Master Thesis

A Transfer Learning approach to Aspect Based Sentiment Analysis for airline customer feedbacks

Gabriele Catanese

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Isa Maks, Lotte van Bakel
2^{nd} reader: Roser Morante

Submitted: June 29, 2021

Abstract

This thesis project focused on the application of Transfer Learning to Aspect Based Sentiment Analysis for airline customer feedbacks data in English. Aspect Based Sentiment Analysis is a task that focuses on the detection of positive, negative and neutral opinions referring to a specific product or service. The retrievement of this information from the clients feedbacks provides an in-depth insight into the passengers satisfaction, serving as a key tool for an efficient improvement of the Customer Experience by the airline company that commissioned this work. As a method, two main steps were involved: the data annotation, where the domain specific ABSA annotation guidelines were put together and followed for the manual annotation of almost 2500 sentences; the classification, where two large language model, BERT (Devlin et al. (2019)) and RoBERTa (Liu et al. (2019b)), were adopted as Transfer Learning method and fine-tuned over the annotated dataset. Two separate classifiers, one for Aspect Category Detection, and one for Sentiment Polarity classification, were trained for each language model. A comparative analysis of the systems proved the superiority of the RoBERTa model (0.67 f1-score on Aspect Category Detection, and 0.83 accuracy on Sentiment Polarity classification). The importance of the methodology described in this thesis is underlined by the gap in related work on this domain, making this work a stepping stone for future research on Transformer-based approaches applied to airline customer feedbacks.

Keywords: Aspect Based Sentiment Analysis, CX, Airline data, NLP, BERT, RoBERTa, Transfer Learning

Declaration of Authorship

I, Gabriele Catanese, declare that this thesis, titled *A Transfer Learning approach to Aspect Based Sentiment Analysis for airline customer feedbacks* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 30-06-2021

Signed: Gabriele Catanese

Acknowledgments

I would like to thank my supervisors dr. Isa Maks and Lotte van Bakel for the constructive suggestions, guidance, and understanding that they provided me with during the entire duration of this thesis project. I would also like to thank Jet Kanters for her collaborative presence for the consolidation of the internship milestones.
I would like to extend my gratitude to team of Underlined for creating a stimulating and positive working environment, to the working students Marcell, Marloes, and Fenna, for annotating the data, and to the director, Theo van der Steen, for believing in the project and in me, by providing all the necessary resources for the success of this work.
I am also grateful to all the CLTL staff of the VU Amsterdam for the invaluable knowledge that they passed on during the course, and especially to my mentor, Pia Sommerauer, for her precious guidance when moving my first steps in this Master's program.
I would also like to acknowledge Marcell for the constructive feedback and for the stimulating exchange of ideas that contributed to the success of this project.
Finally, a special thanks to my family for the invaluable support, their love, and the life lessons that brought me here. To them I owe the realization of any of my ambitions.

List of Figures

Contents

Chapter 1

Introduction

In the past years, the customer experience (CX) field has been continuously developing new strategies in order to face the growing clients expectations. The continuous comparison between services and companies through digital means brought to a business concept where the successful business is not the most convenient, rather the one that adapts faster in relation to the opinion of its customers. Especially in these times where the spread of word can reach any angle of the world and millions of potential clients in a matter of seconds through the Internet, businesses started to realize the potential of AI as an efficient tool of reaction. Fighting technology with other technology became the norm, by efficiently addressing the consumers requests of improvements through cutting edge solutions, and reaching new standards of customer's satisfaction. The AI sub-field that focuses on linguistic data, Natural Language Processing (NLP), has been employed as one of the main methods in CX. This is due to the fact that customer feedback data does not come only in the form of surveys with multiple choice questions or stars ratings, but also through text. Reviews, tweets, and open comments are just some of the examples of the contents that companies try to analyze in order to extract useful information for their business goals. However, automatizing these processes becomes a priority, considering the always growing quantities of data in a digitalized world. For this reason, the design of scalable strategies, and the reduction of unsustainable methods which rely on costly resources, represent the winning card in the ever-changing customer oriented market.

1.1 Problem definition

This thesis project focused on the application of NLP techniques to CX in collaboration with an affirmed company in the field, Underlined[1]. In fact, their main goal is helping businesses increasing their customers satisfaction using data-driven solutions and text mining tools. Our collaboration involved the request of one of their clients, a famous Dutch airline company, consisting in the development of a tool capable of extracting useful information from English customer feedback data. This data contained the opinions of the passengers in open text comments left in a post-flight survey.
Considering that among the active researches at Underlined, the one related to Emotion Mining gained increasing attention over the past years, a study related to that

[1]Underlined: `https://underlined.nl/`

field would have been an interesting path to explore. Although the data showed to not contain explicit emotions, the presence of opinions and sentiment was especially clear. For this reason, the direction of the project went to Aspect Based Sentiment Analysis (**ABSA**).

