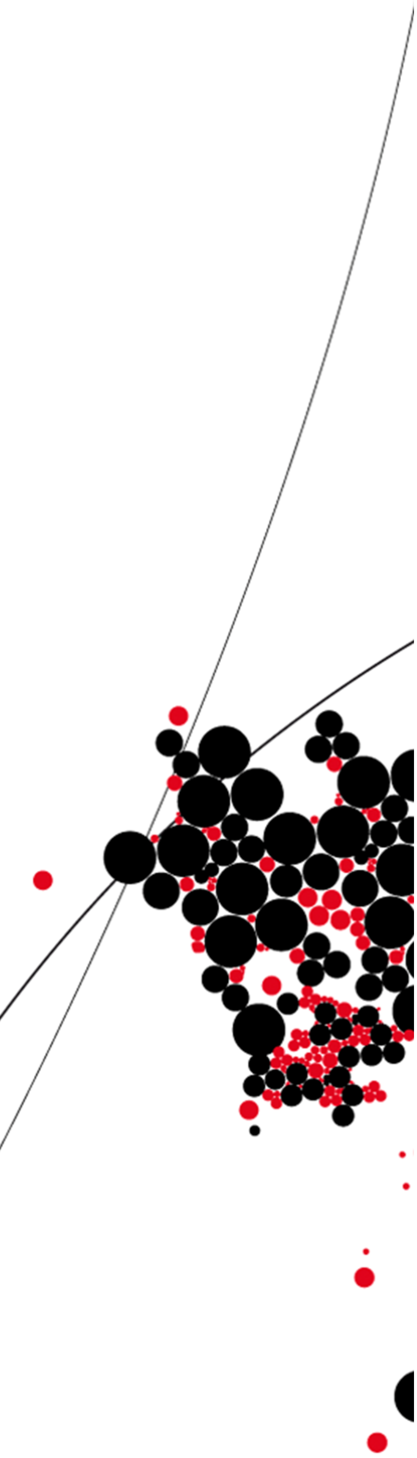
UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering,
Mathematics & Computer Science



Aspect Based Sentiment Classification of Multilingual Customer Reviews

M.Sc. Thesis
Yash Gupta

Industrial Supervisor:
Berk Yenidogan
Data Scientist
Mercedes Benz Customer Assistance Center
Maastricht N.V.

Evaluation Committee:
Dr. IR. Maurice Van Keulen (Committee Chair)
Department of Data Management & Biometrics

Dr. Ing. Gwenn Englebienne
Dr. Shenghui Wang
Department of Human Machine Interaction

Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Abstract

This work aims to find suitable techniques to improve the performance of a state of the art system [1] for the task of aspect based sentiment analysis [2] of customer reviews for a multi-lingual use case. The authors of [1] provide improvement in performance when compared to baseline with the help of auxiliary sentences and state two reasons for this increase. The first one is the increase in the size of training set exponentially and the second is better sense for sentence pair classification for the BERT model when compared to single sentence classification. Three motivated changes are experimented with the state of the art design and training techniques to verify if the reasons stated by authors are actually the reasons behind the increase in performance and also improve the performance of the state of the art system [1]. The baseline systems are developed as demonstrated by authors of [1] but unlike the authors, two baseline systems are developed one with a pre-trained BERT model [3] and one with a pre-trained BERT-multilingual model [3]. To conduct experiment 1, systems are fine-tuned with the above mentioned models on sentence pair classification with auxiliary sentences to perform ABSA [2]. The systems are trained first with authors' approach and then with auxiliary sentences in the language of the review. To conduct experiment 2, both BERT and BERT-multilingual models are fine-tuned via multi-task learning (which took place in effect while fine-tuning with auxiliary sentences with the authors' approach) without auxiliary sentences.

After experimentation, it is concluded that the state of the art [1] can indeed be redesigned to train with multi-task learning (without auxiliary sentences) to provide better results. It is also concluded that the reason behind the increased performance in the state of the art system [1] is multi-task learning which takes place in effect when trained with auxiliary sentences and a better sense of sentence pair classification for the model and not the increased size of training set. Instead, it is observed that the increased data hinders the learning potential of the systems.

The dataset for experimentation is provided by Daimler A.G. subsidy, Mercedes-Benz Customer Assistance Center Maastricht N.V. which contains multilingual customer reviews labelled for different aspects of their business.

Acknowledgement

First, I would like to thank Dr. Gwenn Englebienne, Dr. Maurice van Keulen and Dr. Shenghui Wang of University of Twente, Netherlands for their supervision during this thesis, but also the preparatory Research Topics course. Their feedback and critical analysis of my work has proven to be a vital part for completion of this project. They also helped and supervised me to build research questions that could be aligned with the particular use case. Secondly, I'm grateful to a group of colleagues known as the Data Science Team at Mercedes Benz Customer Assistance Centre, Maastricht. Especially, Berk Yenidogan and Aranka de Barbanson who maintained their support during the project and provided with resources necessary to carry out this project. Berk also supervised with the project's structuring and documentation phase. I would also thank another colleague Emiel de Heij who supervised me during the implementation phase of the experiments needed to answer the research question of this thesis.

Secondly, I am very thankful for the support and encouragement received by my family throughout my studies, as well as my friends, on whom I could always count for necessary distractions when working on my thesis. Especially during the pandemic when family and colleagues both seemed distant, just across screens.

Contents

Appendices

List of Tables

List of Figures

Chapter 1

Introduction

Natural language processing is being used by corporations to grasp consumer insights. One of the most used concepts is Sentiment Analysis (SA), which uses the computational logic and processing powers of machines [4] to classify a given text into a fixed set of sentiment classes. Businesses use a trained sentiment analysis system to analyse consumer sentiment trends and gain insights into the market from customer reviews. The trained SA system assigns one single sentiment to a review/input. However, large corporations like Daimler A.G. subsidy Mercedes-Benz, receive reviews associated to multiple products/services and sentiments. Implying that a single customer review could belong to multiple sentiments associated to multiple products. In this case, organisations prefer to use an aspect based sentiment analysis [2] (ABSA) system. An ABSA system classifies an input review as multiple <aspect,sentiment> pairs, hence solving the above mentioned problem. This work aims to build an ABSA system for Mercedes-Benz Customer Assistance Center Maastricht N.V. and answer the research questions mentioned further in this chapter.

One of the state of the art approaches to build an ABSA system is mentioned in the article titled *'Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence'* [1] by S. Chi et al. The systems trained with this approach outperform other ABSA systems on the SemEval 2014 [2] dataset. Performance comparison is done by evaluating them on two tasks stemming out of aspect based sentiment analysis; namely, aspect category detection and aspect polarity detection. The authors' findings establish two systems, BERT-pair-NLIB (Natural Language Inference - B) and BERT-pair-QAB (Question/Answering - B) that outperform all other prevalent systems on the task of aspect category detection and aspect polarity detection respectively. Both approaches change the task of single sentence classification to sentence pair classification by generating auxiliary sentences using <aspect,sentiment> pairs. The authors state two reasons behind the state of the art results. First, that the two systems generate auxiliary sentences using all <aspect,

sentiment> pairs to train, hence exponentially increasing the amount of data available for training. Second, the use of sentence pair classification for fine-tuning the BERT-model which is also how the BERT model is pre-trained. Experimentation is defined to verify these reasons and also provide possible modifications to design and training techniques of these systems.

1.1 Motivation and Research Questions

The goal is to establish similar classification performance on a real world use case by adapting S. Chi et al's approach [1], represented by RQ. This use case consists of a multi-lingual dataset unlike the S. Chi et al's use case which consisted of only English reviews.

RQ. How can the state of the art approach be adapted to achieve the best performance on a multilingual dataset?

An important thing to note, is the way the BERT model is fine-tuned by the S. Chi et al' approach. They suggest that the fine-tuning process using sentence pair classification works better because the model has a sense of classifying sentence pairs by finding a relationship of their co-existence because of its pre-training technique. They make use of this to change the task of ABSA to a binary classification task. For instance, for a review R, an auxiliary sentence A is created by using a possible <aspect, sentiment> pair. The model is trained to classify if both the sentences R and A can exist together. Hence, changing the formation strategy of auxiliary sentences should have an affect on the performance of the pair approaches. The state of the art systems make use of English auxiliary sentences only. Assuming that the BERT model does classify sentence pairs better than single sentences in this case, the language of the auxiliary sentences plays an important role for the model during training. To provide a better sense of sentence pair co-existence, auxiliary sentences can be created in the language of review. This is expected to improve the system's performance for multi-lingual reviews. The successful improvement and effect of this change can be concluded after answering the question RQ 1.

RQ 1. To what extent does using auxiliary sentences in the language of the review improve the performance compared to using only English auxiliary sentences?

An auxiliary sentence is formed by a possible <aspect, sentiment> pair helping the model to learn about aspects and sentiments at the same time to predict a

{’Yes’, ’No’} result. This approach can also be viewed as multi-task learning since the systems is learning two tasks at the same time(aspect classification and sentiment classification). Hence, it can be argued that the increase in performance of state of the art approach is due to multi-task learning with transfer learning from the BERT model. It is also a possibility that the high amount of data that is generated with auxiliary sentences, hinders the learning potential of systems by generalizing them more during training with data with no new information. Hence, an approach should be evaluated to check if the same or better performance can be achieved with multi-task learning and transfer learning without the auxiliary sentences. The systems trained would be used to answer RQ 2.

RQ 2. To what extent does training the system with multi-task learning and transfer learning without auxiliary sentences improve the performance of the system compared to using auxiliary sentences?

The state of the art approach uses the BERT model as the base model for the SemEval dataset which has customer reviews in only English language. The BERT model uses word-piece tokenization to form tokens out of it’s input before processing. Hence, every word in the input be it in English or any other language is broken down by the model to word-pieces that have a semantic meaning to the model. So, if the model gets an input in any other language, for instance Italian, it will possibly break all words to characters or very small pieces that do not have much semantic meaning by definition in case of BERT model. However, if a BERT-multilingual model is fed the same input, it would form bigger word-pieces which would hold semantic meaning to the model. Therefore, changing the base model from English pre-trained BERT model to pre-trained BERT-Multilingual model should have a positive effect on the systems’ performances. The results from this change are used to answer RQ 3.

RQ 3. To what extent can using a pre-trained multilingual BERT model improve the performance compared to using the English pre-trained BERT model?

1.2 Scientific Contributions

The answer for the research question and it’s sub-questions will lead to the best possible adaption of the S. Chi et al’ approach to build an ABSA system for a multilingual use case.

The answer for RQ 3 will lead to the choice of the base model for fine-tuning and developing a system. This will also help to analyse and compare the performance of the BERT model and BERT-multilingual model on a multi-lingual dataset. The answer to RQ 1 will help to strengthen the understanding of sentence pair classification. The authors of [1] state that sentence pair classification helps to fine-tune the base model better and hence, the model trained with auxiliary sentences in the language of review, is expected to perform better. Moreover, the answer to RQ 2 will help to identify the main reason behind the increase in performance from the author's approach [1]. If the augmented training data is hindering the training approach, systems trained to answer RQ 2 are expected to perform better than the other models. In all, the answers will help to develop an ABSA system adapted from the state of the art approach.

Chapter 2

Background and Related Work

In this section, the background of this project is described. Also, related articles describing to develop ABSA [2] systems are discussed.

2.1 Background

This section details on how sentiment analysis (SA) systems are developed and used deliver consumer insights. It also mentions the limitations of using an SA system and how they can be tackled using an aspect based sentiment analysis (ABSA) system. Later, it describes the traditional and modern methods generally used to develop text classification systems for tasks like SA and ABSA.

2.1.1 Sentiment Analysis Systems

Machines are trained to identify sentiments involved in a given text and then classify it to one of the pre-defined sentiment classes. Sentiment classes vary from project to project but usually one of these two sets, 1) {'positive' , 'neutral' , 'negative} 2) {'highly positive' , 'positive' , 'neutral' , 'negative' , 'highly negative'} is used as the target set. The text is classified by a system [5] using a mathematical function returning a net polarity of the text. The function is then made more precise by the system as the function is optimized during training over data. Many methods are prevalent to encode the words into a numerical format to prepare numerical data (from textual data) for training the system.

The state of the art methods include word2vec [6] and doc2vec [7], which have proven to be very efficient for training neural networks and recurrent neural networks [5]; GloVe has helped deliver state of the art results as well [8] by capturing fundamental count data and forming linear sub-structures within the text.

Once the words are converted to numerical vectors i.e. quantified, they are fed to a machine [4] classifying texts into sentiment classes. For example, consider the sentence, “It was a great day today!” The sentence would first be encoded into a vector containing numerical values capturing the semantic and syntactic relationship between the words present in the sentence. A trained system using a machine-learning model like the Deep Average Network [9] in the background would classify the formed vector into a class from a set of pre-defined classes. Table 2.1 visualizes the input and output of a trained SA system considering it is trained on three sentiment classes {‘positive’, ‘neutral’, ‘negative’}.

Input Text	Output
It was a great day today!	‘Positive’

Table 2.1: Input/output of a trained SA system(2)

The decision of the system is mostly driven by the word “great” in the presented case. The technique of SA is put to an industrial use in a very efficient manner to generate business insights [4] from customer data. SA is used to analyze thousands of customer reviews at a single go to get a grasp of customer feedback of the products and services offered by businesses. Customers usually use describing words/adjectives in their reviews that help the machine learning models to identify the existing sentiment.

2.1.2 Aspect Based Sentiment Analysis Systems

There are a few drawbacks of using SA on the industrial level. One of them being the inability of SA approach to identify multiple sentiments involved in one single document or review. Businesses offer a vast variety of products and services to their consumers and hence receive feedback about all of them at once. There might be some cases where the same consumer reviews multiple products/services in a single review with multiple sentiments involved. For example, a restaurant receives a feedback, “The food was good, but the service was disastrous.” Table 2.2 represents the input and output when considering the same SA system.

