Aspect Category Sentiment Analysis on a Challenging Dataset

Kiran Negi
School of Computing,
Dublin City University
Dublin, Ireland
kiran.negi2@mail.dcu.ie

Saumitra Das
School of Computing,
Dublin City University
Dublin, Ireland
saumitra.das2@mail.dcu.ie

***Abstract*—The advanced world with the internet makes everyone enable to provide his/her opinion, and a review depicts the sentiment of a user towards various aspects of a product or service. Aspect-Category Sentiment Analysis (ACSA) aims to predict the pre-defined aspect categories present in a review and their respective sentiment polarities. Most sentences in current ACSA datasets contain only one category or multiple categories of the same sentiment polarity, which causes ACSA tasks to deteriorate into sentence-level analysis of sentiment. In this paper, we have performed experiments on a large-scale Multi-Aspect Multi-Sentiment (MAMS) dataset, where each sentence consists of at least two different aspect categories with different sentiment polarities. As traditional word representations of the sentences will not produce good results on such a dataset, we have experimented with CNN and BERT classifiers using context-independent and context-dependent word embeddings respectively. They have achieved competitive results on the MAMS dataset as well as SemEval-2014 Restaurant Review Dataset for the Aspect Category Detection (ACD) task. The BERT classifier outperforms the best submission of SemEval-2014 for ACD, and it proves that the use of contextual word embedding is the most effective method for this task. We also presented an attention-based Long Short-Term Memory (LSTM) Network for Aspect Category Sentiment Classification (ACSC). The mechanism concentrates on multiple parts of a given sentence when multiple aspects are used as input. Using LSTM architecture, we observed that assigning weights to the concerned aspects helps in utilizing all the elements present in a sentence. As part of our paper, we try to demonstrate adequate performance of the LSTM mechanism for ACSC.**

***Index Terms*—Aspect Category Sentiment Analysis (ACSA), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Bidirectional Encoder Representations from Transformers (BERT), Long Short-Term Memory (LSTM)**

I. Introduction

Natural Language Processing (NLP) is a field of Artificial Intelligence which is primarily used to decipher the human language, and sentiment analysis is a field within NLP. Sentiment analysis helps in identifying the sentiment of a sentence and often classifies them as positive, negative and neutral. For example, Sentiment analysis will classify the sentence *"The food is delicious."* as positive and *"The mango shake is horrible."* as negative. It has become popular in the field of research and business as people are now able to express their thoughts more openly than ever with the in-hand access of internet and social media platforms. Understanding customers' feelings towards a specific product, brand or commercial establishment are essential for companies and organizations. And Sentiment analysis serves the purpose of understanding sentiments by analyzing the reviews and comments of the users.

Traditional sentiment analysis classifies the overall sentiment of a text and does not specify what the sentiment is about. So, traditional sentiment analysis will not work well on the review *"The food is delicious, but the mango shake is horrible."* as the user is expressing different sentiments towards different subjects (also known as aspects). In order to predict sentiments associated with specific aspects in a text, a fine-grained Aspect Based Sentiment Analysis (ABSA)

is proposed [1] to understand reviews better than traditional sentiment analysis.

ABSA comprises two sub-tasks: Aspect Category Sentiment Analysis (ACSA) and Aspect Term Sentiment Analysis (ATSA). Aspect-category is a group of similar attributes of an entity, whereas aspect-term is a word or phrase that appears in the text explicitly and represents an aspect-category [2]. For example, in the text *"The mango shake is horrible."*, *"mango shake"* is an aspect-term which indicates the aspect-category *"food"*.

ACSA is an underexploited subtask of ABSA. There are two subtasks of ACSA. The first is to identify the aspect category discussed in a review from a pre-defined set of aspect categories, known as Aspect Category Detection (ACD). The critical challenge in this subtask stems from the fact that in most cases, the aspect category may not appear in the sentence explicitly. For example, the review *"The restaurant serves delicious sushi."* speaks on the aspect category *"food"* in an implicit manner. The second is to identify the sentiment polarity (positive, negative, neutral) of each aspect category, known as Aspect Category Sentiment Classification (ACSC). As for the previous example, the sentiment polarity is positive for the aspect category *"food"*.