Aspect Based Sentiment Analysis is a related task to emotion mining that focuses on the detection of positive, negative and neutral opinions referring to a specific product or service. The information that need to be retrieved for this task consist of the Aspect Term (AT), the Aspect Category (AC) and the Sentiment Polarity (SP). Following the definition of Pontiki et al. (2014), the Aspect Term Extraction involves the identification of a single or multiword term contained in the sentence that provides the information about the category that is object of opinion. The Aspect Category Detection focuses on the recognition of one of the predefined categories which is receiving a judgment in the sentence. Finally, the Sentiment Polarity classification consists in assigning a polarity label that is usually selected between "positive", "negative" or "neutral", by identifying it through attitudes, emotions etc. of the customer towards the aspect term within the sentence. The retrievement of this information from the clients feedbacks would provide an in-depth insight into the passengers satisfaction, serving as a key tool for an efficient improvement of the service.

Nevertheless, some challenges need to be faced when dealing with a task like this. The first is the unavailability of annotation on the dataset provided by the client. This involves a further annotation step to be carried out before a machine learning system can be trained and evaluated. Related to this point, there is also the tight time and resources restrictions of this thesis project, making the consideration of an extensive manual annotation process unlikely. Finally, the extreme domain specificity of this task makes the example of related work, mainly focused on hospitality and tech products reviews, irrelevant for a direct comparison of results. This is especially true for the Aspect Category sub-task, as the services provided by a company are extremely domain dependent.

On the other hand, the potential benefits involved in the successful outcome of the project served as a motivation for the continuation of the research. In fact, by conducting a study regarding ABSA on this dataset, the fine-grained details upon the customer experience were primarily taken into consideration, where not only the sentiment of the feedback, but also the related aspects are retrieved. This would allow the company to observe and act effectively on its reputation among the consumers. Furthermore, the restricted availability of resources incentivized the development of a method that would have allowed a successful outcome with small amounts of manually annotated data. In this way, considering the possibility of saving time and money on this costly process, the design of a more sustainable solution was involved. Following this direction, the exploration of state-of-the-art NLP tools becomes the obvious choice, as recent studies demonstrated that fine-tuning large language models, like BERT (Devlin et al. (2019)), as a Transfer Learning technique, leads to efficient performances even with the use of relatively small amounts of annotated data. Therefore, the implementation of such methods might demonstrate the potential of new technologies applied to real life business scenarios. Finally, the adoption of these premises on a domain that has been poorly explored by related work might serve as a stepping stone for further research

concerning feedback data on airline companies.

1.2 Research question & Solution

Taking into account the mentioned challenges and potential benefits, the Research Question was formulated as follows:

Can Transfer Learning, applied to a small set of annotated data, be a solution for domain-specific Aspect Based Sentiment Analysis?

To answer this question, a solution consisting of two main steps was designed as represented in Figure 1.1.

