them as outside aspect term, i.e., *O*. The second example (i.e. '*atención del personal*' (attention of the staff)) may also qualify for the similar reason.

Aspect sentiment classification: For aspect sentiment classification, we observed two most common sources of errors across languages, i.e., lack of polar information inside the defined context window (± 5 neighbouring words) and presence of the sarcastic or metaphoric phrase in the review. We list a few error cases in Table 6. The first example belongs to the Spanish language, which contains an aspect term '*calidad-precio*' (*quality-price*). The actual sentiment towards the aspect term is *positive*; however, in the absence of clue words (i.e. '*restaurantes de referencia de Zaragoza*' (*recommended restaurants of*

Table 6: Common error pattern for aspect sentiment classification.

Language	Review	Aspect Term	Actual Sentiment	Predicted Sentiment	Possible Reason
Source (ES)	*En lo referente a **calidad-precio** y dentro de su categora, desde mi punto de vista, debe ser uno de los restaurantes de referencia de Zaragoza.*	*calidad-precio*	*Positive*	*Neutral*	Lack of polar information inside context window
Translation (EN)	*Regarding **quality-price** and within its category, from my point of view, it must be one of the recommended restaurants of Zaragoza.*	*quality-price*			
Source (EN)	*Finally, my wife stood face to face in front of one of the **staff** and she asked, Are you waiting for a table?".*	*staff*	*Negative*	*Positive*	Sarcasm
Source (EN)	*The **lemon chicken** tasted like sticky sweet donuts.*	*lemon chicken*	*Negative*	*Positive*	Metaphor

Zaragoza)) inside the context window, our proposed system predicts its sentiment as *neutral*.

Predicting sentiment for the sarcastic and metaphoric text are usually challenging due to the difference in its *textual-meaning* and *actual-meaning* (i.e., what is said is not meant or vice-versa). Our system also finds it non-trivial to correctly classify an aspect term in the presence of sarcastic (second example of Table6) or metaphoric (third example) text. In the second example, the staff's unresponsiveness behaviour irked the writer, who had to ask for a table sarcastically. Similarly, in the third example writer was not amused by the quality of *lemon chicken* and compared it with the *sticky sweet donuts* as figure-of-speech.

5 Conclusion

In this paper, we have proposed a language-agnostic deep neural network approach for solving the problems of aspect-based sentiment analysis. Our system employs Bi-LSTM network for learning the sentence embeddings, which is assisted by a few handcrafted features. To show the effectiveness, we evaluated the proposed approach on six languages (i.e. *English*, *Spanish*, *French*, *Dutch*, *German* and *Hindi*) and two problems (i.e. *aspect term extraction* and *aspect sentiment classification*). We also evaluated different ensemble architectures to combine sentence embeddings and handcrafted features. Comparisons with the existing system suggest that our proposed approach attains the state-of-the-art performance for almost each of the language/problem pair.

6 Acknowledgement

Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

Akhtar, M. S., A. Ekbal, and P. Bhattacharyya (2016). Aspect based Sentiment Analysis in Hindi: Resource Creation and Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 23-28, 2016*, Portoro, Slovenia, pp. 2703–2709. European Language Resources Association (ELRA).

Akhtar, M. S., A. Kumar, A. Ekbal, and P. Bhattacharyya (2016). A Hybrid Deep Learning Architecture for Sentiment Analysis. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers, December 11-16, 2016*, Osaka, Japan, pp. 482–493.

Alvarez-López, T., J. Juncal-Martinez, M. Fernández-Gavilanes, E. Costa-Montenegro, and F. J. González-Castano (2016). GTI at SemEval-2016 Task 5: SVM and CRF for Aspect Detection and

Unsupervised Aspect-based Aentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, CA, USA, pp. 306–311.

Araque, O., I. Corcuera-Platas, J. F. Snchez-Rada, and C. A. Iglesias (2017, July). Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications. *Expert Syst. Appl. 77*(C), 236–246.

Baccianella, S., A. Esuli, and F. Sebastiani (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, pp. 2200–2204.

Balamurali, A. R., A. Joshi, and P. Bhattacharyya (2012). Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, pp. 73–82.

Biemann, C. and M. Riedl (2013). From Global to Local Similarities: A Graph-Based Contextualization Method using Distributional Thesauri. In *Proceedings of the 8th Workshop on TextGraphs in conjunction with Empirical Methods on Natural Language Processing*, Seattle, WA, USA, pp. 39–43.