Input Text	Output
The food was good, but the service was disastrous.	‘Negative’ or ‘Neutral’

Table 2.2: Input/output of a trained SA system(2)

As the sentence consists of both positive and negative sentiments involved, the system completely ignores one of the sentiment. This creates a roadblock for organizations to recognize all the sentiments involved in their customer feedback. In addition, it does not allow them to zero-in on the specific product, service or department that is not receiving a positive feedback.

Aspect Based Sentiment Analysis (ABSA) [2], [1] is an approach taken to overcome the above mentioned roadblock. The approach tries to capture long-term dependencies between words in a document to identify multiple aspects and associated sentiments present in the review. Table 2.3 represents the output of a trained ABSA system where the same sentence is used as input to classify between three aspects {‘food’, ‘service’, ‘location’} and three sentiment classes {‘positive’, ‘neutral’, ‘negative’}.

Input Text	**Output**
The food was good, but the service was disastrous.	food’ : ‘Positive’ ; ‘service’ : ‘Negative’ ; ‘location’ : ‘None’

Table 2.3: Input/output of a trained ABSA system

The system is trained to identify multiple aspects present in the sentence from a pre-defined set of aspects and then assign a sentiment to them. The reason for the aspect ‘location’ receiving ‘None’ as output and not ‘Neutral’ is that the sentence does not say anything about the location of the restaurant. Therefore, an ideal system for organizations to develop and deploy would be an ABSA system that equips them to identify the aspects receiving negative sentiments. The ideal system would allow managers and organizations to instantly recognize propositions not being accepted by consumers in a positive manner. This would help them optimize operations towards a more customer centric approach providing intelligence and insights from consumer data.

2.1.3 Traditional Methods of Text Classification

The first article named “The Cross-Out Technique as a Method in Public Opinion Analysis” [10], [11] related to sentiment analysis dates back to the year 1940. The article helped to analyze sentiments of multiple reviews at once and triggered a new phase in opinion analysis. As the field progressed over the years, techniques were used to analyze public sentiments in masses after world wars and other political and socio-economic events. In addition, the industry started relying on the approach to understand their customer better. By mid 1990s, the industry started using logical capabilities and computing powers of machines to process tasks, for instance, “Elicitation, Assessment, and Pooling of Expert Judgments Using Possibility Theory” [12] was published in 1995 , which helped in expert opinion analysis by pooling similar reviews together. This progress can be credited to the fast and revolutionizing developments of processing engines and chips that can leverage the large processing capabilities to generate insights.

This development has led to the rise of application of machine learning techniques and methods to perform tasks like humans in the industrial domain. Organizations now use systems to generate business intelligence insights from large quantities of textual data at a single go [4]. In a nutshell, the task is to represent textual data in a numerical format, and train a system to identify patterns which it uses to classify data. The techniques were also supported by the constant development of techniques like word2vec [6], doc2vec [7] and GloVe [8]. All the three techniques aim to represent words or documents with a vector that would be used to train systems. Word2vec formed vector representation of words by capturing affect and context of neighboring words. In addition, a window can be defined to determine how many neighboring words have to be considered to form a word's vector representation. This window can also be defined in a skip-gram format implying that not only continuous words can be considered for creating vector representations. Doc2vec took the same approach but delivered a vector representation for a whole document and not just a word. The way it did that was by keeping word vectors from word2vec and assigning special indexes/vectors to paragraph topics or paragraph ids. All these topic vectors provide a representation of the order of paragraphs presents in the document. This enabled the vector to represent the words as well as paragraphs/documents. GloVe made use of fundamental count data related to the presence of words in a document and corpus (all the textual data) along with capturing semantic relationships by forming sub-patterns in text. After the development of such state of the art techniques, natural language processing took a big turn.

Traditionally, the task of sentiment analysis started out with feed forward neural networks. They take the text as a bag of words formed by vector representations achieved by embedding models like word2vec. All word vectors are summed up or averaged out to form an input representation of a bag carrying all words, which is then fed to a neural network. One of the examples of a feed forward network for text classification task would be the Deep Average Network [9]. Figure 2.1 below represents the data flow and architecture for a DAN. Another extension of DAN was fasttex [13], which inputted text in a bag or words format just like DAN,moreover it also incorporated a new feature capturing local word order information. These systems are then trained on specific tasks to output sentiments of text.

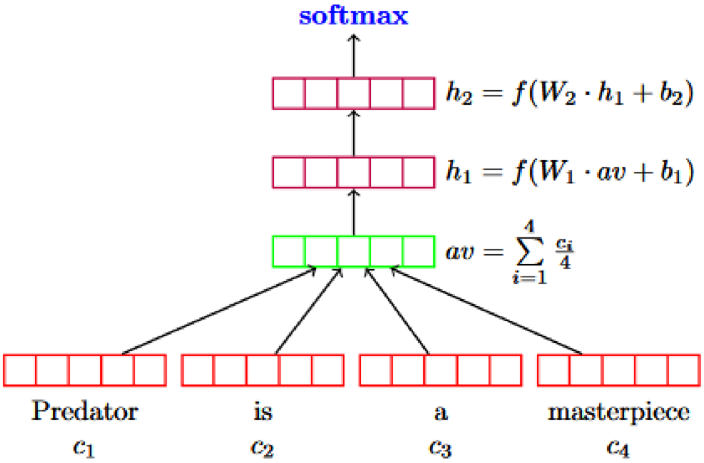

softmax
$h_2 = f(W_2 \cdot h_1 + b_2)$
$h_1 = f(W_1 \cdot av + b_1)$
$av = \sum_{i=1}^{4} \frac{c_i}{4}$
Predator
c_1
is
c_2
a
c_3
masterpiece
c_4

Figure 2.1: Architecture of Deep Average Network

Following feed forward networks, were recurrent neural networks. Instead of taking text as bag of words for input, RNNs read words sequentially in a given text. This helps to capture dependencies between words in a more precise manner and capture long term dependencies are realized. However, vanilla RNNs end up with exploding or vanishing gradients while training not being able to capture long term dependencies often. This problem was solved by adding a memory cell storing historical information of words. The amount of information in these cells is controlled by three gates, the input, output and forget gates. This architecture helped systems to capture long-term dependencies formed in a text, which is imperative for tasks like aspect based sentiment analysis. It was termed as the Long Short Term Memory RNN or LSTM [14], [15], [16]. The LSTM-RNNs were also improved by transforming the architecture from a chain model to a tree model creating a cell to store historical information for multiple child cells. Another interesting development in this context was the development of Multi Timescale LSTM or MT-LSTM which incorporated the time of occurrence of text as one feature and stored this information in a memory cell. The connections of the networks would be activated only if they belonged to a certain time period. Later, a bi-directional LSTM or bi-LSTM [17] was also proposed which incorporated two-dimensional max pooling for attaining information about textual features. Since the amount of information that could be captured while training increased, the performance of the classifiers was enhanced.

Using simple networks like DAN and Fasttext might render good results for sentiment analysis tasks. However, they fail to capture long-term dependencies inside a given input text since they process the encoding of all words at once. Hence, if trained for aspect identification such networks would not perform to deliver desired results as they would not understand the relationship between words. The only context that the network would understand would be that provided by the embedding formed from words and documents. Coming to recursive nets, the processing flow becomes sequential and every word is processed one after the other. This allows the network to capture some dependencies between words. However, the serial processing is highly expensive and costs a ton of time and resources. Moreover, even bi-LSTM would fail to capture dependencies between words present at the two terminals of a given long input sentence/document. So, if the above-mentioned networks are trained to perform aspect identification and sentiment classification they would not show good results. In addition, the process of training RNNs would be very time consuming.

2.1.4 Transformers and BERT

In 2016, Yang et. al. [18] proposed attention mechanisms that could reduce the amount of processing required and captured word dependencies in a much better way. The classification mechanism works in two steps mainly where 1) the document was interpreted in a hierarchical manner and 2) special attention was provided to important instances present at sentence level, document level and word level whereas unimportant parts of the text were not provided with such attention. This reduced the amount of iterations required to train neural networks as the number of iterations required to train them reduced drastically. Due to this advancement, further developments were made to train light weight neural networks and recursive neural networks to perform tasks of text classification. Y. Liu et. al. [19] and T. Shen et. al. [20] propose the application of attention mechanisms to train bi-LSTMs and RNN/CNN respectively. However, the recursive nature of such networks were highly time consuming.

The only bottleneck now was the sequential and long processing nature of RNNs. Although, if replaced by CNNs, the processing in a sequential manner becomes less cost effective, the computational cost to capture relationships between words in a sentence also grows with increasing length of the sentence. This is why, in 2017, Vaswani et. al. from Google proposed a Transformer [21] architecture which comprises of an encoder and a decoder. Instead of representing documents in a hierarchical way, the architecture proposed to quantify the relationship between each word present in a given text document and then provide attention to the most important relationships at the encoder end. At the decoder end, these matrices containing relationships between words are transformed into a key value pair. The key is formed by the output already produced by the decoder and the pair is optimized over training. This architecture rendered the recursive nets highly inefficient as the transformer could train in a semi-supervised fashion without recursive processing. Figure 2.2 below represents the proposed transformer architecture.

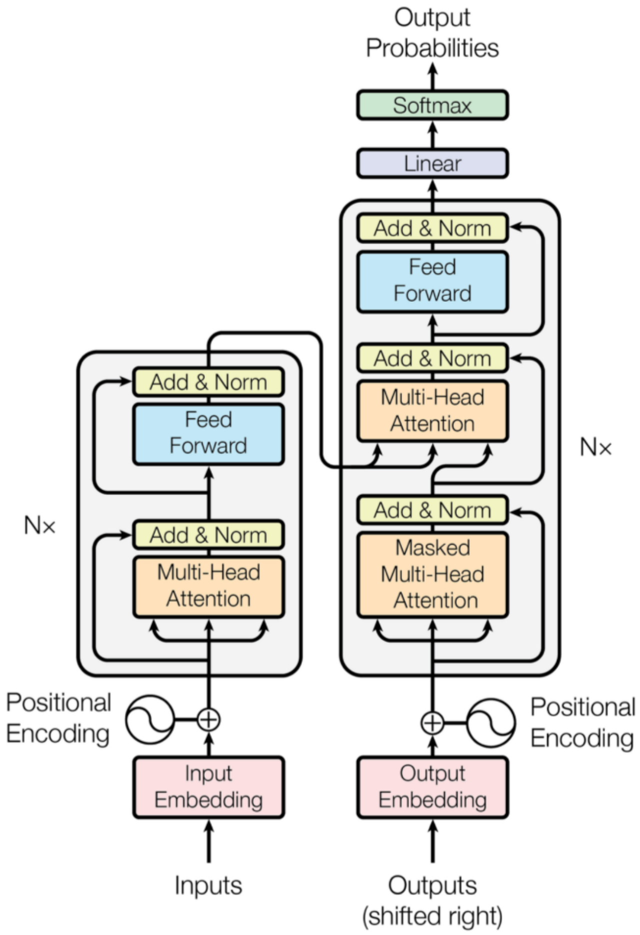


Figure 1: The Transformer - model architecture.

Figure 2.2: Transformer Architecture

The transformer architecture sparked a new revolutionary era in the space of natural language processing. The processing cost and time to train smart systems running on the transformer architecture reduced drastically. In 2018, Devlin et. al. extended the transformer architecture to form BERT [3], a pre-trained bi-directional transformer trained on huge amounts of textual data on next word prediction tasks and sentence pair classification task. This pre-trained system was used to fine tune many tasks as specific as aspect based sentiment classification and the fine-tuned system also delivered [1] state of the art results.

One of the main examples of fine-tuned systems for aspect based sentiment classification is ‘Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence’ [1] proposed by Chi, Sun & Huang, Luyao & Qiu, Xipeng. The system delivers state of the art results on SemEval 2014 Task 4 and Sentihood 2016 for aspect identification and sentiment classification and target extraction and sentiment classification respectively. The approach considers the generation of auxiliary sentences so that the final output vector received from BERT can be fed to classification layers. The idea behind the architecture is to exponentially increase the amount of data available for training. However, the data augmentation techniques render auxiliary sentences that contain little to no information about a specific aspect or a sentiment related to it. In addition, the mentioned datasets are in English and hence the performance of this system is not evaluated for a multi-lingual task. We base our project out of this article and form research questions that analytically evaluate the performance of this system and propose ways to improve the performance on multi-lingual tasks.

2.2 Related Work

This section mentions and critically analyses the existing solutions to aspect based sentiment analysis task. [22] provides and overview of systems performing aspect based sentiment analysis and evaluation methods. The basic aim of ABSA is to identify a sentiment communicated by multiple reviews concerned to a particular aspect. However, the aspect and sentiment might be explicitly or implicitly defined in a given text. For example, the text; "You can come out on top here!"; has an implicit aspect but an explicit sentiment. The text fails to mention an aspect but communicates a positive sentiment. In this work, we ignore such examples which do not explicitly mention or imply an aspect from a predefined set of aspects. Implicit sentiments do not pose any challenge to the proposed solutions. [22] The solutions proposed for ABSA can be categorized into three categories, namely, knowledge based approaches, machine learning based approaches and hybrid approaches.

Usually, [22] knowledge based approaches make use of a lookup sentiment dictionary. The keys of this dictionary are words from the corpus (not all) and the value is the sentiment associated with that word. The system accumulates the sentiment originating from different words from it's knowledge base and classifies a document accordingly. [23] mentions sentic flow, a technique that provides the system the ability to keep flow of sentiments from one concept to the other. To carry out this task, the system accumulates sentiments related to words from it's knowledge base and attaches the sentiment to a concept graph made from the corpus. This way, the system then finally declares the sentiment related to different concepts and hence performing the task of ABSA. The development of knowledge based solutions would need knowledge of multiple linguistic domains since the task is to perform ABSA on a multi-lingual environment. Hence, we do not include knowledge based approaches and transitively the hybrid approaches in our proposed solutions.