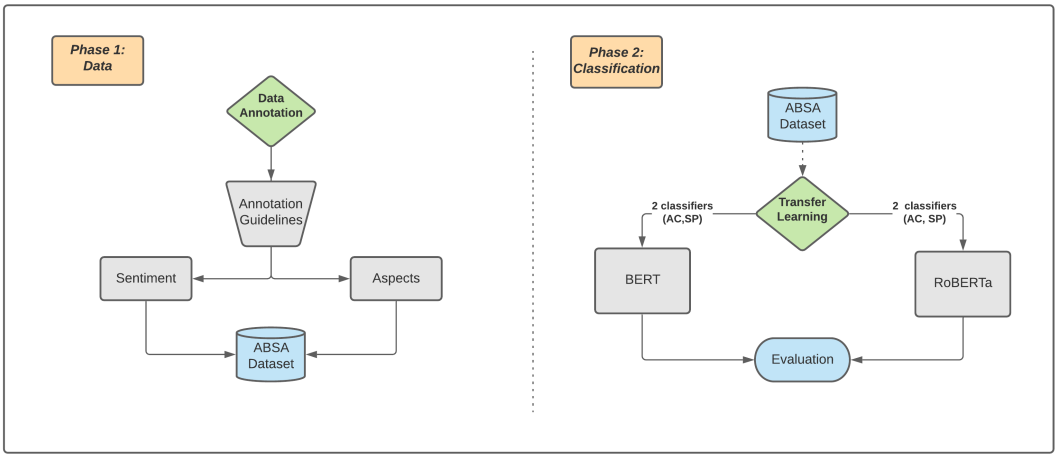

Figure 1.1: Solution flow of this Thesis Project

The first phase focuses on the data annotation. Here, the domain specific ABSA annotation guidelines are put together, where the instructions for the annotation of Aspects and Sentiment are provided. After the manual annotation process, a dataset consisting of almost 2500 sentences is created.
In the second phase, the focus is on the classification. As a Transfer Learning method, two large language model, BERT (Devlin et al. (2019)) and RoBERTa (Liu et al. (2019b)), are adopted and fine-tuned over the annotated dataset. Two separate classifiers, one for Aspect Category Detection, and one for Sentiment Polarity classification, are trained for each language model. Finally, a comparative analysis of the systems is carried out to evaluate which is the best performing.
The importance of the methodology described above stems from the fact that related work on the airline domain focused on traditional machine learning or deep learning approaches applied to customer reviews or tweets. On the other hand, this project concentrated on a Transformer-based approach, which to the best of my knowledge was never applied before on customer feedbacks of this domain. This type of user-generated data differs from reviews and tweets in the fact that the comments were requested by and addressed only to the company in a post-flight survey, with the improvement of

the service as a goal. Customer reviews and tweets, instead, are spontaneous, they do not share the same goal with the previously mentioned texts, and are addressed to an audience consisting of other potential consumers. Therefore, the content and the topics differ between each other, making the considerations present in related work less relevant, and underlining the importance of further study in this field.

1.3 Outline of the chapters

The discussion over each of the steps involved in this work is structured as follows. **Chapter 2** provides a literature review over the past and recent approaches to Aspect Based Sentiment Analysis. **Chapter 3** focuses on the analysis of the data and the annotation study. **Chapter 4** describes the classification method and the design of the systems. **Chapter 5** provides a comparative evaluation of the systems, and the analysis of the obtained results. Finally, **Chapter 6** contains the conclusions on the project and a discussion over future directions.

Chapter 2

Related Work

This chapter provides a literature review on different past and recent approaches to Aspect Based Sentiment Analysis. Here are included works that address ABSA as a task, focusing mostly on Aspect Category Detection and Sentiment Polarity, but not directly the domain which is object of this thesis. In fact, the related work that is currently available on the airline domain is scarce and non-existent for Transformer-based approaches. This factor makes this work a stepping stone for future research.

2.1 The Task

Aspect Based Sentiment Analysis is defined as a classification task. Text classification consists in assigning a class label to a single or a sequence of linguistic units. This process can be carried out on a document level, sentence level, phrase level, word level and so on. The nature of text is difficult to interpret for machines as it is unstructured, as opposed to structured data, which is organized in tables or similar formats. In order to retrieve useful information, the text needs to be splitted into individual words (or tokens) that are usually called features. These features are represented through vectors, which are numerical representations. The correlation between those vectors is interpreted by the machine to categorize the text. In order to enable the machine to interpret the text representation, different method were developed. The earliest approach was the definition of handcrafted rules that require high linguistic knowledge and resources. Later, rule-based systems started being supported or replaced by Machine Learning techniques, which were developed with scalability in mind. In fact, rule-based systems had the limitation of being restricted to a certain domain or task, while machine learning overcame this through probabilistic and statistical algorithms. These work by receiving a set of training examples that consist of an encoded representation of the features and their corresponding label. The information extracted from this representation is then learned by assigning probability scores to combinations of features and labels. The acquired knowledge is finally applied to perform a classification over a test set which was unknown to the system.

In the case of ABSA, the classification involves the Aspect Term (**AT**), the Aspect Category (**AC**) and the Sentiment Polarity (**SP**). Following the definition of Pontiki et al. (2014) for the International Workshop on Semantic Evaluation (**SemEval**) Task 4 of 2014, AT Extraction (also called opinion target extraction or **OTE**) involves the

identification of a single or multiword term contained in the sentence that provides the information about the category that is object of opinion. The AC Detection (**ACD**) focuses on the recognition of one of the predefined categories which is receiving a judgment in the sentence. The SP classification consists in assigning a polarity label that is usually selected between *positive*, *negative* or *neutral*, by identifying it through attitudes, opinions, evaluations, emotions, or feelings etc. of an opinion holder towards the aspect term within the sentence (Pontiki et al. (2014)). Given the example:

"The snacks that were distributed during the flight were really tasty!"