ACSC came into limelight in the SemEval-2014 [3] workshop where it was first introduced as Task 4 and reappeared in the SemEval-2015 [4] and SemEval-2016 [5] competitions. The SemEval Restaurant and Laptop Review datasets are considered as the benchmark datasets for the ACSA task, as other significant datasets only focus on the Aspect term Sentiment Analysis (ATSA) task [3][4][5]. However, most sentences in the SemEval datasets consist of only one aspect category or multiple aspect categories with the same polarity. For example, only 454 instances have multiple aspect categories with different sentiment polarities out of 4738 instances in the SemEval2014 Restaurant Review Dataset. Due to this structure of the dataset, ACSA degenerates to sentence-level sentiment analysis as most of the reviews contain the same polarity towards a single or multiple aspect categories. Furthermore, a sentence-level sentiment classifier can achieve good results on that dataset without considering the aspect categories [6], making ACD and ACSC tasks easy.

Multi-Aspect Multi-Sentiment (MAMS) dataset is introduced to advance and facilitate research in the field of aspect-based sentiment analysis. Each sentence in MAMS dataset consists of at least two different aspect categories with different sentiment polarities. As a simple sentence-level classifier will fail to achieve good results on MAMS dataset, this dataset is considered more challenging than the SemEval-2014 Restaurant review dataset [6].

Deep learning-based approaches have exhibited competitive results for text classification [7], and recently, pre-trained language model BERT has achieved state-of-the-art results on many text classification datasets [8][9]. There are no submissions at any SemEval workshops that have used BERT for the ABSA tasks. It is also well-established that the use of word embedding models is very efficient in representing data for NLP tasks [7]. In this paper, we have experimented with both contextual word embeddings using transformer-based model BERT and context-independent ones (Glove and Word2Vec) trained using deep learning-based model CNN, for ACD. Experimental results show that neural word embeddings are very efficient and have produced excellent results. And the contextual language model has outperformed the context-independent ones on MAMS as well as SemEval 2014 Restaurant Review Dataset, and achieved better results than the top SemEval-2014 workshop submissions for ACD.

Moreover, deep learning neural networks have accomplished impeccable performance in NLP tasks such as paraphrase identification [10], question answering [11], machine translation [12], and text summarization [13]. In few of the works, Target dependent LSTM (TD-LSTM) is beneficial for aspect based sentiment classification [14], but these target based models cannot take aspect level information into consideration.

Application of Attention is an effective method to achieve the state of the art results as described in machine translation [15], sentence summarization [13]. For the ACSC task, we presented an attention mechanism to demonstrate a model to take care of the critical part of a given sentence corresponding to an aspect. In this paper, we investigate the correlation between an aspect and its respective polarity.

The paper demonstrates the attention-based aspect category sentiment classification. The mechanism can focus on key parts of a sentence/review when multiple aspect categories are involved. Our experimental results indicate that the approach improves the output results as compared to existing methods, and thus dictates it works well for ACSC.

II. RELATED WORK

A. Aspect Category Detection

- *Classical Methods*: Previous approaches to ACD framed the task as a multi-class classification problem. They relied on Conditional Random Fields (CRF) classifier which leverages various features such as Name Entity Recognition (NER), POS tagging, parsing, semantic analysis, as well as domain-dependent lexicons trained on reviews from YELP and Amazon data. SVM, binary MaxEnt and logistic regression models are also used for individual aspect categories with various types of n-gram, BoW and TF-IDF features [3][4]. The top scorer of SemEval-2014 workshop used one multi-class SVM classifier for all the categories [3].

  The old-fashioned Bag-of-Words (BOW) strategy was extensively used for data representation in the SemEval workshops [3][4][5]. This approach builds a vocabulary from the given dataset, where each word in the vocabulary becomes a feature. TF-IDF is mostly used for the feature generation of words in the vocabulary.
- *Word Embedding Methods*: Some approaches to ACD formulated the task as a multi-label classification problem. They depended on word embeddings for data representation and modelled them using CNN and logistic regression classifiers [16] [17].