There is a recent rise in machine learning techniques to perform ABSA. With the help of attention neural networks [24], the systems can view at some part of the text with high attention. Also, the focus of this attention also changes from input to input. [25] proposes a method named Content Attention Based Aspect based Sentiment Classification (CABASC) model. The model uses a weighted memory module taking into the ordering of words and their correlations with each other. This solution out performed prevalent methods for ABSA like support vector machine (SVM) and a Long Short-Term Memory model (LSTM) on the SemEval [2] 2014 dataset. This solution also outperformed recurrent attention networks [26] using deep bidirectional LSTMs, multi-hop attention and position based attention mechanisms to generate custom memories for a particular aspect from a given text.

The paper [27] explains a method called Left-Center-Right separated neural network with Rotatory attention (LCR-Rot). The proposed solution uses three LSTM models corresponding to left context, right context and target phrase. The model would identify aspects related to the words from both ends and also use a rotatory technique to model relationship between aspects and the target phrase. The LCR-Rot model also outperformed the CABASC model [28].

The recent state of the art methods make use of a transformer architecture [21]. The transformer trains with self attention as described in section 2. Most approaches make use of the BERT model [3] by fine-tuning the pre-trained model on specific tasks like ABSA. In [29] the authors use the technique of machine reading comprehension to perform ABSA. They collect many customer reviews to form passages of the BERT language model which is then able to answer questions about aspects mentioned in the reviews. This solution achieved state of the art results in 2019. Although this solution is easy to execute, it fails to provide a technique to handle the challenge of data scarcity present for a particular aspect. For instance, if only a small number of reviews mention aspect ’A’, while many mention aspects ’B’, ’c’, and ’D’, the passage formed by accumulating reviews will not have a balanced representation of all aspects. Hence, the system will fail to answer anything precisely about aspect ’A’.

Another solution [1] that provided state of the art uses auxiliary sentences to perform ABSA. The system generates auxiliary sentences to change the task from single sentence classification to sentence pair classification with the BERT language model. The system provides state of the art results on SemEval [2] 2014 dataset. The authors credit the high performance score to the technique of matching pre-training and fine-tuning techniques of sentence pair classification. However, the generated auxiliary sentences do not contain much information about any aspect or sentiment present in the review. The same level of performance might be possible to achieve with the help of multi-task learning and transfer learning approaches with the BERT language model. As the formation of auxiliary sentences increases the amount of training data exponentially, it does not provide any relevant information to the model. Hence, this work aims to investigate if the good performance of the model can be credited to auxiliary sentences or not. Another interesting aspect of this approach is to investigate the effect of multi-lingual auxiliary sentences for a multi-lingual dataset.

Chapter 3

Technical Contributions

This chapter mentions the technical contributions made to carry out this project. The project is based on a classification technique [1] that augments data before training and evaluating systems. Some motivated modifications have been suggested to the data augmentation technique for training and evaluating systems in experiment 1. This change is motivated to provide a better sense of understanding for sentence pair classification to the models by changing the language of auxiliary sentences to that of the review. Moreover, some changes are devised in the state of the art architecture by changing the learning technique to multi-task learning without auxiliary sentences in experiment 2. This modification is suggested to keep the same training technique as S. chi et al. [1] but without the auxiliary sentences. In all, all systems are trained with both BERT and BERT-multilingual model. Unlike the approach of authors of [1], the BERT-multilingual model is also used to perform ABSA.

3.1 Data Augmentation

The authors of [1] use auxiliary sentences to increase the size of training set and provide a sense of sentence pair classification to the base BERT model. Each record generates a*s number of new training records from one single record where a is the number of possible aspects and s is the number of possible sentiments. Each auxiliary sentence is formed by a possible <aspect, sentiment> pair. The sentences are formed in English by the Natural Language Inference - B (NLIB) technique and also the Question/Answering - B (QAB) technique proposed by authors of [1]. A modification is made with the NLIB-lang and QAB-lang approaches where the auxiliary sentence is created in the language of the review for sentence pair classification. This change aims to provide a better sense of sentence co-existence to the base BERT and BERT-multilingual models. These approaches are used to train systems for experiment 1.

3.2 Architectural Adjustments - Multi Task Learning

The auxiliary sentences mentioned in the last section are formed by each <aspect, sentiment> and then all sentences are paired up with a review to perform sentence pair classification. This implies that the model learns to classify a review as a particular aspect and sentiment at the same time. Hence, it can be said that the model is trained with multi-task learning using auxiliary sentences. However, it is a possibility that the auxiliary sentences limit the performance of systems by generalizing them more to NO new information. Hence, an architecture is devised to train the model with multi-task learning but without auxiliary sentences. This architecture has a*s number of output neurons where a is the number of possible aspects and s is the number of possible sentiments. This would enable the model to classify each input as an <aspect, sentiment> pair i.e. to an aspect and a sentiment at the same time. The model is not provided with any auxiliary sentences and the task is carried out by single sentence classification.

3.3 Model Adjustments - BERT Multilingual

All experiments have been carried out with both the BERT model and the BERT-multilingual models. Hence, the state of the art [1] system is adjusted to cater to the multilingual use case by replacing the base model from BERT to BERT-multilingual.

Chapter 4

Methodology

This chapter describes the methodology to process data, setup experiments and evaluate results. The data is first described and pre-processed to prepare it for training the machine learning models. To answer sub-questions RQ 1 and RQ 2, two experiments are designed namely Experiment 1 and Experiment 2 respectively. Subquestion RQ 3, is answered by taking into account the results from both these experiments as both experiments are carried out with both BERT and BERT-Multilingual models. The problem statements for both experiments are also described in this chapter in section 4.2. The results of experiments have been reported and discussed in the Chapter 5.

4.1 Data Description

This section provides a description of both the datasets being used for setting experimentation setup. One of the datasets is used for training the all the systems and the other is used to evaluate all the trained systems. The datasets are provided by Daimler A.G.. The dataset is labelled for sentiments with different business aspects. The dataset's comparison can be made using this section with that of the SemEval dataset described in Appendix A. The structure of both datasets is similar however they differ in number of aspects and number of languages present in the dataset.

This dataset for the project has been provided by Mercedes-Benz Customer Assistance Center Maastricht N.V. a subsidiary of Daimler A.G. It contains records received from customers of Mercedes-Benz Customer Satisfaction Survey. The dataset has 1600 records with 52 columns. Each record can be classified into classes from a set of 21 classes. For the concerned project only natural text data i.e. input of the customer in 'customer_feedback' field would be used to train the system. The 21 classes namely are, 'CSR - Speed of the answer', 'CSR - Solution provided', 'CSR - Friendly/helpful', 'CSR - Competent/ Professional', 'CSR - Understanding of expectations', 'CSR - Communication quality', 'CAC / Process - Call/ email process', 'CAC / Process - Waiting time', 'CAC / Process - Case Ownership', 'CAC / Process - Speed of the solution', 'CAC / Process - Solution provided', 'Dealer/ Overflow Provider - Speed of the solution', 'Dealer/ Overflow Provider - Solution provided', 'Dealer/ Overflow Provider - Friendly/ helpful', 'Dealer/ Overflow Provider - Competent/ Professional', 'MPC/HQ - GDPR/Website', 'MPC/HQ - Company Policy', 'MPC/HQ - Friendly/ helpful', 'Product/ Service - Vehicle quality', 'Product/ Service - Service (CMS) quality', 'Product/ Service - Accessory quality'. There are total six languages namely English, Italian, Spanish, German, Dutch and French, in which a customer review might exist. A column specifies the language of the review in the dataset. Out of the 1600 records 372 records or reviews have not been labelled for any of of the mentioned classes. Hence, these reviews are removed before splitting data for training and evaluation. Therefore, a total set of 1228 multi-lingual reviews is available for training and evaluating our systems.

4.1.1 Grouping Classes

The size of the dataset is very low for any model learn about 21 aspects. Hence, some classes/aspects would be merged together to form a broader definition of aspects. However, merging classes just for creating a good distribution by ignoring the business representations of such classes would render the project impractical. Hence, certain business requirements have to be met in order to make use of the system. To form broader aspects and keep business goals aligned with the project, three classes or aspects are formed from the above mentioned aspects. The aspects are 'Customer Assistance Center', 'Dealer/ Retailer' and 'Products or services or head-quarters'. The sentiments for aspect 'Customer Assistance Center' are formed by merging sentiments of aspects 'CSR - Speed of the answer', 'CSR - Solution provided', 'CSR - Friendly/helpful', 'CSR - Competent/ Professional', 'CSR - Understanding of expectations', 'CSR - Communication quality', 'CAC / Process - Call/ email process', 'CAC / Process - Waiting time', 'CAC / Process - Case Ownership', 'CAC / Process - Speed of the solution', and 'CAC / Process - Solution provided'. The sentiments for the aspect 'Dealer / Retailer' are formed by merging sentiments of aspects 'Dealer/ Overflow Provider - Speed of the solution', 'Dealer/ Overflow Provider - Solution provided', 'Dealer/ Overflow Provider - Friendly/ helpful', and 'Dealer/ Overflow Provider - Competent/ Professional'. Finally the aspect 'Products or services or head-quarters' are formed by merging sentiments of the aspects MPC/HQ - GDPR/Website', 'MPC/HQ - Company Policy', 'MPC/HQ - Friendly/ helpful', 'Product/ Service - Vehicle quality', 'Product/ Service - Service (CMS) quality', and 'Product/ Service - Accessory quality. All records have been labelled to the sentiments {’positive’, ’neutral’, ’negative’, ’none’}. Ideally, sentiment for a aspect should be labelled ’conflicting’ to an aspect if sentiments of any two sub-aspects of that aspect have conflicting labels. However, for this use-case and data-distribution, such records are also labelled as ’neutral’.

After all records have been assigned sentiments for the three broad aspects, the dataset is split to form training and evaluation sets. The evaluation set takes 20% of the whole dataset and hence the training set forms 80% of the whole dataset.

4.1.2 Training Set

After the split and grouping classes, there are a total of 982 reviews available for training. Each record in the training set is labelled to an <aspect, sentiment> pair, where aspect belongs to the set {’Customer Assistance Center’, ’Dealer/ Retailer’, ’Products or services or head-quarters’} and sentiment belongs to the set {’positive’, ’neutral’, ’negative’, ’none’}. The table 4.1 below, represents the number of reviews labelled to each <aspect, sentiment> pair and figures 4.1 - 4.4 visualize these numbers.

	CAC	**Dealer/ Retailer**	**Products/Services/HQ**
Positive	218	166	25
Negative	315	263	265
Neutral	26	9	6
None	423	544	686
Total classified	595	463	311

Table 4.1: Distribution of reviews over aspect and sentiment

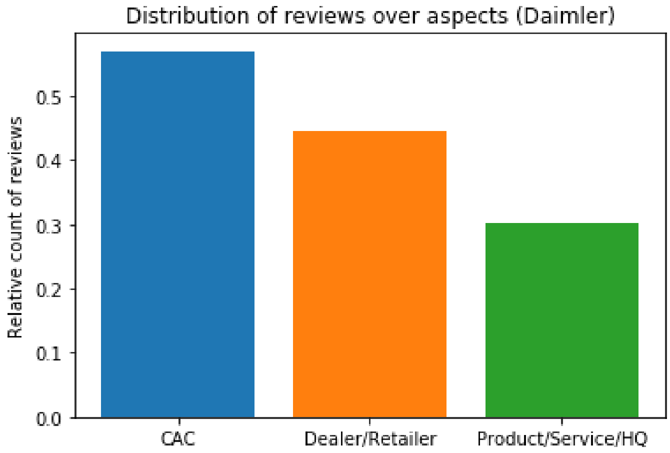


Figure 4.1: Distribution of reviews over aspects (Daimler)

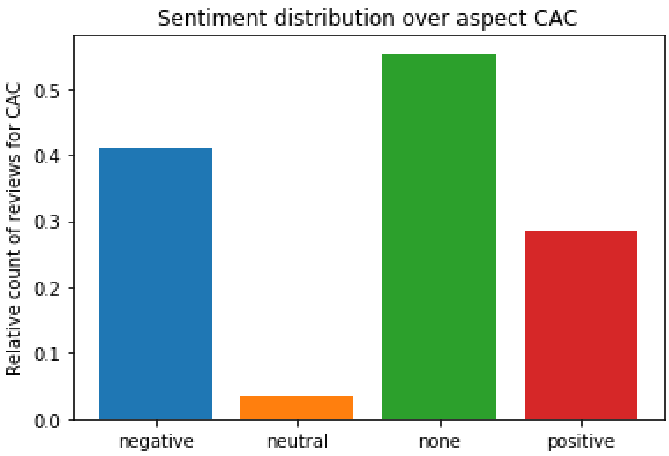


Figure 4.2: Sentiment distribution over aspect CAC

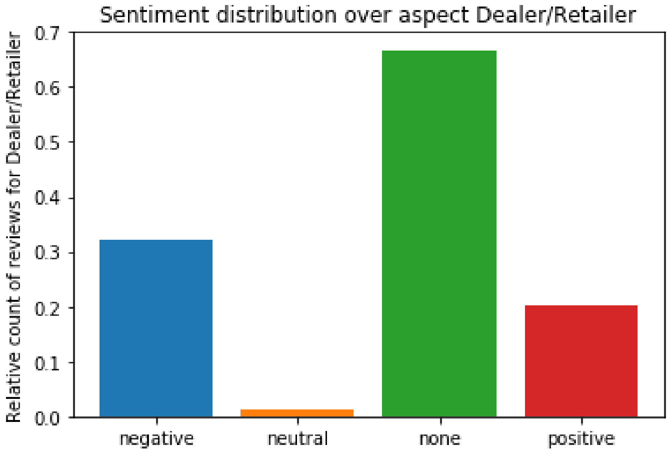


Figure 4.3: Sentiment distribution over aspect Dealer/Retailer

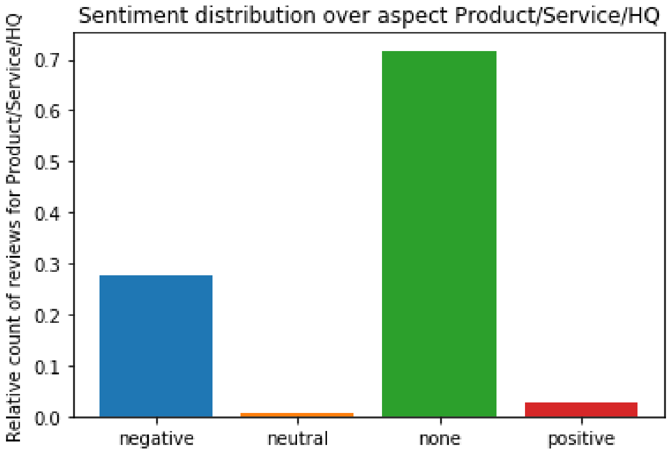


Figure 4.4: Sentiment distribution over aspect Product/Service/HQ

The training set has reviews in six languages, namely, English, German, Dutch, Spanish, Italian and French. The table 4.2 below represents the number of reviews available for training for each language and the figure 4.5 visualises this distribution over the training set.