"*the snacks*" is the **AT**, *Food* is the **AC**, and *Positive* is the **SP**.

It needs to be specified that, over the years, the ABSA task as a whole has been redefined multiple times and researchers have been also experimenting with different interpretations and approaches to it. On the other hand, a common element to most of the related work is the use of the same electronic devices or restaurant review dataset provided for the SemEval (2014-2016). In these, each sentence contained one or multiple AT and related AC. Benchmark dataset were also released for the same domains. Nevertheless, there is still a huge gap in research concerning different domains or datasets.

2.2 First approaches

Early approaches to the task involved the use of handcrafted rules to perform OTE, large linguistic databases for ACD and sentiment lexicons for SP. Hu and Liu (2014) adopted a technique called "association mining" which consisted in identifying aspects through frequency of words or sequences of words. The assumption is that aspects would usually be expressed with recurrent expressions. When no frequent patterns are recognized, the closest noun is classified as aspect. For what concerns SP, a sentiment lexicon extended through WordNet was created for sentiment classification. Their association mining technique achieved 0.8 precision and 0.72 recall on a digital products customer review dataset.
On the base of this work, a different application of WordNet to ABSA is found in the system of Carenini et al. (2005). WordNet is a well known lexical database for English that organizes words into synonym sets, where also the dichotomy hyponymy/hypernymy and meronymy/holonymy is used to represent relatedness of words under the same lexical concept (Miller (1995)). In the work of Carenini et al. (2005), WordNet is used over the output of the unsupervised method of Hu and Liu (2014) to retrieve the AC through lexical similarity. For the evaluation, given the uniqueness of this work in which custom dataset and metrics were employed, making a comparative analysis with other systems becomes rather complicated.

The International Workshops on Semantic Evaluation (SemEval) 2014 Task 4 (Pontiki et al. (2014)) was a turning point for the research on ABSA. The NLP community started giving new importance to the advancements on this field, in fact, the competition received 163 submissions from 32 international teams. The datasets used for this occasion, consisting of laptop and restaurant customer reviews, represent still the main

resources for related studies. Furthermore, the standard metrics for efficiency evaluation were established as: precision (P), recall (R), f1-score ($F1$) and accuracy ($Acc.$).

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2PR}{P + R}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

TP for True Positives and TN for True Negatives are the correct prediction for both positive and negative cases; FP for False Positives are the prediction that are identified by the system and not by the annotators; FN for False Negatives for the opposite case, hence, when the human annotated class is not detected by the system. As described in Jurafsky and Martin (2020), the percentage of the elements that the system detected that are in fact positive is calculated by Precision; the percentage of elements actually present in the input that were correctly identified by the system is measured by Recall. Finally, the F1-score is a weighted harmonic mean of precision and recall. The F1-score is usually preferred for OTE and ACD, and accuracy for SP.
Brun et al. (2014) participated to the SemEval 2014 with a system using both handwritten rules and machine learning to extract useful input features. Fundamental components were Part of Speech tagging, chunking, Named Entity Recognition and syntactic parsing. In order to retrieve the AC, semantic similarity with a list of domain specific words extended through WordNet and food related terms from Wikipedia was calculated and classified through a Logistic Regression algorithm. A minimum probability threshold of 0.25 was set to avoid misclassification when multiple categories were present in the sentence. The SP was classified through lexicon and assigned separately to the AT and AC following the SemEval 2014 guidelines. The result achieved by this system for ACD was 0.82 f1-score; it also reached a 0.78 accuracy for Category Polarity and 0.77 Term Polarity. Another successful participant to the SemEval 2014 was the system presented by Kiritchenko et al. (2014). It used five binary one vs-all Support Vector Machine classifier over each of the predetermined AC. The input features included stemmed tokens, different types of ngrams, and cumulative scores calculated through a domain-specific lexicon for each term related to a category that appeared in the sentence. They achieved the best performance with 0.88 f1-score on ACD. The results on SP were 0.82 accuracy for Category Polarity and 0.78 for Term Polarity.
A common technique used in the research related to this task is topic modeling, with Latent Dirichlet Allocation (LDA) as main method. It has been applied in many works as in Poria et al. (2015) and García-Pablos et al. (2018). The reason behind this choice is that through statistical clustering, it is possible to identify both aspect terms and

group them into categories. On the other hand, as mentioned in Schouten and Frasincar (2016), the effective separation of categories into fine-grained ones and the need of human supervision on categorization is a downside of this method.