Pre-trained word embeddings are the embeddings trained on large datasets, saved and used for other tasks. They capture the semantic and syntactic meaning of a word as they are already trained on large datasets. The pre-trained word embedding provides one only vector for each word that encodes the meaning of the word. Most word embedding techniques rely on feed-forward neural network architecture for the learning process [18].

- *Pre-trained Language Models*: Bert is the first deeply bidirectional and unsupervised language representation model which focuses on contextualized word representations. It uses transformers that are attention-based, to get the word context features from a text by jointly conditioning on both left-to-right and right-to-left context. Therefore, this model provides different vectors for the same word depending on the context of the word [8].
- *Our Approach*: We address the ACD task with one-vs-all classifiers, i.e. creating one classifier for each category that we are trying to extract from the data. Initially, we have experimented with classical SVM model using POS tagging, CountVectorizer and TF-IDF features. The SVM model performed poorly with all those features on the MAMS Dataset and achieved F1-scores of less than 60. It proves that traditional word vector representation of features will not work efficiently on the MAMS dataset and so in this paper, we didn't discuss SVM implementations in detail. Subsequently, we switched our approach to a feed-forward deep-learning model and experimented with pre-trained word embedding models which are context-independent. We have used a simple 1D CNN classifier which achieved competitive results on the MAMS as well as SemEval-2014 Restaurant Review dataset. Finally, we used MobileBERT [19], which is a thin version of BERT, for the contextual representation of the data and category classification. This model outperformed all the other models for the ACD task.

B. Aspect Category Sentiment Classification

Sentiment classification at aspect level is taken as a generic classification problem in theory. As mentioned earlier, instead of a generic classification problem, ACSC is a fine-grained classification task. The traditional mechanisms focus on the polarity of a given sentence as a whole and ignore the number of aspects or entities present in a sentence. In classical methods, they use to manually build a design for a set of features. The lexicon based features are built due to the presence of a number of sentiment lexicons [20]. These experiments involve sentiment lexicons and BOW vectors. The resultant output of such mechanisms are highly dependent on the feature quality, also building a set of features is a tedious task.

Sentiment classification with Neural networks has advanced as compared to feature engineering methods. Few of the classical approaches in the neural networks are Recursive Neural Network (RNN) [21]. There is also one quite effective method known as Tree-based LSTM for many NLP tasks but this method faced issues with syntax parsing, a common issue in language usage.

Target oriented mechanisms such as TD-LSTM and TC-LSTM [14] have gained better performance in sentiment classification which are target dependent. TC-LSTM uses a target vector which averages the word vectors of the containing target phrase. But this is not an adequate approach to demonstrate the semantics of the target which results in less good performance.

In the presence of these challenges in the traditional methods, it remained a challenge to distinguish multiple polarities at aspect level. Hence, as part of this paper we try to develop a model that focuses on the aspect information for sentiment classification.

TABLE I
STATISTICS OF MAMS DATASET

	Positive		Negative		Neutral		Total	
Category	Train	Test	Train	Test	Train	Test	Train	Test
Food	754	184	255	59	1298	338	2307	581
Price	72	18	114	32	136	33	322	83
Service	174	55	329	73	128	34	631	162
Ambience	181	34	90	31	53	3	324	68
Miscellaneous	237	59	196	45	531	161	964	265
Menu	64	19	39	7	372	101	475	127
Place	125	37	139	43	430	89	694	169
Staff	322	80	922	232	129	22	1373	334
Total	1929	486	2084	522	3077	781	7090	1789

III. DATASETS

MAMS dataset is created by Jiang et al [6] with the help of three NLP researchers for ACSA. An instance of MAMS dataset is shown in Figure 1. The eight pre-defined aspect categories present in the MAMS dataset are: *food, service, staff, price, ambience, menu, place and miscellaneous*. Five of the aspect categories: *food, price, service, ambience, and anecdotes/miscellaneous* are adopted from SemEval-2014 Restaurant review dataset, and three more are added to avoid confusion between categories. Only the sentences which consist of at least two unique categories with different sentiment polarities are present in the dataset.