Language	Number of reviews
English	382
German	44
Spanish	150
French	124
Italian	102
Dutch	150

Table 4.2: Distribution of training reviews over language

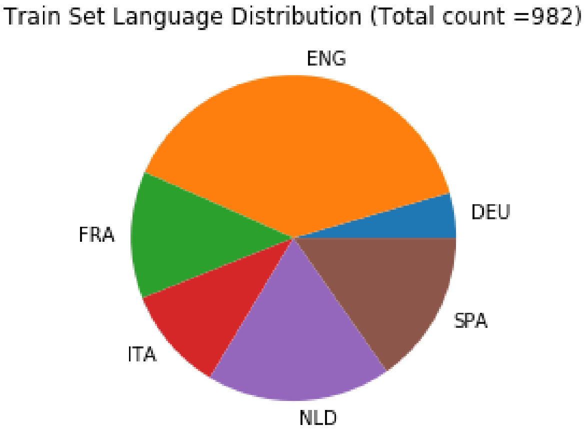


Figure 4.5: Distribution of training reviews over all languages

4.1.3 Test Set

After the split and grouping classes, there are a total of 246 reviews available to evaluate trained systems. Each record in the test set is also labelled to a <aspect, sentiment> pair, where aspect belongs to the set {'Customer Assistance Center', 'Dealer/ Retailer', 'Products or services or head-quarters'} and sentiment belongs to the set {'positive', 'neutral', 'negative', 'none'}. The table 4.3 below, represents the number of reviews labelled to each <aspect, sentiment> pair and figures 4.6 - 4.9 visualize these numbers.

	CAC	Dealer/ Retailer	Products/Services/HQ
Positive	61	34	3
Negative	87	51	68
Neutral	10	7	3
None	88	154	172
Total classified	158	92	74

Table 4.3: Distribution of test reviews over aspect and sentiment

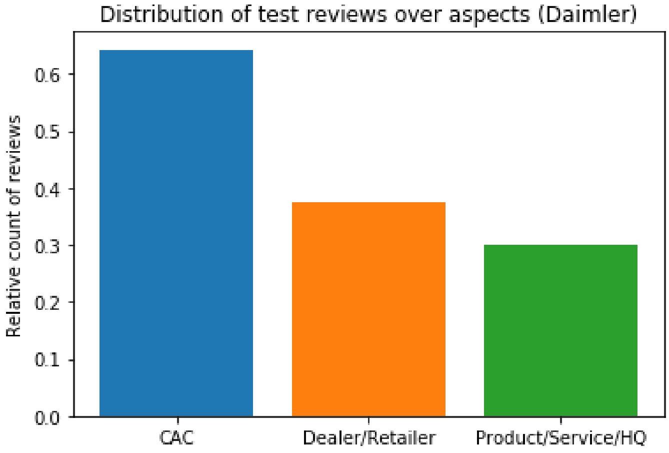


Figure 4.6: Distribution of test reviews over aspects (Daimler)

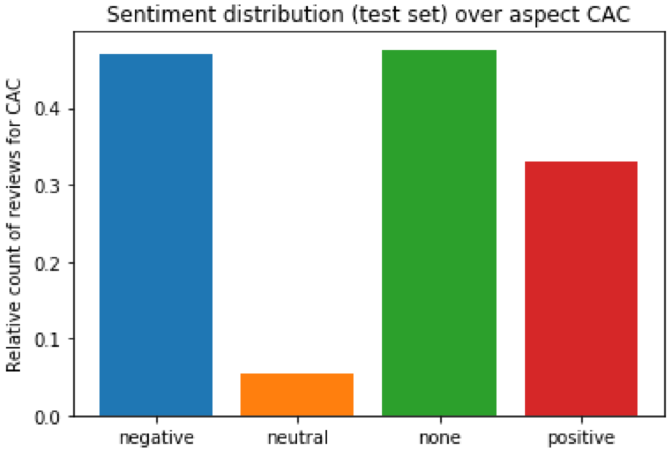


Figure 4.7: Sentiment distribution over aspect CAC (test set)

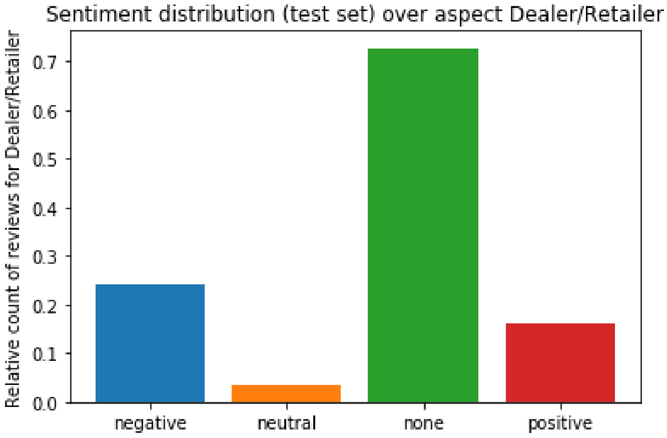


Figure 4.8: Sentiment distribution over aspect Dealer/Retailer (test set)

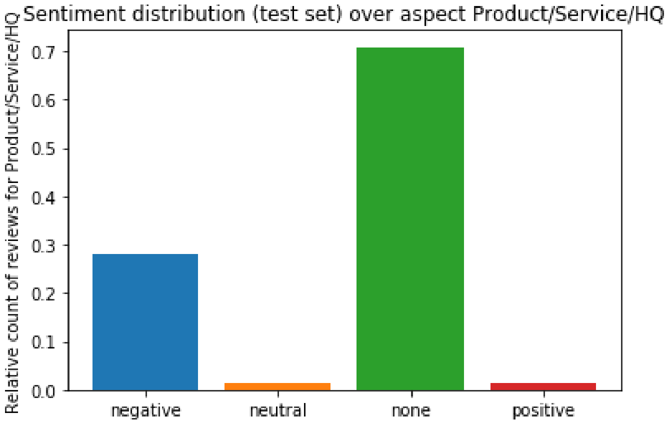


Figure 4.9: Sentiment distribution over aspect Product/Service/HQ (test set)

The test set also has reviews in six languages, namely, English, German, Dutch, Spanish, Italian and French. The table 4.4 below represents the number of reviews available for evaluating for each language and the figure 4.10 visualises this distribution over the training set.

Language	Number of reviews
English	104
German	7
Spanish	27
French	31
Italian	21
Dutch	56

Table 4.4: Distribution of test reviews over language

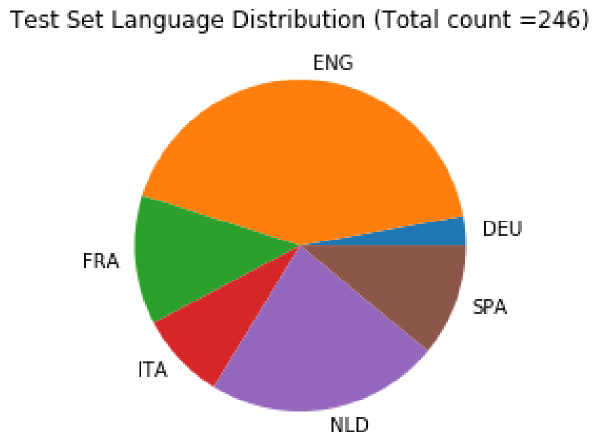


Figure 4.10: Distribution of test reviews over all languages

4.2 Problem Formulation

The problem for Experiment 1 is defined as a 5-class classification problem. Given a document/review D and an aspect A, predict the sentiment class Y from the set {‘positive’, ‘negative’, ‘neutral’, ‘none’}. The aspect A belongs to the set {’Customer Assistance Center’, ’Dealer/Retailer’, ’Product/Services/HQ’}. The document/review D is a customer review in any language in the set {English, German, Dutch, Spanish, Italian, French }. The proposed setup is relevant to performing and learning from SemEval Task 4’s subtask 3 (Aspect Category Detection) and subtask 4 (Aspect Category Polarity). It is important to note that for the model in experiment 2, the problem is 9-class classification problem. That is, given a document D predict class Y from a set of all <aspect, sentiment> pairs. There are 3 possible aspects and each of them can have 3 possible sentiments (excluding ’none’). Hence, a 9-class classification problem. Detailed description of problem statement and approaches follows.

4.2.1 Generation of auxiliary sentences

Aligned to the problem statement for experiment 1, the state of the art system, “Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence” [1] defines four ways of defining auxiliary sentences. The motivation behind generation of these auxiliary sentences is to utilize the full potential of the pre-trained BERT [] system to produce state of the art results by providing augmented data for the transformer to learn. Another thing to note would be that these auxiliary sentences are constructed with the intent to help in natural language inference and extract sentiments from all aspects present in a given document. Table 4.5 below shows the sentence that would be used as an example for the process of generating auxiliary sentences. This work focuses on two approaches delivering state of the art results, proposed by S. Chi et al., namely BERT-pair-NLIB (Natural Language Inference - B) and BERT-pair-QAB (Question Answering - B) to generate auxiliary sentences. The word "pair" is used represent that the task changes from single sentence classification to sentence pair classification with auxiliary sentences. The BERT-pair-NLIB delivered state of the art results on the task aspect category detection and BERT-pair-QAB outperformed all models on the task aspect polarity detection.

Document (D)	“Amazing service with regards to roadside assistance. 15 min waiting period only.”
Aspect Set	’CAC’, ’Dealer/Retailer’, ’Products/Services/HQ’
Aspect (A)	‘Dealer/Retailer’
Sentiment (Y)	"Positive"

Table 4.5: Example Document for CAC Dataset

The experimentation setup for experiment 1 would need the generation of auxiliary sentences for both the BERT model and the BERT multilingual model to benchmark the results for their comparison. The sentences will be generated in English and in the language of the review for this experiment, directly following the approach of authors of [1] propose. These sentences will act as a complement for a review to form sentence pairs for classification.

Sentence for QAB – A new sentence will be generated for a <aspect,sentiment> pair. This implies, a total of 12 possibilities (3 aspects, 4 sentiments), and hence 12 sentences. Each of these sentences couple with a review to form 12 training records from 1. The sentences that would be formed for the example in table 4.5 are, "The polarity of aspect Customer Assistance Center is positive.", "The polarity of aspect Customer Assistance Center is negative.", "The polarity of aspect Customer Assistance Center is neutral." and so on, for every <aspect, sentiment> pair. The label set for these records would be ’1’,’0’, ’1’ being the label when <aspect,sentiment> pair mentioned in the auxiliary sentence exists in the review. The records in the evaluation set are also used to generate new records for evaluating the QAB system.

Sentence for NLIB – Similar to QAB, this approach also generates 12 new records for training from one record in the training set. However, unlike QAB the sentence follows the format "aspect name - sentiment". Taking the example in table 4.5, the NLIB sentences for this record would be, "Customer Assistance Center -

positive", "Customer Assistance Center - negative", "Customer Assistance Center - neutral", "Customer Assistance Center - none" and so one for all aspects and sentiments. The classification for this approach also changes to '1','0' classification like in QAB. The records in the evaluation set are also used to generate new records for evaluating the NLIB system.

To answer RQ 1, the records in training and evaluation set are used to generate auxiliary sentences like QAB and NLIB, however for this case the language of auxiliary sentences is also that of the review it is paired up with. The two approaches are named **QAB-Lang** and **NLIB-Lang**.

4.2.2 Multi-Task Approach

The experiment for RQ 2 does not require the generation of auxiliary sentences. The data is processed to remove the 'none' sentiment label. Since, all records belong to at least one sentiment in 'positive', 'negative', 'neutral'. It is also a possibility that a record is labelled with two different sentiments for two different aspects. Hence, problem statement for experiment 2 and RQ 2 would be answered by setting up a multi-label 9-class classification problem trained on multi-task learning, since the system will learn to classify a record for an aspect and sentiment together.

4.3 Input Representation

The records in all datasets are transformed to create input for the BERT models. The model's input format is: [CLS] SeqA [SEP] SeqB [SEP], where [CLS] is the classification token and [SEP] is the separator token for the BERT pre-trained system. While training the BERT model without auxiliary sentences, there are not two sequences to input and hence the input format is set to [CLS] SeqA [SEP]. The length of the whole input can be up-to 512 tokens including all the special tokens ([CLS],[SEP]). The output for the [CLS] token or the pooler output is feed forwarded to a classification layer (output layer).

4.4 Experimentation and Evaluation Setup

This section describes the experimentation setup proposed to answer research questions. All setups use the input representation as described in section 4.3. The methodology for interpreting results from systems and evaluating these prediction results is also described in this section.

4.4.1 Baseline - Single Systems

Baseline systems are defined with both BERT and BERT-multilingual systems. To perform ABSA just with BERT models, a single BERT model is fine-tuned to classify a record as one of the possible four sentiments from {'positive', 'negative', 'neutral', 'none'} but only for one aspect. Hence to create an ABSA systems with single BERT models, *n* BERT-models are fine-tuned for *n* possible aspects. In this use-case, three BERT models are fine-tuned for three possible aspects. The output from these three systems is compiled to form final classification of the ABSA system. The interpretation technique is described further in this section. This system is referred to as **BERT-single**. Similarly, three BERT-multilingual models are fine-tuned to form another ABSA system. This system is referred to as **BERT-multilingual-single**. The performance of these systems is used as baseline for designed experiments. The performance evaluation technique of the systems is also described further in this section.