2.3 Neural Network Approaches

Following the previous edition, the SemEval 2015 Task 12 (Pontiki et al. (2015)) provided an extended dataset, with the addition of a hotel reviews set, and a redefinition of the annotation guidelines. It received 93 submissions from 16 international teams. It showed a predominance of features-heavy systems based on traditional machine learning algorithms for ACD, like the second best performer using a Maximum Entropy model (Saias (2015)), and sentiment lexicon based approaches for SP. Over the years, traditional machine learning proved to be an effective solution for the ABSA; however, it also showed its limitations, among which the need of large amounts of annotated datasets, high requirements of linguistic knowledge for time consuming features extraction and low reproducibility of results on different domains (Fernández-Gavilanes et al. (2016)). The advent of deep learning tried to overcome some of these limitations.
Neural networks (NN) are deep-learning models that were created with the aim of imitating the architecture of the brain. In fact, a representation of neurons called "nodes" are grouped in "layers" which are connected between each other in a network and activated when receiving an input signal. After computing a classification on it, consisting in multiplying each input by its weight, perform a sum over them and applying a non-linear function to the result, the individual nodes send an output to the following layer (Goldberg (2015)).

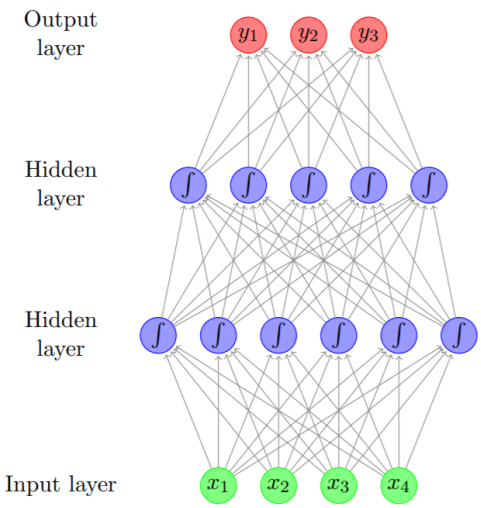

Figure 2.1: Feed Forward Neural Network. Source: Goldberg (2015)

The above image is a good representation of a simple feed-forward neural network (FFNN) where a the input x_n is fed to the model, processed by the hidden layers and provided as output y_n. In depth research over the potential of NN showed that,

depending on the tasks, some architectures might be preferable over others. As described in Goldberg (2015), the use a non-linear functions (e.g. sigmoid function) in fully connected networks (e.g. the multi-layer perceptron) can be used as a solution in binary and multiclass classification problems. Convolutional neural networks (CNN) demonstrated high efficiency when strong local clues regarding class membership are expected. This becomes helpful in a task like sentiment analysis, as the model creates a ngram representation of the input that preserves the order information providing a more accurate interpretation of the text. On the other side of the spectrum, a different type of architecture, the Recurrent networks (RNN), proved to be successful in a variety of tasks and, especially, in language modeling. This is a complex architecture that allows inputs of arbitrary size, preserving long-distance dependencies and context. For this reason, RNN are usually preferred in sequence classification or sequence-to-sequnce tasks (e.g. machine translation) overcoming the limitations of the Markov assumption. This architecture represented also the basis for the development of state-of-the-art Transformer-based language models like BERT (Devlin et al. (2019)) and RoBERTa (Liu et al. (2019b)) that will be described in detail in Chapter 4. Although, a number of deep learning systems for ABSA made use of big sets of handcrafted features, the advantage of these models is that they are able to capture linguistic information (e.g. structure, contex) directly from the preprocessed input, allowing also a greater generalization power. This is also possible thanks to the different input representation, the so-called "word embeddings". Sparse one-hot vector representations, like the Bag of Words, were discovered to not be suitable for NN because of their high-dimensionality consisting of many zero values. Furthermore, the order information was not preserved, creating a lack of structural and semantic information. Word embeddings overcame these limitation through a dense vector representation of words based on the concept that related words might share the same vector space. From here the popular statement "you shall know a word by the company it keeps" (Firth (1957)). These word representations are obtained through a training process that makes use of algorithms like word2vec (Mikolov et al. (2013)), GloVe (Pennington et al. (2014)) and fasttext (Bojanowski et al. (2017)) over large linguistic corpora. Through the training process, context and syntactic information is preserved and a greater generalization power is enhanced. On the other hand, there are also downsides in the use of NN. Just to mention a few, the necessity of high computational power to perform their complex operations and the need of large sets of data to enable learning of correlation between features.