Brun, C., J. Perez, and C. Roux (2016). XRCE at SemEval-2016 Task 5: Feedbacked Ensemble Modeling on Syntactico-Semantic Knowledge for Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, CA, USA, pp. 277–281.

Çetin, F. S., E. Yıldırım, C. Özbey, and G. Eryiğit (2016). TGB at SemEval-2016 Task 5: Multi-Lingual Constraint System for Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, CA, USA, pp. 337–341.

Chen, Y. and S. Skiena (2014). Building Sentiment Lexicons for All Major Languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, MD, USA, pp. 383–389.

Ding, X., B. Liu, and P. S. Yu (2008). A Holistic Lexicon-Based Approach to Opinion Mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, Stanford, CA, USA, pp. 231–240.

Ghosal, D., M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya (2018, October-November). Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 3454–3466. Association for Computational Linguistics.

Ghosh, A., G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes (2015). Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 470–478.

Glorot, X., A. Bordes, and Y. Bengio (2011). Deep Sparse Rectifier Neural Networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, Ft. Lauderdale, FL, USA, pp. 315–323.

Graves, A., S. Fernández, and J. Schmidhuber (2005). Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. In *Proceedings of the 15th International Conference on Artificial Neural Networks: Formal Models and Their Applications - Volume Part II*, Warsaw, Poland, pp. 799–804.

Hutto, C. J. and E. Gilbert (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM-14)*, Ann Arbor, MI, USA, pp. 216–225.

Jagtap, V. and K. Pawar (2013). Analysis of Different Approaches to Sentence-level Sentiment Classification. *International Journal of Scientific Engineering and Technology (ISSN: 2277-1581) Volume 2*, 164–170.

Ji-Ung Lee, Steffen Eger, J. D. and I. Gurevych (2017). UKP TU-DA at GermEval 2017: Deep Learning for Aspect Based Sentiment Detection. In *Proceedings of the GermEval 2017 Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, Germany, pp. 22–29.

Kim, S. and E. Hovy (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, Sydney, Australia, pp. 1–8.

Kingma, D. and J. Ba (2014, 12). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, Vancouver, BC, Canada.

Kiritchenko, S., X. Zhu, C. Cherry, and S. Mohammad (2014). NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 437–442. Association for Computational Linguistics and Dublin City University.

Kumar, A., S. Kohail, A. Kumar, A. Ekbal, and C. Biemann (2016). IIT-TUDA at SemEval-2016 Task 5: Beyond Sentiment Lexicon: Combining Domain Dependency and Distributional Semantics Features for Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, CA, USA, pp. 1129–1135.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, Lake Tahoe, NV, USA, pp. 3111–3119.

Mishra, P., V. Mujadia, and S. Lanka (2017). GermEval 2017 : Sequence based Models for Customer Feedback Analysis. In *Proceedings of the GermEval 2017 Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, Germany, pp. 36–42.

Mohammad, S. M., S. Kiritchenko, and X. Zhu (2013, June). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.

Naderalvojoud, B., B. Qasemizadeh, and L. Kallmeyer (2017). HU-HHU at GermEval-2017 Sub-task B: Lexicon-Based Deep Learning for Contextual Sentiment Analysis. In *Proceedings of the GermEval 2017 Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, Germany, pp. 18–21.

Pang, B., , and L. Lee (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pp. 115–124.

Pang, B. and L. Lee (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval 2*(1-2), 1–135.

Pang, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86. Association for Computational Linguistics.

Pontiki, M., D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A.-S. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, et al. (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, CA, USA, pp. 19–30.

Pontiki, M., D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar (2014, August). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 27–35.

Poria, S., E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency (2017). Multi-level multiple attentions for contextual multimodal sentiment analysis. In *Data Mining (ICDM), 2017 IEEE International Conference on*, pp. 1033–1038. IEEE.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research 15*, 1929–1958.

Toh, Z. and J. Su (2016, June). Nlangp at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, pp. 282–288. Association for Computational Linguistics.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th ACL*, pp. 417–424.

Waltinger, U. (2010, May). GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, pp. 1638–1642.

Wilson, T., J. Wiebe, and P. Hoffmann (2005). Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, BC, Canada, pp. 347–354.

Wojatzki, M., E. Ruppert, S. Holschneider, T. Zesch, and C. Biemann (2017). GermEval 2017: Shared Task on Aspect-based Sentiment in Social Media Customer Feedback. In *Proceedings of the GermEval 2017 Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, Germany, pp. 1–12.

Zadeh, A., P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency (2018). Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, pp. 2236–2246.