4.4.2 Experiment 1 - Pair and Pair Lang Systems

To successfully answer the research question, its sub questions need to be answered. Both pre-trained models are imported and fine-tuned with auxiliary sentences in English first. The auxiliary sentences are formed by the NLIB and QAB approach. The systems are developed by training two BERT models independently with the NLIB and QAB approaches. These systems are referred to as **BERT-pair-NLIB** and **BERT-pair-QAB** systems. The same setup is then devised with two BERT-multilingual models referred to as **BERT-multilingual-pair-NLIB** and **BERT-multilingual-pair-QAB**. All models are trained on a binary classification task as described in section 4.2.1.

To check how the change in language of auxiliary sentences affects the performance of the model, the models both BERT and BERT-multilingual are fine-tuned with (reviews and) auxiliary sentences in the language of review. Hence, two systems are developed by two BERT models trained with the NLIB-Lang and QAB-Lang approaches. Also, two BERT-multilingual models are trained with these approaches. These models are referred to as **BERT-pair-NLIB-Lang, BERT-pair-QAB-Lang, BERT-multilingual-NLIB-Lang and BERT-multilingual-QAB-Lang**. All pair models, namely, BERT-pair-NLIB, BERT-pair-QAB, BERT-multilingual-pair-NLIB, BERT-multilingual-pair-QAB, BERT-pair-NLIB-Lang, BERT-pair-QAB-Lang, BERT-multilingual-NLIB-Lang and BERT-multilingual-QAB-Lang are trained for a {'1', '0'} label set. Figure 4.11 represents the architecture for systems developed with BERT and BERT-multilingual in this experiment. A detailed comparison of performances of above mentioned models is presented in chapter5. The comparison of performances of these models will conclusively answer RQ 1.

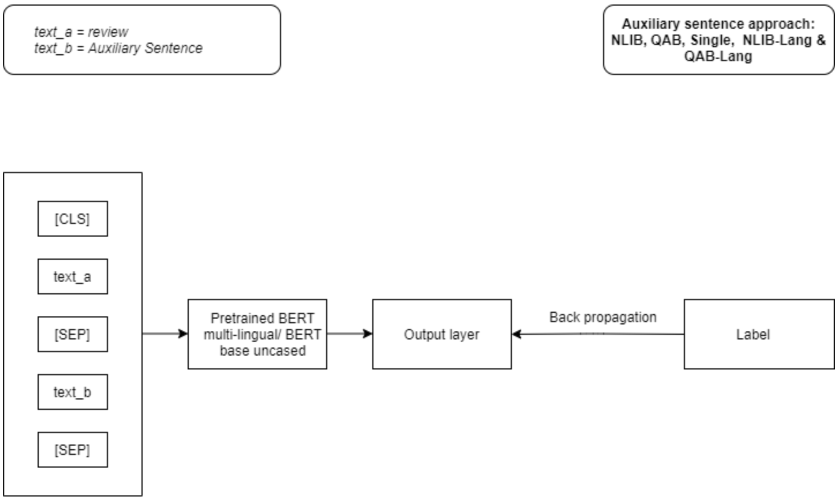

text_a = review
text_b = Auxiliary Sentence
Auxiliary sentence approach: NLIB, QAB, Single, NLIB-Lang & QAB-Lang
[CLS]
text_a
[SEP]
text_b
[SEP]
Pretrained BERT multi-lingual/ BERT base uncased
Output layer
Back propagation
Label

Figure 4.11: Architecture of BERT and BERT-multilingual systems - Experiment 1

4.4.3 Experiment 2 - Mutli Systems

To answer RQ 2, models are fine-tuned with multi-task learning with the BERT model and BERT-multilingual-model. To achieve this training approach, the pre-trained models are fine-tuned with nine output neurons that output the probability of an <aspect, sentiment> pair. This way the model learns about both aspect and sentiment at the same time but without auxiliary sentences. The ABSA systems with the BERT model and the BERT-multilingual model are referred to as **BERT-multi** and **BERT-multilingual-multi** respectively. Both the models are fine-tuned on the label set {’CAC-positive’, ’CAC-negative’, ’CAC-neutral’, ’Product-positive’, ’Product-negative’, ’Product-neutral’, ’Dealer-positive’, ’Dealer-negative’, ’Dealer-neutral’}. The comparison of performance of these two systems and those developed in experiment 1 is documented in chapter5. This comparison will answer RQ 2. The figure 4.12 represents the architecture of both the systems developed in this experiment.

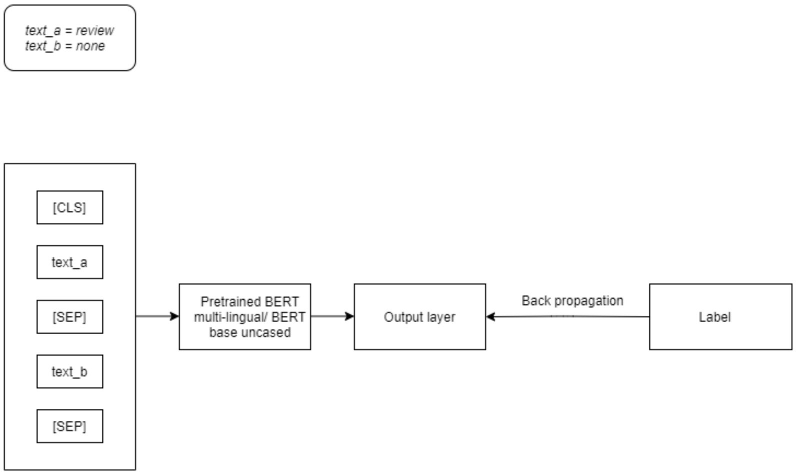


Figure 4.12: Architecture of BERT and BERT-multilingual systems - Experiment 2

4.4.4 Interpretation and Evaluation

Although all systems are being trained in different ways, they are trained to deliver on the task aspect based sentiment classification. Hence, evaluating them at the core definition of the task is imperative. The task of aspect based sentiment analysis has two characteristics for evaluation namely, aspect category detection and aspect polarity detection. Hence, all systems are evaluated on these two sub-tasks to answer the research question. To evaluate all systems on these characteristics and to the same standard, an interpretation methodology is devised to first interpret predictions of all the systems.

To interpret the prediction results for a test record first the prediction probabilities for all possible <aspect,sentiment> pairs is gathered, except for when sentiment = 'none'. The probability for 'none' sentiments is left out to strategically eliminate some foul cases discussed further. Also, every record belongs to at least one sentiment of an aspect and hence the final prediction need not consider 'none' sentiment probabilities. The maximum prediction probability for a particular aspect is then identified. For instance, the maximum probability for an outcome out of the outcomes {<'CAC', 'positive'>, <'CAC', negative>, <'CAC', 'neutral'>}. This action is performed for all three aspects namely, 'CAC', 'Products', 'Dealer'. If the probability is greater than 0.5, the record is labelled the corresponding <aspect, sentiment> pair. Interpreting the results in this manner avoids conflicting outcomes for example <'CAC', 'positive'> and <'CAC', 'negative'> and also includes two possible aspects detected for instance <'CAC', 'positive'> and <'Dealer', 'Negative'>. However, there exists a case when this interpretation technique fails to interpret any result. This is the case when all the maximum prediction probabilities identified are less than 0.5. In this case, the record is assigned the <aspect, sentiment> pair with the highest prediction probability provided that sentiment is not ’none’. Hence, eliminating the possibility of a record not to be classified for any <aspect, sentiment> pair. All records falling in this case are referred to as ’zero prediction records’. This interpretation technique takes into account the possibility of detection of multiple aspects and sentiments too as described above.

Once the predictions from all systems is interpreted, the prediction results are evaluated against the test set on the tasks aspect category detection (ACD) and aspect polarity detection (APD). To evaluate all systems, we consider the metrics of macro recall, macro precision and macro-f1 scores of classification. The macro-recall is defined as the sum of individual recall scores calculated for each target class. The macro recall is the fraction of correctly classified examples for a target class to the total number of examples for that target class whereas the macro precision is the ratio of correctly classified examples for target class to all classified examples for that class. The macro-precision is calculated by averaging precision values for all target classes. By nature and definition, recall and precision have a trade-off for a classification result. To maintain a standard, the harmonic mean of these metrics is used to finally evaluate systems. This metric is referred to as macro-F1 score. Macro scores are taken into account since the evaluation metrics provides equal importance to each individual test record independent of it’s actual class and it’s distribution in the test set. The equations 4.1, 4.2 and 4.3 below represent the calculation for these metrics.

$$\textit{Precision} = \frac{\textit{All Correctly Classified Examples for a target class}}{\textit{All Classified Examples for this class}} \tag{4.1}$$

$$\textit{Recall} = \frac{\textit{All Correctly Classified Examples for a target class}}{\textit{All Examples for this class}} \tag{4.2}$$

$$\textit{macro-Recall} = \textit{avg(Recall for all target classes)} \tag{4.3}$$

$$\textit{macro-Precision} = \textit{avg(Precision for all target classes)} \tag{4.4}$$

$$\textit{macro-F1} = \frac{\textit{2*macro-Precision*macro-Recall}}{\textit{macro-Precision + macro-Recall}} \tag{4.5}$$

These metrics are used to evaluate all systems on both tasks namely, aspect category detection and aspect polarity detection. For aspect category detection, the target label set is {'CAC', 'Product', 'Dealer'} and for aspect polarity detection, the target label set is {'positive', 'negative', 'neutral'}. The metrics since averaged in a 'macro' manner, provide equal importance to all target classes while evaluation. To evaluate all systems on the complete task of ABSA, a metric is proposed taking into account the micro-F1 score from both its children tasks in essence, aspect category detection (ACD) and aspect polarity detection (APD). The metric is referred to as 'ABSAEval'. It is calculated by taking the harmonic mean of micro-F1 scores of the both the sub tasks. Equation 4.6 represents the formula to calculate this metric.

$$\mathit{ABSAEval} = \frac{\mathit{2*macroF1(ACD)*macroF1(APD)}}{\mathit{macroF1(ACD) + macroF1(APD)}} \tag{4.6}$$

In the end, all models are evaluated on language level classification To carry out this comparison, language level classification accuracy is calculated for every model for each language. Equations 4.7 and 4.8 represent this metric for each language on the tasks ACD and APD respectively.

$$\textit{Language level accuracy ACD} = \frac{\textit{Correctly classified records for ACD for the language}}{\textit{Total number of records in that language}} \tag{4.7}$$

$$\textit{Language level accuracy APD} = \frac{\textit{Correctly classified records for APD for the language}}{\textit{Total number of records in that language}} \tag{4.8}$$

The metrics are calculated for all systems developed for both experiments and reported in the next chapter.

Chapter 5

Results and Discussion

This chapter reports results for experiments devised to answer the research question and discusses the reported results. All systems are trained on a single GPU for 5 epochs with a learning rate of 2e-5 with a batch size of 10 records. As mentioned in section 4.4.4, all systems are evaluated on the two sub-tasks namely ACD and APD. The results reported include microF1 scores, precision scores, recall scores, ABSAEval scores and number of zero prediction records for all models. The metrics will be used to compare the performance of all trained systems and answer the research question and its sub-questions provided there exists a statistically significant difference between the performances of the models. This is imperative as there is absence of multiple test sets to compare performances of the systems over multiple tests. Hence, the outcome of experiments could be an occurrence by chance. To eliminate this doubt, T-tests are performed over groups of models to answer sub-research questions. T-tests are performed to establish if the difference between two samples is statistically significant or exists by chance. Hence, to compare learning techniques, multiple system performances are taken as a single sample to establish significant differences in performances to answer RQ 3.

5.1 Baseline - Single Systems

The performance of BERT-single and BERT-multilingual-single are reported in tables 5.1 - 5.4.

Model	Precision	Recall	F1-Score
BERT-single	0.70	0.69	0.70
BERT-multilingual-single	0.65	0.67	0.66

Table 5.1: Aspect Category Detection - Baseline

Model	Precision	Recall	F1-Score
BERT-single	0.26	0.51	0.30
BERT-multilingual-single	0.30	0.50	0.31

Table 5.2: Aspect Polarity Detection - Baseline

Model	ABSAEval
BERT-single	0.42
BERT-multilingual-single	0.42

Table 5.3: ABSAEval - Baseline

Model	Number of zero prediction records
BERT-single	26
BERT-multilingual-single	59

Table 5.4: Number of zero prediction records - Baseline

These results are used to compare further trained systems. It can be observed from the results that for the ABSA task, both BERT model and BERT-multilingual model perform at similar levels since both have an ABSAEval of 0.42. The BERT model outperforms the BERT-multilingual model on ACD by 4 points of F1-score but the BERT-multilingual model outperforms the BERT model on APD by 1 point of F1-score. Both the models are used as baseline for further experiments and comparisons.

5.2 Experiment 1 - Pair and Pair Lang Systems

Table 5.5 compares the performance of all models described in section 4.4.2 on the task of aspect category detection and table 5.6 compares the performance on the task of aspect polarity detection. The table 5.7 presents the final ABSAEval metric calculated for all models. The presence of zero prediction records cannot be ignored and exists for all systems. The table 5.8 reports the number of zero prediction records for all systems out of a total of 246 records in the test set. Subsection 5.2.1 discusses the reported results to answer RQ 1 further in this section.