The first deep learning model to be ranked as top performer of the SemEval was the feed forward neural network of Toh and Su (2015). They used a set of feature-heavy one-vs-all binary sigmoidal classifiers, with one FFNN classifier for each AC. Their system achieved 0.50 and 0.62 f1-score for respectively laptop and restaurant domain in ACD.

Ruder et al. (2016) developed a system based on a CNN using 300-dimensional GloVe Common Crawl word embeddings for both ACD and SP, obtaining top performances on the SemEval 2016 Task 5 (Pontiki et al. (2016)). They reached a f1-score of 0.68 on restaurant domain and 0.45 on laptop ACD. The accuracy for SP was of 0.82 for the restaurant domain and 0.78 for the laptop one.

State-of-the-art results were achieved by Sun et al. (2019) by using BERT (Devlin

et al. (2019)) in a rather creative way and seeing the ACD classification as a Question Answering (QA) or Natural Language Inference (NLI) task. In their auxiliary sentence approach, BERT's ability to receive a pair of input sentences was leveraged in the fine-tuning process and made comparative analysis with other solutions previously adopted. The SP classification was carried out through three binary BERT classifiers (one for each polarity). They achieved 0.92 f1score for ACD and 0.95 accuracy for SP on the SemEval 2014 dataset, proving that either their method or even the simple use of BERT as single sentence classification, brought improvement on state-of-the-art results.

2.4 Airline domain

As mentioned previously, there is a scarce availability of related work on ABSA for the airline domain. Furthermore, to the best of my knowledge, none of the works currently published directly address airline specific customer feedback data as it is treated in this thesis project. In fact, previous research in this field mostly focused on airline tweets, which are a good example of user generated text, but lack the specificity of customer review related to the experience with the company and its services. The receiver is also different in this case, because in the case of a survey the message is addressed directly to the company for future improvement, while in a social media environment the message is mainly directed to a digital "audience". Nevertheless, it is undoubtably interesting to look at approaches like the one of Ashi (2019), to get an insight into a work that addressed the airline domain and the conclusions that were drawn out of it. The research addressed by this paper focused on airline domain tweets in Arabic. A manual annotation of SP (two classes, negative and positive) and AC (13 classes identified by an observation of the dataset) was carried out on a set of 5k tweets. The following table (Table 2.1) gives an overview of the AC that were identified and that also inspired some of the categories of this thesis project.

Aspect	**Tweets topics**
Schedule	Flight schedule, rescheduling, timing, delays
Destinations	Airline destination and routes
Luggage & cargo	Luggage, air cargo, luggage allowance, luggage delays
Staff & crew	All staff such as pilots and flight attendants
Airplane	Airplane seating, cabin features, maintenance
Lounges	First-class and frequent flyer service and airport lounges
Entertainment	In-flight entertainment, other media, and wi-fi
Meals	In-flight meals and in-flight services
Booking services	Airline website, mobile app, and self-service machines
Customer service	Customer communications and complaint management
Refunds	Ticket refunds and compensations
Pricing	Ticket pricing and seasonal airline offers
Miscellaneous	Represents all tweets with gratitude or complaints about the airline as general with no relation to other 12 aspect categories

Table 2.1: Aspect categories from Ashi (2019)

The systems that were implemented for the tasks used a SVM algorithm combined

to 300 dimensions fastText (Bojanowski et al. (2017)) Wikipedia word embeddings. They achieved high results on both tasks with a 0.79 f1-score on ACD and a 0.89 accuracy on SP.

The project that is presented in this thesis is an experimental approach to the task, because apart from addressing airline customer feedback data, it also focuses on a Transformer-based method that has rarely been applied in related work treating this domain.

Chapter 3

Data & Annotation Study

This chapter contains all the details regarding the data that was used for this project. It gives an overview on its characteristics, the annotation guidelines that were followed and, finally, some statistics related to its content.

3.1 Characteristics

The dataset provided by the client consisted of a .csv file of 175.626 rows each containing a customer feedback with related information about the flight and the passenger for a total of 32 columns. The feedbacks consisted of English open text comments that were written at the end of a survey by each passenger after the flight.