The statistics of the MAMS dataset is presented in Table I. A total of 8879 reviews are present in the MAMS dataset, which is 1.87 times of SemEval-2014 Restaurant Review dataset. All reviews contain multiple categories with different sentiment polarities. MAMS dataset is more challenging than other datasets as MAMS reviews can not be classified correctly without modelling the relationship between contexts and aspects [6]. For comparison, we have also evaluated our models on the SemEval-2014 Restaurant Review dataset.

IV. MODEL

A. CNN for ACD

The model architecture we use is a slight variant of the CNN structure used by Collobert et al [22], successfully applied by many others [7] for text classification and achieved

```
<sentence>
    <text>The bartender was skilled, the owners were very friendly, but the wait for my burrito was longer than I would have liked.</text>
    <aspectCategories>
        <aspectCategory category="staff" polarity="positive"/>
        <aspectCategory category="service" polarity="negative"/>
        <aspectCategory category="food" polarity="neutral"/>
    </aspectCategories>
</sentence>
```

Fig. 1. Example sentence with aspect categories and sentiment annotations for MAMS dataset

state-of-the-art performance on many benchmark sentiment classification datasets [23].

The CNN model is constructed with an embedding layer, a one-dimensional convolutional layer and a max-pooling layer. Pre-trained word vector embeddings represent the embedding layer, and they are fine-tuned during the training stage. We have experimented with Glove [24] and word2vec [25] as an embedding layer separately. The embedding layer takes the input indices of the words such as $w_i \varepsilon \{1, 2, \ldots, V\}$ and generates the embedding word vectors $v_i \varepsilon R^D$, where *D* represents the dimension size of the embedding vectors and *V* denotes the size of the word vocabulary. Therefore, the input text is denoted by a matrix $X = [v_1, v_2, ..., v_L]$ produced by the embedding layer which is padded to length *L*. The one-dimensional convolutional layer acts as a filter $W_c \epsilon R^{D*k}$ applied to a window of size *k* to produce a new feature *c*. For a whole text, a feature map

$$c = [c_1, c_2, ..., c_{L-K+1}] \tag{1}$$

is generated as we slide the filter window. The general equation for a feature map is given by

$$c_i = f(X_{i:i+K-1} * W_c + b_c) \tag{2}$$

where $b_c \ \varepsilon R$ is the bias term, *f* is a non-linear activation such as Rectified Linear Unit (ReLU) for our model and * denotes the convolution operation. Max-over-time pooling layer takes the maximal value among the generated feature maps and thereby condensing the feature maps to its most essential feature. Finally, a sigmoid layer uses the feature map to predict the category classification of the input sentence for each category.

The context-independent word embeddings Glove and Word2Vec do not take into account the order of the words in a review while training. These models combine all the different senses of a word into one vector. For example, in the sentence *"The restaurant is located pretty far from the flat that I am renting, but I love the flat noodles they serve."*, the word *flat* has different meanings based on the sentence context. Still, these models will collapse them into one vector for flat in their output.

B. BERT for ACD

The MobileBERT model architecture [19] for text classification, is used for the ACD task. The model is equipped with bottleneck structures and a carefully designed balance between self-attentions and feed-forward networks. It contains a deep bidirectional encoder trained on large scale corpora: Wikipedia and the BookCorpus.

Bert pre-trained language model is based on the Transformer framework. The MobileBERT model is a slight variant of the BERT architecture. The encoder is built with 24 Transformer blocks (hidden layers), 4 self-attention heads, bottleneck size of 512, 4 feed-forward layers and a hidden size of 128. BERT takes a sequence input of no more than 512 tokens and outputs the sequence representation. The sequence has one or two segments where the sequence's first token is often [CLS] which includes the embedding of the special classification, and another unique token [SEP] is used for segment separations.

BERT takes the final hidden state *h* of the first token [CLS] as a representation of the entire sequence for text classification tasks. To estimate the likelihood of the label *c*, a simple softmax classifier is applied to the top of BERT such that

$$p(c|h) = softmax(Wh) \tag{3}$$

where *W* is the task-specific parameter matrix. We fine-tune BERT's parameters as well as *W* jointly by improving the log-probability of the correct label [8].

The BERT model is bi-directional that takes into account the order of the words in the text from left to right as well as right to left. It will generate different vectors for all different senses of a word. As for the example *"The restaurant is located pretty far from the flat that I am renting, but I love the flat noodles they serve."*, BERT model will generate two different vectors for the word *flat* as it represents two different contexts of the word.