Model	Precision	Recall	F1-Score
BERT-pair-NLIB	0.37	0.37	0.36
BERT-pair-QAB	0.38	0.39	0.37
BERT-pair-NLIB-Lang	0.39	0.33	0.36
BERT-pair-QAB-Lang	0.42	0.33	0.36
BERT-multilingual-pair-NLIB	0.40	0.41	0.39
BERT-multilingual-pair-QAB	0.47	0.34	0.33
BERT-multilingual-pair-NLIB-Lang	0.44	0.47	0.44
BERT-multilingual-pair-QAB-Lang	0.42	0.42	0.41

Table 5.5: Aspect Category Detection - Experiment 1

Model	Precision	Recall	F1-Score
BERT-pair-NLIB	0.57	0.55	0.56
BERT-pair-QAB	0.57	0.54	0.55
BERT-pair-NLIB-Lang	0.56	0.54	0.55
BERT-pair-QAB-Lang	0.49	0.44	0.45
BERT-multilingual-pair-NLIB	0.54	0.53	0.54
BERT-multilingual-pair-QAB	0.49	0.49	0.49
BERT-multilingual-pair-NLIB-Lang	0.56	0.53	0.54
BERT-multilingual-pair-QAB-Lang	0.56	0.53	0.55

Table 5.6: Aspect Polarity Detection - Experiment 1

Model	ABSAEval
BERT-pair-NLIB	0.44
BERT-pair-QAB	0.44
BERT-pair-NLIB-Lang	0.43
BERT-pair-QAB-Lang	0.40
BERT-multilingual-pair-NLIB	0.45
BERT-multilingual-pair-QAB	0.39
BERT-multilingual-pair-NLIB-Lang	0.48
BERT-multilingual-pair-QAB-Lang	0.47

Table 5.7: ABSAEval - Experiment 1

Model	Number of zero prediction records
BERT-pair-NLIB	32
BERT-pair-QAB	136
BERT-pair-NLIB-Lang	158
BERT-pair-QAB-Lang	246
BERT-multilingual-pair-NLIB	56
BERT-multilingual-pair-QAB	230
BERT-multilingual-pair-NLIB-Lang	138
BERT-multilingual-pair-QAB-Lang	180

Table 5.8: Number of zero prediction records - Experiment 1

5.2.1 Pair vs Pair-Lang

In case of pair systems, the BERT-pair-NLIB works best with 0.44 as ABSAEval and the BERT-Multilingual-pair-NLIB-Lang works best with an ABSAEval of 0.48 for pair-lang systems. To compare the effect of changing the language of auxiliary sentences and answer RQ 1, all pair systems are compared. The table 5.9 compares the performance of these systems on the task of Aspect Based Sentiment Analysis presenting the ABSAEval metric. It can be observed that changing language of the auxiliary sentences to the language of the review increases the performance of the pair systems by 3 points considering best performing systems in both groups.

System type	Pair	Pair-Lang
BERT-NLIB	0.44	0.43
BERT-QAB	0.44	0.40
BERT-Multilingual-NLIB	**0.45**	**0.48**
BERT-Multilingual-QAB	0.39	0.47

Table 5.9: ABSAEval - Comparison of BERT-pair

From table 5.9, it can be observed both NLIB and QAB approaches perform differently in case of both groups of systems (BERT based systems and BERT multilingual based systems). It can also be observed that the ’lang’ approaches perform better with BERT multilingual based systems and outperform other systems. Here, a T-test is not feasible to identify a significant difference as the sample size of groups formed for comparison would be too small. This is because while comparing pair approaches to pair-lang approaches for the BERT based systems, both groups would get only 2 outcomes. The change in performance can be visualized by the figure 5.1.

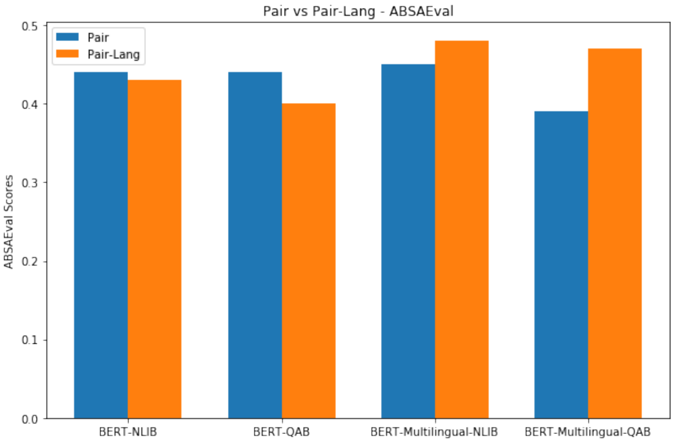


Figure 5.1: Pair vs Pair-Lang - ABSAEval

Reading figure 5.1, it can be observed there is always a change in performance of systems when the language of auxiliary sentences is changed to the language of the review. The performance increases with the BERT-Multilingual model and decreases with the BERT model due to this change. This suggests that the models have a better sense of sentence pair classification. Since the model was able to classify sentence pairs better when it could capture that, they are in the same language. The BERT-Multilingual model captures this fact that sentences in the same language are more likely to exist together which increases the performance of the system. The BERT model cannot identify this relationship between both the sentences as it is only pre-trained on English language datasets. This reason justifies the increase in performance that can be seen for auxiliary sentences in the language of review with BERT-multilingual as the base model and the decrease in performance of its counterpart.

Another point to note because of changing language of the auxiliary sentences is the difference of change of performance between different models. Considering the BERT based systems first, there is decrease in performance by 0.1 when comparing NLIB and NLIB-Lang approaches and 0.4 for QAB and QAB-Lang approaches. The reason for the decrease in performance has been described above. This decrease in performance is not similar because of the amount of augmented data. In case of all learning approaches considered (NLIB, QAB, NLIB-Lang and QAB-Lang), the size of the training set and test set increases 4-folds. However, the size and composition of the auxiliary sentences for both the NLIB and QAB approaches is different. There are extra 6 words in every auxiliary sentence for the QAB approach when compared to the NLIB approach. This suggests that although the QAB approach increases the amount of data available for training and improves performance by providing the model a sense of sentence pair classification, it hinders the model to learn better because of extra data present in the auxiliary sentences. This is also the reason behind the NLIB approach outperforming the QAB approach in all cases. Another way to look at it is that using the NLIB approach, the system was effectively learning about aspects and sentiments together in essence via multi-task learning. The extra data in QAB approach hinders the learning process and hence the QAB approach lacks in performance when compared to the NLIB approach. This is again the reason for the outlying behaviour or QAB and QAB-lang approaches observed in section 5.4. However, the decrease in performance because of 6 extra words per record in a training set that is 4-fold of its original size, should have been more than what is displayed by the systems trained with QAB approach. The reason why this high amount of extra data with no new information does not make the performance lag behind by a significant amount in some cases is because of the self-attention mechanism [21] of the transformer architecture on which both the base models are based. The mechanism helps the model to identify that the extra data is useless to some extent. However, it still blocks the complete learning potential as discussed in section 5.3.1. **In a nutshell, changing the language of auxiliary sentences to the language of the review increases the performance of pair-systems by 3 points on the ABSAEval metric for this use case, provided that the base model is able to capture this information. Also, it can be suggested that augmenting data with the help of auxiliary sentences might hinder the complete learning potentials of the systems.**

5.3 Experiment 2 - Multi Systems

The tables 5.10 and 5.11 compare the results of both the fine-tuned models (BERT-multi and BERT-multilingual-multi) on the test set for the tasks of aspect category detection and aspect polarity detection respectively. The table 5.13 records the number of zero prediction records (out of a total of 246 test records) for both these systems and table 5.12 records the ABSAEval score for both the systems. This section also includes a subsection (5.3.1), discussing results to answer RQ 2.

Model	Precision	Recall	F1-Score
BERT-multi	0.71	0.61	0.66
BERT-multilingual-multi	0.68	0.60	0.64

Table 5.10: Aspect Category Detection - Experiment 2

Model	Precision	Recall	F1-Score
BERT-multi	0.58	0.57	0.57
BERT-multilingual-multi	0.57	0.56	0.57

Table 5.11: Aspect Polarity Detection - Experiment 2

Model	ABSAEval
BERT-multi	0.61
BERT-multilingual-multi	0.60

Table 5.12: ABSAEval - Experiment 2

Model	Number of zero prediction records
BERT-multi	66
BERT-multilingual-multi	31

Table 5.13: Number of zero prediction records - Experiment 2

5.3.1 Multi vs Pair

In effect, all pair approaches learn about aspects and sentiments at the same time. This implies that all models are learning via multi-task learning but as observed in section 5.2.1, the learning is hindered by extra data in the auxiliary sentences. This can also be observed from table 5.14 where the multi models outperform all their counterparts. The difference between ABSAEval scores of the best performing multi system and the best performing pair system is of 0.13 or 13 points. The multi systems (systems trained on multi-task learning without auxiliary sentences) outperform pair systems and single systems for both the BERT model and BERT Multilingual model. The reason behind this is that the model still learns with multi-task learning as it targets output probability of every <aspect, sentiment> pair at the same time but without auxiliary sentences. This avoids the impediment generated by augmented data as discussed in section 5.2.1. Both the multi systems based on BERT model and BERT-Multilingual model perform similarly as discussed in section 5.4. **This concludes that the multi-task learning without auxiliary sentences outperforms the technique of sentence pair classification or multi-task learning with auxiliary sentences due to less irrelevant data hindering the learning process.**

5.4 BERT vs BERT-Multilingual

To answer RQ 3, we compare the performances of BERT and BERT-multilingual systems trained in both experiments using the ABSAEval metric. The table 5.14 represents this comparison. The table provides the ABSAEval metric of evaluation of all types of systems trained for both experiments. From the table, it can be observed that both the BERT and BERT-multilingual models deliver similar results. To statistically verify this, the independent samples t-test is conducted. This particular t-test is chosen as the systems trained on the BERT model and BERT-multilingual model are all trained in an independent environment. Hence, two samples are created namely, Sample 1 and Sample 2. Sample 1 is represented by the ABSAEval scores of all the systems trained on the BERT model and Sample 2 is represented by the ABSAEval scores of all the systems trained on the BERT-multilingual model. A confidence interval of 95% is initialized implying the alpha value for the test is 0.05. The degree of freedom is calculated by summing up total number of samples from both samples and subtracting 2. If the p-value associated to this degree of freedom and t-value is less than 0.05, the null hypothesis (that both samples are not statistically different) can be rejected.

System type	BERT Model	BERT-multilingual Model
single	0.42	0.42
pair-NLIB	0.44	0.45
pair-QAB	0.44	0.39
pair-NLIB-Lang	0.43	0.48
pair-QAB-Lang	0.40	0.47
multi	**0.61**	**0.60**

Table 5.14: ABSAEval - Comparison of BERT and BERT-multilingual

5.4.1 T-Test

Null hypothesis: There is no significant difference between Sample 1 and Sample 2.

Alternate hypothesis: There is a significant difference between Sample 1 and Sample 2.
Confidence interval = 95%

Sample 1: {0.42, 0.44, 0.44, 0.43, 0.40, 0.61}
Sample 2: {0.42, 0.45, 0.39, 0.48, 0.47, 0.60}
Degrees of freedom = 10 (6 examples in each sample => 6+6-2)

Mean of Sample 1 = 0.457
Standard deviation of Sample 1 = 0.070
Mean of Sample 2 = 0.468
Standard deviation of Sample 2 = 0.066

t-value = -0.270
p-value = 0.792 (» 0.05)
Since the p-value is very high, the null hypothesis can not be rejected.

From the t-test described in section 5.4.1, it can be seen that there is no statistically significant difference between the ABSAEval scores of BERT model systems and BERT-multilingual model systems with a confidence of 95%. Hence, it can be said that both models perform in a similar fashion on the task of Aspect Based Sentiment Analysis. However, this can also be the case if there exists a trade-off between performances on the sub-tasks ACD and APD and both types of systems perform differently on the sub-tasks. To verify if the performance on the sub-tasks is similar as well, the performances of both types of systems on the sub-tasks is compared in the table 5.15. The figures 5.2 and 5.3 visualize these metrics to compare both groups.

From figures 5.2 and 5.3 it can be seen that both groups have similar trends in both sub-tasks, except for QAB and QAB-lang systems. Also, there is no approach for which both the models perform in contrast when comparing their performance on APD and ACD. Hence, the non-significant difference between both Samples 1 and 2 can not be credited to the trade-off b/w the groups' performances on sub-tasks ACD and APD. The reason for the outlying behaviour of QAB and QAB-lang systems is discussed in section 5.2.1.

System type	BERT Model	BERT-multilingual Model
	ACD / APD	**ACD / APD**
single	**0.70** / 0.30	**0.66** / 0.31
pair-NLIB	0.36 / 0.56	0.39 / 0.54
pair-QAB	0.37 / 0.55	0.33 / 0.49
pair-NLIB-Lang	0.36 / 0.55	0.44 / 0.54
pair-QAB-Lang	0.36 / 0.45	0.41 / 0.55
multi	0.66 / **0.57**	0.64 / **0.57**

Table 5.15: macro-F1 (ACD/APD) - Comparison of BERT and BERT-multilingual

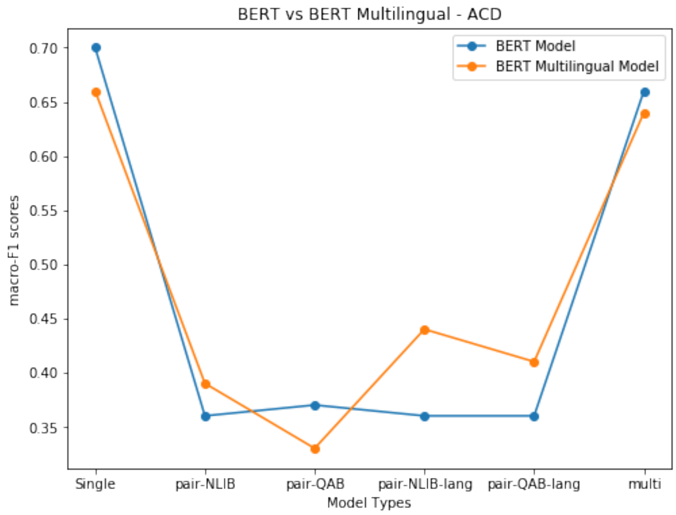


Figure 5.2: macro-F1 scores (ACD)

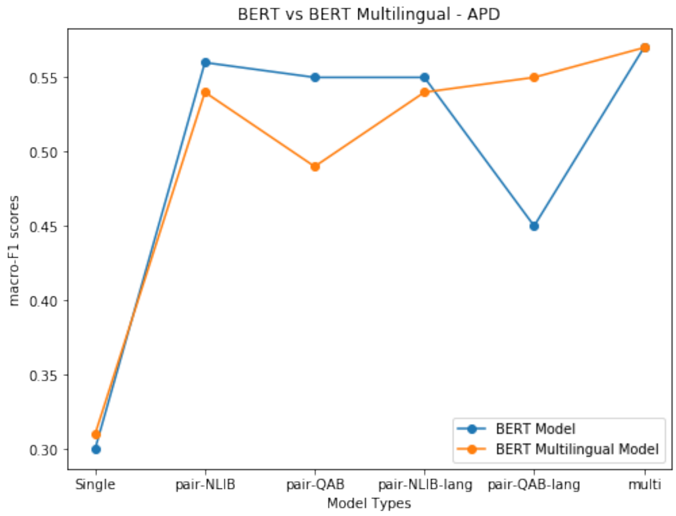


Figure 5.3: macro-F1 scores (APD)

The BERT-multilingual model was expected to outperform the English BERT model since it was expected to bigger word piece tokens of words in multilingual reviews as seen in the table 5.16, which represents the tokenization of some Italian words.