As we might expect, these comments present the typical linguistic features of user generated text. Hence, they contain typos, incorrect use of punctuation, grammar mistakes, colloquialisms and emojis. However, unlike social media text (e.g. tweets), the receiver of the message is more specific, the comment is not spontaneous, but it is requested by a survey, it contains explicit opinions on determined aspects and it has their improvement as a goal. For what concerns the domain, it is strictly related to the airline subject and flight-related topics, unlike most of the related work on ABSA. In fact, the research on this field usually focused on the data provided by the SemEval, which contained reviews concerning restaurants, hotels and laptops. This will have important consequences also in the selection of a relevant set of AC, as they can be completely different depending on the domain (the details of this step will be discussed in the next section). This also entails that a direct comparative analysis with other systems for ABSA might be problematic, if not completely impossible.

In order to prepare the data for the annotation and the classification task, a smaller set was selected and each feedback was splitted into sentences using spaCy (`https://spacy.io/`). Finally, one set of 50 sentences and 3 different sets of 1000 sentences (one for each annotator) were created. This sets were formatted as .csv files and looked as in Table 3.1.

The choice behind the relatively small number of sentences composing this dataset is motivated by the tight time availability for the annotation and development process for this project. On the other hand, this also suggested the possibility of exploring the

Sentence_ID	Feedback_ID	Sentence	Aspect_Category	Sentiment	Aspect_Term
1	129288273	The snacks that were distributed during the flight were really tasty!			

Table 3.1: Example of clean annotation dataset

potential of new NLP tools and verify their efficiency when in combination with the resources at hand.

3.2 Annotation

The annotation step consisted in having three working students of the company, with proficiency in the English language, annotate the same set of 50 sentences as a first step. As mentioned before, the limiting time availability brought us to come up with a solution that involved the annotation of a smaller set in order to calculate the Inter Annotator Agreement on it and, if the agreement was high, moving on to a bigger set of around 1000 sentences that was different for each annotator. In this manner, it was possible to check the reliability of the guidelines and the competence of the annotators, while building a set of sufficient size for the project. The annotators performed the task using Microsoft Excel (`https://www.microsoft.com/en-ww/microsoft-365/excel`) as annotation tool on a .csv file.

The annotation guidelines were inspired by the ones used for the SemEval 2014 Task 4 (Pontiki et al. (2014)) and their purpose was to provide detailed instructions on how to identify Aspect Category, Sentiment Polarity and Aspect Term within the sentences. Although, the AT was also included in the guidelines and the annotations, its extraction was not included in the current work. The choice of performing a sentence-level annotation was also inspired by Pontiki et al. (2014), although for future work, the exploration of different solutions like a phrase-level annotation, to handle multiple classes of AC or SP within the same sentence, or a token-level one on the AT side, following the BIO schema, might bring interesting results.

The **Aspect Category** was identified as the category of the aspect discussed in the sentence. It provides the information regarding the category of what is being discussed within the sentence and is receiving a judgment. The selection of a relevant set of categories for this domain and task was a crucial step for a successful outcome of the project. The AC described in the table below were partially inspired by the work of Ashi (2019). In addition, they were refined through the computation of a TF-IDF using the scikit-learn implementation (`https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`) over the complete dataset and the observation of representative keywords that were generated. Among the categories we also find a *Multiple* label that was assigned when multiple AC were present within the same sentence. Although this shows the limitation of the sentence-level approach that was also present in the SemEval, it helps in the identification of problematic cases for future research.

The **Sentiment Polarity** is expressed through attitudes, opinions, evaluations, emotions, or feelings etc. of an opinion holder towards the aspect term within the sentence. A *Mix* label was also added to handle those cases where both positive and

Aspect	Description
a. Service	this includes the general service provided by the company or more specific services like cleanness, disability support, range of products offered, type of flight (economy, business, etc.), website, booking process
b. Company	this includes all the direct mentions to the company or also indirect through comparison with other companies
c. Staff	this includes all the mentions to the cabin crew, pilots or any other employee of the company (on-board and off-board), including customer service staff
d. Price	this includes the mentions to the price of the ticket or the services provided by the company, including food and products sold on board. If a mention to the price of the food, for example, is made within the sentence, the d category should be preferred over the g category
e. Travel	this includes the mentions to the flight, delayed or on-time departures and arrivals, turbulences and comfort
f. Aircraft equipment	this includes all the parts and equipment of the aircraft (e.g. the seats, the design, architectural choices) and the flight bundle (e.g. movies, headphones)
g. Food	this includes all the mentions to food or rinks included in the ticket price or sold on board
h. Safety	this includes all the mentions to the safety equipment, measures and services, like security checks, provided by the company. Also the mentions to the emergency exits should be included here (and not in f). The waiting time at the security checks is not included as it is not related to the safety, but to the boarding process (i)
i. Boarding	this includes all the mentions to the boarding process, like the waiting time, respect of the priorities, documents check and gates organization. Also the mentions to the waiting time at the airport and at the security checks should be included here
j. Luggage	this includes all the mentions to the handling, storing, tracking, weighting, and accidents about luggage
k. Information	this includes all the mentions to the information provided by the company regarding the flight and services, and through any channel (e.g. announcement on board or at the gates, website)
l. Others	this includes all the mentions to categories that are not present in the list
m. Multiple	this includes all the mentions to multiple categories in a sentence. Obviously, if one of the mentioned categories in the sentence is not the object of any judgment, it should not be considered
n. NA (no aspect)	this should be used when no aspect category is present, like in informative or anecdotal sentences that do not contain a judgment (e.g. “I was happy that day.”)