C. Attention-based LSTM for ACSC

An extension of conventional neural network, RNN, has the well known gradient vanishing issue while back propagating the early layers. The LSTM approach was developed and later achieved better performance in the sentiment classification and various NLP tasks [26]. The LSTM architecture consists of a cell and three gates.

$$w_1, w_2, ..., w_N \tag{4}$$

represents word vectors for a sentence with length *N*. The hidden vector is represented by

$$h_1, h_2, ..., h_N \tag{5}$$

The computation corresponds to each cell can be done as follows:

$$X = \left(\frac{h_t - 1}{x_t}\right) \tag{6}$$

$$f_t = sigmoid(W_f X + b_f)) \tag{7}$$

$$i_t = sigmoid(W_i X + b_i) \tag{8}$$

$$o_t = sigmoid(W_o X + b_o) \tag{9}$$

$$c_t = f_t ct_1 + i_t tanh(W_c X + b_c) \tag{10}$$

$$h_t = o_t tanh(ct) \tag{11}$$

In above equations, W_i,W_f,W_o ε R^{d*2d} are the weighted matrices with corresponding biases represented as b_i,b_f,b_o ε R^d[27]. The biases are learnt during the training phase which then parameterized the input transformations. x_t includes the LSTM inputs which represents W_i: word embedding vector. The last hidden layer h_N is the sentence representation and it gets inputted to the SoftMax layer. The class labels in our study are: *positive, negative, neutral*.

To classify the polarity of a sentence, aspect information is critical. We may come across scenarios where we can get totally opposite polarities when considered all aspects. Thus, we propose an embedding vector for every aspect.

The traditional LSTM approach fails to detect critical parts of a sentence for sentiment level classification. To overcome this limitation, we attempt to design a mechanism with attention which captures the critical part of a sentence with respect to an aspect. The Attention based LSTM is explained as follows Assume H ε $R^{d*N}$is a matrix consisting of hidden vectors (h_1, h_2....h_N) produced by LSTM[27]. *'d'* is size of hidden layers and sentence length is *N*. v_a is embedding of an aspect and e_N ε R^N is a vector containing *1s*. The mechanism produces an attention weight vector alongwith *r* which is a weighted hidden representation

$$M = \left(tanh \frac{w_h * H}{W_v * v_a * e_N} \right) \tag{12}$$

$$\alpha = softmax(w^T * M) \tag{13}$$

$$r = H\alpha^T \tag{14}$$

where projection parameters are represented by M ε $R^{(d+da)N}$, α ε R^N, r ε R^d. W_h ε R^{d*d}, W_v ε R^{d*d} and w ε R^{d+d}[27]. A vector containing attention weights is a weighted representation with given aspect is represented by *r*. The final representation of a sentence [28] is as follows:

$$h* = tanh(W_p * r + W_x * h_N) \tag{15}$$

where W_p and W_x represents projection parameters, and $h^*\varepsilon$ R^d. The projection parameters are learned during training. h^* represents the features of a sentence. To make it more effective, we append aspect embedding to each input word vector. This mechanism hidden representations in output can have the information from (v_a) which is the input aspect. Hence it enables us to model the interdependence of input aspects and corresponding words.

V. METHODOLOGY

A. Aspect Category Detection

The objective of this sub-task is to identify all the aspect categories discussed in a review. Each restaurant review in MAMS Dataset has at least two different aspect categories present. The eight pre-defined aspect categories present in the MAMS dataset are: *food, service, staff, price, ambience, menu, place and miscellaneous*. And the five pre-defined aspect categories present in the SemEval-2014 Restaurant Review dataset are: *food, price, service, ambience, and anecdotes/miscellaneous*.

1) Preprocessing: We lower-case, remove stopwords and tokenize the MAMS and SemEval 2014 Restaurant Review Dataset for the ACD task.

2) Hyperparameters: We randomly split off 20 percent of the whole dataset and used it as a validation set.