Words	“Servizio straordinario”
BERT Tokenization	[’ser’, ’vi’, ’zio’, ’st’, ’rao’, ’rdi’, ’nar’, ’io”]
BERT-Multilingual Tokenization	[’servizio’, ’str’, ’ao’, ’rdi’, ’nario’]

Table 5.16: Tokenization example

The reason why both the models BERT and BERT multilingual perform similarly can be credited to the cross-lingual abilities of BERT-multilingual. The paper 'How multilingual is Multilingual BERT?' [30], discusses how word piece tokenization helps BERT achieve crosslinguality along with multilinguality. It mentions that word piece tokenization takes place with the BERT-multilingual model as mentioned also in chapter 1. To achieve crosslinguality some pieces of the word are mapped to a shared space for all languages. However, the other pieces of the words need language specific data for the model to correctly make semantic sense. In this case, the BERT-multilingual model gets bigger tokens to train on. However, some of them are mapped to a shared space to achieve crosslinguality while the dataset does not contain a decent amount of language specific data for the multilingual model to make complete semantic sense. On the other hand, the BERT model achieves similar level of classification performance on the ABSA task. This can be credited to the BERT model's ability to generate syntactic and some semantic meaning from small tokens with the help of its bi-directional self attention mechanism [21]. A language level classification performance comparison is discussed in section 5.4.2.

5.4.2 Language level comparison

Tables 5.17 and 5.18 report the language level classification accuracy of all models for the task ACD and APD respectively. The tables need to be joined with the table 5.16 on column 'No.' to interpret results. It can be observed that the multilingual model outperforms the English model at language level. Also, it can be seen that the multilingual model captures the existence of two sentences in the same language with the 'lang' approaches and as observed in section 5.2 improves results at language level as well. Another thing to note can be that the multilingual model performs better on APD when compared to ACD. This hints that some tokens of sentimental words like 'positive', 'negative' and 'neutral', belong to the shared space of all languages enabling the model to classify records in all languages correctly for sentiments.

Hence, it can finally be concluded that both models perform similarly for the given use case. However, the BERT multilingual model is expected to perform better given more language specific data so it can make semantic understanding of words and tokens defining aspects in a record.

System	No.
BERT-single	1
BERT-pair-NLIB	2
BERT-pair-QAB	3
BERT-pair-NLIB-Lang	4
BERT-pair-QAB-Lang	5
BERT-multi	6
BERT-multilingual-single	7
BERT-multilingual-pair-NLIB	8
BERT-multilingual-pair-QAB	9
BERT-multilingual-pair-NLIB-Lang	10
BERT-multilingual-pair-QAB-Lang	11
BERT-multilingual-multi	12

Table 5.17: Model to Number Mapping

No./Lang.	German	English	French	Italian	Dutch	Spanish
1	0.43	**0.51**	0.52	0.38	0.41	0.41
2	0.71	0.16	0.48	0.19	**0.34**	0.56
3	0.71	0.21	0.45	0.19	0.39	0.52
4	0.71	0.13	0.42	0.05	0.45	0.52
5	0.71	0.17	**0.65**	0.05	0.45	0.52
6	0.57	0.41	0.52	**0.57**	0.46	0.52
7	**0.86**	0.32	0.48	0.29	**0.52**	0.48
8	0.71	0.17	0.39	0.24	0.45	0.48
9	0.71	0.23	0.42	0.10	0.45	**0.63**
10	0.43	0.19	0.48	0.0	0.45	0.45
11	0.71	0.13	**0.65**	0.05	0.34	0.37
12	0.57	0.44	0.48	0.52	0.45	0.56

Table 5.18: Language level classification accuracy - ACD

No./Lang.	English	German	French	Spanish	Italian	Dutch
1	0.71	0.29	0.35	0.14	0.27	0.19
2	0.71	0.74	**0.84**	0.71	0.84	**0.89**
3	0.71	0.71	**0.84**	0.76	0.84	0.85
4	0.57	**0.77**	0.81	0.71	0.77	0.81
5	0.57	0.61	0.81	0.71	0.59	0.85
6	0.43	0.75	0.81	**0.81**	**0.88**	0.85
7	0.57	0.28	0.29	0.10	0.29	0.19
8	**0.86**	0.70	**0.84**	0.71	0.79	0.81
9	**0.86**	0.57	0.77	0.67	0.80	0.63
10	**0.86**	0.68	**0.84**	0.76	0.84	0.85
11	0.71	0.73	0.81	0.76	80	81
12	0.86	0.73	0.77	**0.81**	0.84	**0.89**

Table 5.19: Language level classification accuracy - APD

5.5 Baseline Comparison

Taking the performance of single systems as baseline, we start comparing other systems. From table 5.14, it can be seen that the BERT-multi system outperforms all other systems with an ABSAEval score of 0.61. Both the single systems perform similarly with an ABSAEval of 0.42. The figure 5.4 represents the change in performance of both systems with the learning techniques in terms of ABSAEval taking best model (NLIB) for the 'pair' and 'pair-lang' categories.

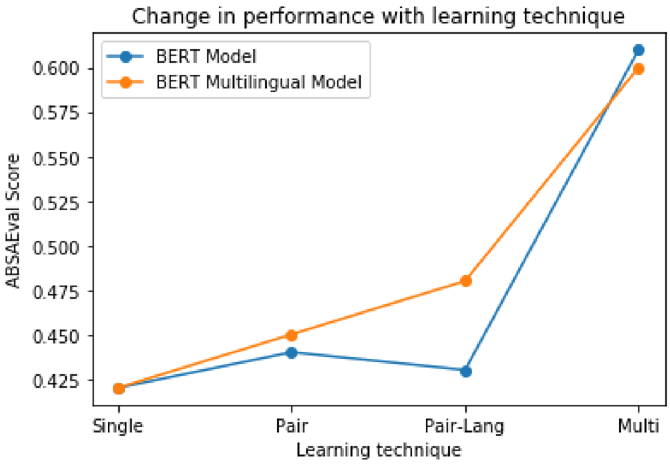


Figure 5.4: Change in performance with learning technique

It can be seen that the approach (pair) proposed by S. Chi et al. [1] improves the performance from baseline. The performance further improves as the language of auxiliary sentences is changed from English to the language of the input review (pair-lang) and performs best without auxiliary sentences(multi). The reason for increase in performance for pair approaches when compared to the baseline can be credited to the fine-tuning technique of sentence pair classification over single sentence classification as described by authors of [1]. The reason for increase and decrease in performance of BERT and BERT-Multilingual model moving from pair to pair-lang approach is because sentence pairs existing in same language help the models with better sentence pair classification, provided the model is able to capture this fact. Since BERT model (pre-trained only in English) does not grasp this concept for pair-lang approach, it's performance decreases while that of its counterpart increases. The reason why the multi models outperform all other models has been discussed in section 5.3.1.

Chapter 6

Conclusions and Future Works

6.1 Conclusions

The conclusions answer the research question and the sub-questions that were formulated in Chapter 1. To answer the research question the sub-questions are answered first. The answers to sub research questions are as follows:

RQ 1. To what extent does using auxiliary sentences in the language of the review improve the performance compared to using only English auxiliary sentences? From section 5.2, it can be concluded that using auxiliary sentences in the language of the review increases the performance of the system by 3 points (ABSAEval) as compared to using English auxiliary sentences. Also, it can be concluded that the increase in performance happens only if the base BERT model is able to capture the information that both sentences are in the same language. Hence, the BERT-multilingual model outperforms the general English BERT model when trained using sentence pair classification, and auxiliary sentences in language of the reviews. In addition, data augmentation with auxiliary sentences increases the systems' performances but also hinders their complete learning potential.

RQ 2. To what extent does training the system with multi-task learning and transfer learning without auxiliary sentences improve the performance of the system compared to using auxiliary sentences? From section 5.3, it can be concluded that training the system with multitask learning without auxiliary sentences improves the performance of the system by at least 13 points as compared to systems trained with auxiliary sentence. It can also be concluded that the authors of [1] provide state of the art results due to multi-task learning achieved by auxiliary sentences and sentence pair classification and not because of augmented data generated for training. Instead, the generated data hinders the learning potential of system.

RQ 3. To what extent can using a pre-trained multilingual BERT model improve the performance compared to using the general pre-trained BERT model? From section 5.4, it can be concluded that both the BERT model and BERT-Multilingual system work similarly and do not produce a significant difference in performance. However, given more language specific data, the BERT-multilingual is expected to outperform the English BERT model.

RQ. How can the state of the art approach be adapted to achieve the best performance on a multilingual dataset? Overall, the state of the art approach [1] proposed by S. Chi et al., can be modified to deliver better performance on the task of aspect based sentiment analysis for a multilingual use case. To achieve the best performance, the system is to be trained via multi-task learning without any auxiliary sentences, implying a system (with BERT multilingual model) trained to deliver probabilities of all the <aspect, sentiment> pairs for single sentence classification without auxiliary sentences.

6.2 Future Works

This section enlists the limitations of the conducted research and also suggests future adjustments to improve. The first limitation of the research is the size of the dataset. The number of reviews present in the dataset and their distribution with respect to languages is not ideal. It can be the case that BERT-multilingual [30] models might out-perform English BERT model which was not observed in this research credited to the small multilingual dataset. Hence, an experimentation could be set up to compare BERT-multilingual and English BERT models with a larger dataset. Another work might aim to generalize the results of this project. In other words, a BERT model with multitask learning can be trained without auxiliary sentences on SemEval 2014 dataset [2].

Bibliography

[1] S. Chi, L. Huang, and X. Qiu, “Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence,” 03 2019.

[2] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, “Semeval-2014 task 4: Aspect based sentiment analysis,” *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 27–35, 01 2014.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019.

[4] P. Rokade and A. D, “Business intelligence analytics using sentiment analysis-a survey,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, p. 613, 02 2019.

[5] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad Khasmakhi, M. Asgari-Chenaghlu, and J. Gao, “Deep learning based text classification: A comprehensive review,” 04 2020.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *Proceedings of Workshop at ICLR*, vol. 2013, 01 2013.

[7] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” *31st International Conference on Machine Learning, ICML 2014*, vol. 4, 05 2014.

[8] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *EMNLP*, 2014.

[9] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, “Deep unordered composition rivals syntactic methods for text classification,” 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)*. Beijing, China: Association for

Computational Linguistics, Jul. 2015, pp. 1681–1691. [Online]. Available: https://www.aclweb.org/anthology/P15-1162

[10] R. Stagner, “The cross-out technique as a method in public opinion analysis,” *The Journal of Social Psychology*, vol. 11, no. 1, pp. 79–90, 1940.

[11] M. Mäntylä, D. Graziotin, and M. Kuutila, “The evolution of sentiment analysis - a review of research topics, venues, and top cited papers,” *Computer Science Review*, vol. 27, 12 2016.

[12] S. Sandri, D. Dubois, and H. Kalfsbeek, “Corrections to "elicitation, assessment, and pooling of expert judgments using possibility theory",” *Fuzzy Systems, IEEE Transactions on*, vol. 3, pp. 479–, 12 1995.

[13] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, “Fasttext.zip: Compressing text classification models,” 2016.

[14] X. Zhu, P. Sobihani, and H. Guo, “Long short-term memory over recursive structures,” in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1604–1612. [Online]. Available: http://proceedings.mlr.press/v37/zhub15.html

[15] J. Cheng, L. Dong, and M. Lapata, “Long short-term memory-networks for machine reading,” *CoRR*, vol. abs/1601.06733, 2016. [Online]. Available: http://arxiv.org/abs/1601.06733

[16] P. Liu, X. Qiu, X. Chen, S. Wu, and X. Huang, “Multi-timescale long short-term memory neural network for modelling sentences and documents,” 01 2015, pp. 2326–2335.

[17] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, “Text classification improved by integrating bidirectional lstm with two-dimensional max pooling,” 2016.

[18] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1480–1489. [Online]. Available: https://www.aclweb.org/anthology/N16-1174

[19] Y. Liu, C. Sun, L. Lin, and X. Wang, “Learning natural language inference using bidirectional lstm model and inner-attention,” 2016.

[20] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, “Disan: Directional self-attention network for rnn/cnn-free language understanding,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11941

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.

[22] K. Schouten and F. Frasincar, “Survey on aspect-level sentiment analysis,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 813–830, 2016.

[23] E. Cambria and A. Hussain, *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*, 1st ed. Springer Publishing Company, Incorporated, 2015.