- CNN: For training this classifier, we use the following parameters which are previously used by Kim [7]: batch size of 10, 15 epochs, word embedding size of 300 and a maximum sentence length of 100. Furthermore, we simplified our CNN model by using 64 feature maps and a kernel window of size 3 as these produce better results for the task. We use Adam optimizer [29] and binary-cross entropy loss function to train our models.
- Pre-trained Word Embeddings: We initialize the review texts by adding an embedding layer using two different pre-trained word embeddings separately: Glove and Word2Vec. Glove consists of 300-dimensional word vectors trained on 840 billion tokens of the Common Crawl corpus using a large vocabulary of 2.2 million words [24]. Google News Dataset is used for creating Word2Vec embeddings, which contains 300-dimensional vectors for 3 million words and phrases [25].
- BERT: We utilize the MobileBERT model [19] with a hidden size of 128, 4 self-attention heads and 24 transformer blocks. For fine-tuning this classifier, we use the following hyper-parameters which are suggested by Sun et al [9]: batch size of 32, max sequence length of 128, learning rate of 2e-5, warm-up proportion of 0.1, training epochs of 4, drop out probability of 0.1 and Adam optimizer.

3) Training Process: We address this subtask as a multi-class text classification problem. Eight binary one-vs-all CNN and BERT classifiers are built, one for each category of MAMS Dataset. We have experimented with non-contextual word embeddings: Glove and word2vec while training the CNN classifier. For each review, class predictions are generated using predict classes function of Keras. MobileBERT model is also fine-tuned for contextual feature generation and category classification. Separate evaluation scores are recorded for each aspect category, and the average value is considered as the final score. We have conducted the same experiments on SemEval-2014 Restaurant Review Dataset using five binary one-vs-all classifiers to compare the results.

TABLE II
RESULTS OF ACD ON SEMEVAL-2014 RESTAURANT REVIEW DATASET

Model	Precision	Recall	F1score	Accuracy	AUC
Glove+CNN	89.61	79.84	84.42	93.00	88.25
Word2Vec+CNN	89.93	79.23	84.11	92.88	87.87
BERT	91.15	89.64	90.37	95.25	93.21

TABLE III
RESULTS OF ACD ON MAMS DATASET

Model	Precision	Recall	F1score	Accuracy	AUC
Glove+CNN	87.28	85.07	85.96	93.63	89.94
word2vec+CNN	87.24	86.28	86.59	93.94	90.58
BERT	87.05	88.28	87.57	94.28	91.54

B. Aspect Category Sentiment Classification

The purpose of this subtask is to identify the aspect polarities related to all the aspect categories discussed in the reviews. Each review in MAMS Dataset has at least two different aspects with different corresponding polarities. The sentiment polarities present in the dataset are : *positive, negative and neutral*.

1) Preprocessing: We lower-case, remove stopwords and tokenize the MAMS and SemEval 2014 Restaurant Review Dataset for the ACSC task.

2) Hyperparameters: In our experiment, we randomly split the data 20 percent of the dataset and used it as a validation set. The word vectors are formatted using Glove [24]. We pre-trained the word embedding vector on an unlabeled corpus with size 840 billion. The remaining parameters use a uniform distribution U(ε,-ε). We take aspect embeddings, hidden layer, and word vectors size as 300. We keep the length of attention weights the same as sentence length. We take inspiration from Theano [30] approach to implement our deep learning model. We use a batch size of 25 to train the model with a momentum of 0.9. We use 0.001 in L2-regularization weight with an initial learning rate of 0.01.

3) Training Process: The task in our study involved determining polarity related to each aspect in a positive/negative/neutral basis. The paper proposes using attention weights assigned to each word related to an aspect to determine its appropriate polarity. We use the Parameter set [A, W_h, W_v, W_p, W_x, w]. Also, [W_i , W_f , W_o, W_c] dimensions are expanded to concatenate the aspect embedding.As part of this mechanism, we optimize the aspect and word embeddings during the training phase. We see 5 percent out-of-vocabulary words which are initialised at random with = 0.01.

TABLE IV
RESULTS OF ACSC

Model	Accuracy	F1
Attention Based LSTM for SemEval 2014	76.78	65.78
Attention Based LSTM for MAMS	73.48	59.57

For our optimization technique use AdaGrad [31]. AdaGrad [32] approach adjusts the learning rate in accordance to the involved parameters. It performs fewer updates for frequent parameters and more updates for less frequent parameters. Also, we initialize all the word vectors using Glove. We pre-trained our word embedding vectors on unlabeled corpus with a size of 840 billion. We take a batch size of 25 to carry out the training phase.

VI. EVALUATION

We have conducted several experiments for the ACD and ACSC sub-tasks, and reported the results in the following tables. Table II and Table III present results of Aspect Category Detection on SemEval 2014 Restaurant Dataset and MAMS Dataset, respectively. Table IV presents results of Aspect Category Sentiment Classification on both the datasets.

A. Aspect Category Detection

In spite of using only the text reviews as input data and initializing them with word embeddings, our models have achieved competitive results for Aspect category Detection. The BERT model applied on SemEval-2014 Restaurant Dataset, achieves the F1-score of 90.37 beating the best submission score of 88.58 in SemEval-2014 competition. The F1-scores of context-independent word embeddings (Glove and word2vec) are slightly better for the MAMS Dataset than the SemEval 2014 Restaurant Dataset. However, the F1-score of context-dependent language model BERT has dropped inconsiderably for the MAMS Dataset as compared to the SemEval-2014 Restaurant Dataset, but still performs better than Glove and word2vec. The performance of the BERT model can be further improved by fine-tuning the language model on domain-specific corpora, which is pre-trained on general text corpora [33]. Another interesting path for future research will be to examine cross-domain correlations for additional domains such as hotels, that are similar to restaurants [34].

B. Aspect Category Polarity

As part of our experiment, we have realized how crucial it is to analyze what word depicts the polarity of a specific aspect. We evaluate this by visualizing the corresponding attention weights. We can evaluate attention weight using equation 14 and adapt the weights accordingly. This shows how this mechanism focuses on specific words impacting a given aspect. Attention weights help pick dynamically the key parts in a sentence. It comes very handy if more than one keyword exists and hence it detects them. Our MAMS dataset is a tough dataset with more than one aspect in a review and different corresponding polarities. We have seen from our experiment in Table IV that the attention mechanism works well with our dataset.

VII. CONCLUSION AND FUTURE WORK

In this paper, we have evaluated a series of experiments for ACD that involved traditional word vectors (BoW), neural

embeddings (Glove and Word2Vec) and pre-trained language model (BERT) approaches on a challenging dataset. The traditional word vectors approach has performed poorly on the MAMS dataset, but the other two approaches have achieved competitive results. We have presented a deep learning and a transformer-based approach which employs a simple 1D CNN and pre-trained language model BERT respectively. By using these models, we have evaluated context-dependent and context-independent word embeddings for representing the reviews of MAMS as well as SemEval 2014 Restaurant Review Dataset. Results show that it is hard to achieve good performance on the MAMS dataset, and context-dependent BERT model outperformed both the context-independent models (Glove and Word2Vec) on both the datasets. We have observed that contextual word representations are the most efficient technique to represent text for ACD. How to leverage the contextual BERT model by fine-tuning it on domain-specific and cross-domain corpora will be our future work for the ACD task.

Furthermore, we presented an attention-based LSTM model for ASCS. The key purpose of the learning is to gain insights on aspect embeddings and allow aspects to compute the corresponding attention weights better. Our model focuses on multiple parts of a sentence when given multiple aspects which makes it more competitive for such fine-grained classification problems. Our study shows that Attention-based LSTM performs better than the existing baseline mechanisms. For future work, there are various existing obstacles surrounding pre-training to go in what granularity, i.e. a word, subword, character level. Also, which structure language model would be appropriate to deal with this issue such as Transformer and LSTM, more detailed study is required in this field [35]. Moreover, instead of using Word2Vec or Glove approaches for the learning of the word embedding, it will be efficient to use context-sensitive word representations.

REFERENCES

[1] Liu, B. and Zhang, L., 2012. A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer, Boston, MA.

[2] B. Liu. Sentiment analysis and opinion mining Synth. Lect. Human Lang. Technol., 5 (1) (2012), pp. 1-167.