[24] Y. Wang, M. Huang, X. Zhu, and L. Zhao, “Attention-based LSTM for aspect-level sentiment classification,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 606–615. [Online]. Available: https://www.aclweb.org/anthology/D16-1058

[25] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, and Z. Wu, “Content attention model for aspect based sentiment analysis,” 04 2018, pp. 1023–1032.

[26] P. Chen, Z. Sun, L. Bing, and W. Yang, “Recurrent attention network on memory for aspect sentiment analysis,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 452–461. [Online]. Available: https://www.aclweb.org/anthology/D17-1047

[27] S. Zheng and R. Xia, “Left-center-right separated neural network for aspect-based sentiment analysis with rotatory attention,” 2018.

[28] O. Wallaart and F. Frasincar, *A Hybrid Approach for Aspect-Based Sentiment Analysis Using a Lexicalized Domain Ontology and Attentional Neural Models*, 05 2019, pp. 363–378.

[29] H. Xu, B. Liu, L. Shu, and P. S. Yu, “Bert post-training for review reading comprehension and aspect-based sentiment analysis,” 2019.

[30] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual bert?” 2019.

Appendix A

Appendix A - SemEval 2014 Dataset

The dataset is a public dataset named SemEval 2014 Subtask 4 data set. This data set is used to benchmark systems for aspect category detection and aspect sentiment detection. The data set comprises of customer reviews in English for a restaurant. There are two sets provided, one for training and one for evaluating the trained system. The training set contains a total of 3044 reviews each classified into multiple classes from the set ’food’, ’ambience’, ’service’, ’price’, ’anecdotes/miscellaneous’. The reviews also have labelled sentiment for an aspect provided that the review is classified for that aspect. The sentiment can be from the set ’positive’, ’negative’, ’neutral’, ’conflicting’. Hence, a review in total has 20 <aspect, sentiment> pairs possible in theory. However, a review can have only 1 out of the 4 sentiment classes for a given aspect. Therefore, each review would have 5 <aspect, sentiment> possibilities. The test/evaluation set contains 800 reviews labelled in a similar fashion. Both the sets are .xml files with a review id, review text, review aspect category, and review aspect category sentiment at each node. Both sets are traversed and data is extracted to form .csv files for both sets. The figure A.1 represents the distribution of data over different aspects in the training set.

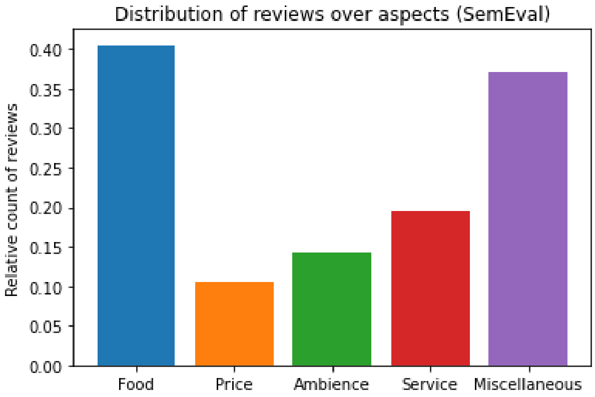


Figure A.1: Distribution of reviews over aspects (SemEval)

The figure A.1 shows that most of the reviews have either been labelled with 'Food' aspect or the 'Miscellaneous' aspect. The figures A.2, A.3, A.4, A.5 and A.6 represent what number of reviews belong to each aspect.

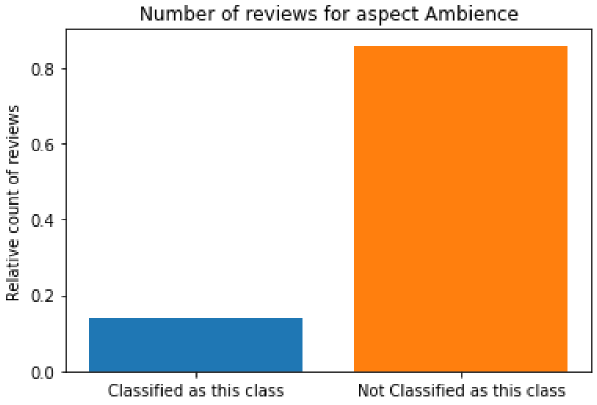


Figure A.2: Number of reviews for aspect Ambience

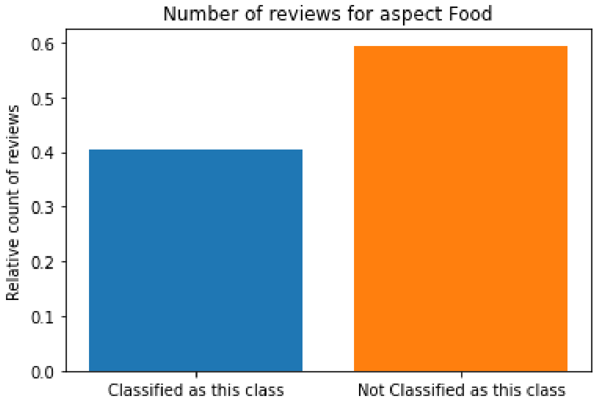


Figure A.3: Number of reviews for aspect Food

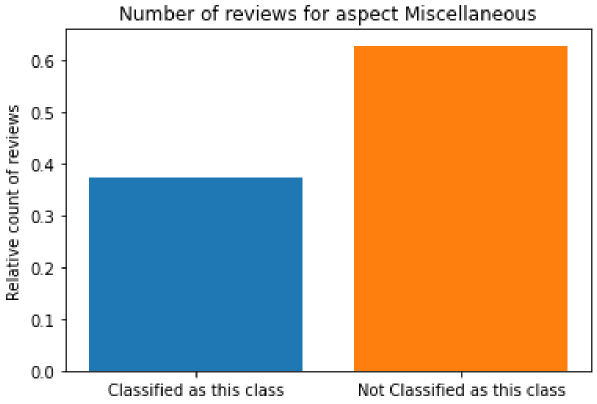


Figure A.4: Number of reviews for aspect Miscellaneous

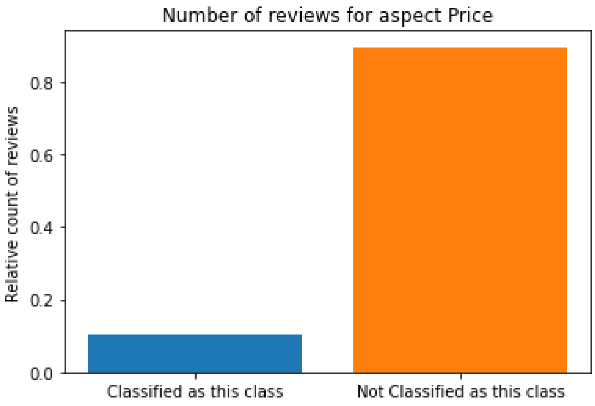


Figure A.5: Number of reviews for aspect Price

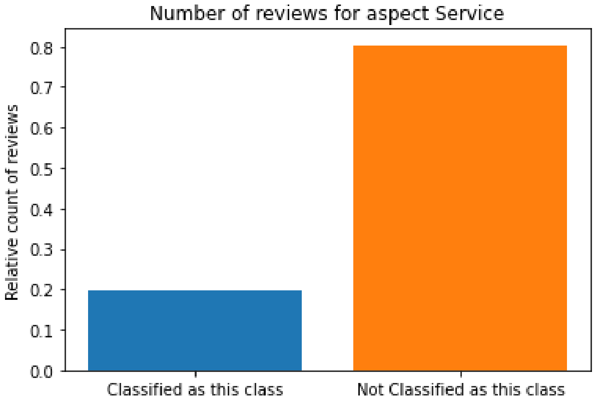


Figure A.6: Number of reviews for aspect Service

The table A.1 represents number of reviews belonging to each aspect and the labelled sentiment. Each aspect can have a sentiment from the set 'positive', 'negative', 'neutral', 'conflicting'. We also create a 'none' tag for a review that does not belong to a particular aspect.

	Food	**Service**	**Ambience**	**Price**	**Miscellaneous**
Positive	867	324	263	117	545
Negative	209	218	98	115	199
Neutral	90	20	24	10	357
Conflicting	67	35	47	17	30
None	1811	2447	2612	2725	1913

Table A.1: Distribution of reviews over aspect and sentiment

The figures A.7, A.8, A.9, A.10 and A.11 visualize the above table values for every aspect.

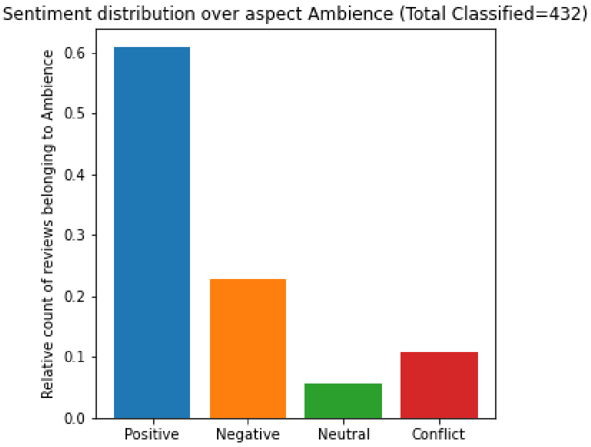


Figure A.7: Sentiment distribution over aspect Ambience

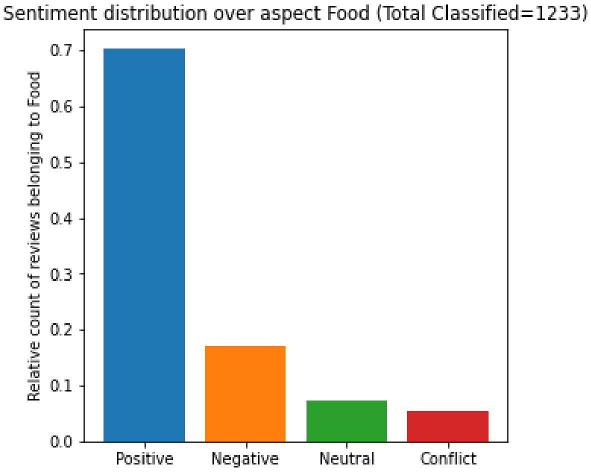


Figure A.8: Sentiment distribution over aspect Food

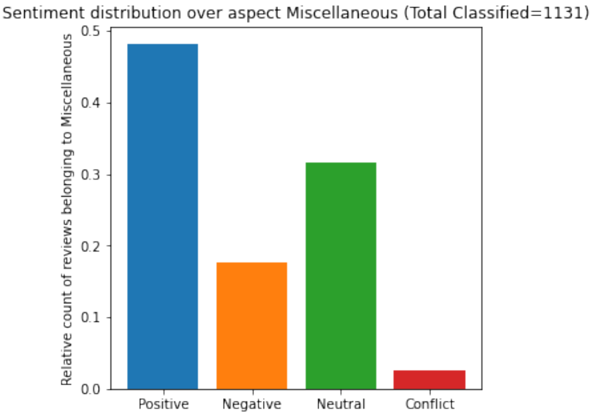


Figure A.9: Sentiment distribution over aspect Miscellaneous

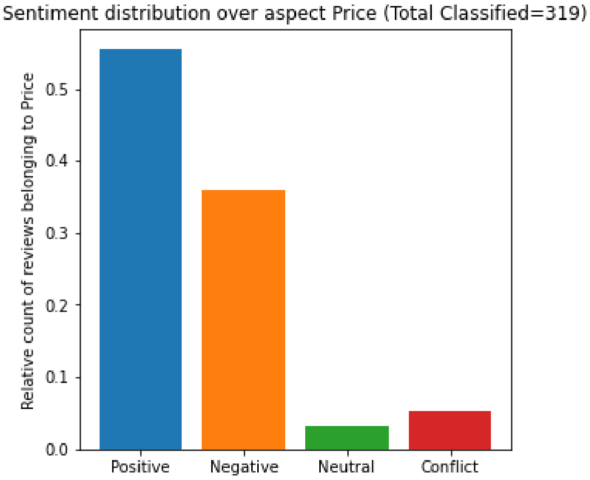


Figure A.10: Sentiment distribution over aspect Price

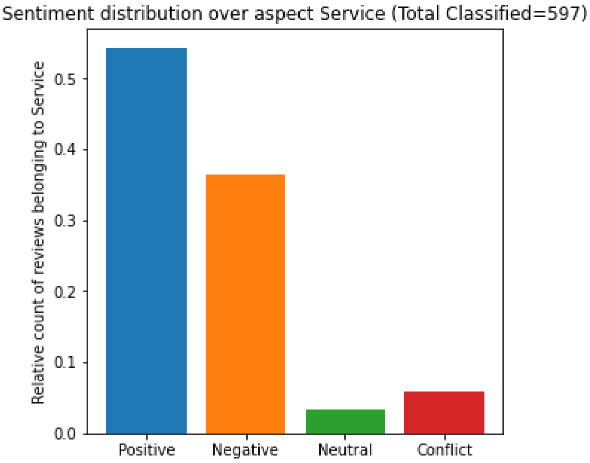


Figure A.11: Sentiment distribution over aspect Service

The figures visualize sentiment distribution and it can be seen that the sentiment ’conflict’ has the least number of records.

Appendix B - Technical Specifications

This chapter mentions the technical specifications used to carry out the project. The table B.1 enlists all specifications

Specification	Purpose
Python	Programming language
Numpy, Pandas	Data management libraries
Matplotlib, Seaborn	Data visualisation libraries
PyTorch	Model designing and training
Sklearn	Model evaluation
HuggingFace	Pre-trained models and tokenizers
Seed value = 22	Reproducing results
Learning rate = 2e-5	Optimizing model parameters
Warmup proportion = 1e-4	Optimizing model parameters
Optimizer = BertAdam	Optimizing model parameters
Epochs = 5	Training of models
Batch size = 10	Training of models
Machine = Standard NC12 by Microsoft Azure	Processing
nGPU = 1	Processing

Table B.1: Technical Specifications and Descriptions