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

[4] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S. and Androutsopoulos, I., 2015, June. Semeval-2015 task 12: Aspect based sentiment analysis. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015) (pp. 486-495).

[5] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O. and Hoste, V., 2016, June. Semeval-2016 task 5: Aspect based sentiment analysis. In 10th International Workshop on Semantic Evaluation (SemEval 2016).

[6] Jiang, Q., Chen, L., Xu, R., Ao, X. and Yang, M., 2019, November. A challenge dataset and effective models for aspect-based sentiment analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 6281-6286).

[7] Kim, Y., 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[8] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[9] Sun, C., Qiu, X., Xu, Y. and Huang, X., 2019, October. How to fine-tune bert for text classification?. In China National Conference on Chinese Computational Linguistics (pp. 194-206). Springer, Cham.

[10] Yin, W., Schütze, H., Xiang, B. and Zhou, B., 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. Transactions of the Association for Computational Linguistics, 4, pp.259-272

[11] Golub, D. and He, X., 2016. Character-level question answering with attention. arXiv preprint arXiv:1604.00727.

[12] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C., 2016. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

[13] Rush, A.M., Chopra, S. and Weston, J., 2015. A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.

[14] Tang, D., Qin, B., Feng, X. and Liu, T., 2015. Effective LSTMs for target-dependent sentiment classification. arXiv preprint arXiv:1512.01100.

[15] Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[16] Ruder, S., Ghaffari, P. and Breslin, J.G., 2016. Insight-1 at semeval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis. arXiv preprint arXiv:1609.02748.

[17] Zhou, X., Wan, X. and Xiao, J., 2015, February. Representation learning for aspect category detection in online reviews. In Twenty-Ninth AAAI Conference on Artificial Intelligence.

[18] Mnih, A. and Hinton, G.E., 2009. A scalable hierarchical distributed language model. In Advances in neural information processing systems (pp. 1081-1088).

[19] Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y. and Zhou, D., 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. arXiv preprint arXiv:2004.02984.

[20] Mohammad, S.M., Kiritchenko, S. and Zhu, X., 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. arXiv preprint arXiv:1308.6242.

[21] Qian, Q., Tian, B., Huang, M., Liu, Y., Zhu, X. and Zhu, X., 2015, July. Learning tag embeddings and tag-specific composition functions in recursive neural network. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 1365-1374).

[22] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P., 2011. Natural language processing (almost) from scratch. Journal of machine learning research, 12(ARTICLE), pp.2493-2537.

[23] Le, H.T., Cerisara, C. and Denis, A., 2017. Do convolutional networks need to be deep for text classification?. arXiv preprint arXiv:1707.04108.

[24] Pennington, J., Socher, R. and Manning, C.D., 2014, October. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[25] Goldberg, Y. and Levy, O., 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

[26] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

[27] Wang, Y., Huang, M., Zhu, X. and Zhao, L., 2016, November. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 conference on empirical methods in natural language processing (pp. 606-615).

[28] Rocktäschel, T., Grefenstette, E., Hermann, K.M., Kočiský, T. and Blunsom, P., 2015. Reasoning about entailment with neural attention. arXiv preprint arXiv:1509.06664.

[29] Kingma, D.P. and Ba, J.L., 2015. Adam: A method for stochastic gradient descent. In ICLR: International Conference on Learning Representations.

[30] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D. and Bengio, Y., 2012. Theano: new features and speed improvements. arXiv preprint arXiv:1211.5590.

[31] Duchi, J., Hazan, E. and Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research, 12(7).

[32] Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M.A., Senior, A., Tucker, P., Yang, K. and Le, Q.V., 2012. Large scale distributed deep networks. In Advances in neural information processing systems (pp. 1223-1231).

[33] Xu, H., Liu, B., Shu, L. and Yu, P.S., 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. arXiv preprint arXiv:1904.02232.

[34] Rietzler, A., Stabinger, S., Opitz, P. and Engl, S., 2019. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. arXiv preprint arXiv:1908.11860.

[35] Zhou, J., Huang, J.X., Chen, Q., Hu, Q.V., Wang, T. and He, L., 2019.Deep learning for aspect-level sentiment classification: Survey, vision,and challenges. IEEE Access, 7, pp.78454-78483.