Table 3.2: Aspect Categories with corresponding descriptions.

negative polarities are present within the sentence. As done in the SemEval, it was specified that “if a sentence conveys both neutral and negative (or positive) opinions about an aspect category, then the negative (or positive) polarities dominate over the

neutral ones" (Pontiki et al. (2014)).

Sentiment Polarity	Description
Positive (1)	a sentence should be annotated with this label when the aspect term (or multiple aspect terms) within it is described with positive or some-what positive qualities or expressions
Negative (-1)	a sentence should be annotated with this label when the aspect term (or multiple aspect terms) within it is described with negative or some-what negative qualities or expressions
Neutral (0)	a sentence should be annotated with this label when the aspect term within it is not described with explicitly positive or negative qualities or expressions. This label should also be used in informative or anecdotal sentences that do not contain any aspect term
Mix (2)	as described for the conflict label in (Pontiki et al., 2014), a sentence should be annotated with this label when the aspect term (or multiple aspect terms) within it is described with both negative and positive qualities or expressions

Table 3.3: Sentiment polarities with corresponding descriptions.

The **Aspect Term** is a single or multiword term contained in the sentence that provides the information about the aspect category. As mentioned previously, we will not go into the details of this element for this work. The complete guidelines can be found in Appendix A. Table 3.4 shows how an annotated sentence looks like.

Sentence_ID	Feedback_ID	Sentence	Aspect_Category	Sentiment	Aspect_Term
1	129288273	The snacks that were distributed during the flight were really tasty!	Food	Positive	The snacks

Table 3.4: Example of annotated sentence

3.3 Inter Annotator Agreement

The Inter Annotator Agreement was computed using Cohen's Kappa scikit-learn implementation (`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html`). Cohen's Kappa is a statistical measure used to calculate pairwise agreement between two annotators. The formula is described as follows, where p_0 is the probability of agreement and p_e is the probability of random agreement:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

The IAA results for respectively the AC and SP are shown in Tables 3.5 and 3.3.

As expected, a higher agreement is shown on the SP compared to the AC. This can be explained by the superior straightforwardness of the first task compared to the other

Annotators	Cohen's Kappa
A1 & A2	0.81
A1 & A3	0.79
A2 & A3	0.76

Table 3.5: Inter Annotator Agreement on Aspect Category

Annotators	Cohen's Kappa
A1 & A2	0.90
A1 & A3	0.88
A2 & A3	0.86

Table 3.6: Inter Annotator Agreement on Sentiment Polarity

and by the number of labels to be assigned. In fact, having a wider number of labels proved to harm the agreement score by augmenting the chances of overlap between them.

3.4 Stats

Given the high IAA in the annotation of the first batch of 50 equal sentences, the annotators moved on to a bigger set that was different for each of them in order to put together a wider number of training and test examples. Annotator 1 and Annotator 2 annotated 1000 sentences each, and 495 were annotated by Annotator 3.

The final dataset consisted of 2495 annotated sentences. Table 3.7 shows some statistics [1].

# Sentences	2495
# Tokens	43824
Most common single word	flight
Most common bigram	business class
# sentences containing actual AC	1451
Average sentence length	17.56
Most common AC	NA - 37%
Most common SP	Neutral - 47%

Table 3.7: General statistics regarding the annotated dataset

The sentiment distribution is shown Figure 3.1. As expected, the *Neutral* class is the most common one, being representative of 47% of the dataset. This was also the case for some related work, because not all the sentences within a feedback actually contain an evident polarity orientation. It also needs to be specified that all the sentences that were not explicitly expressing either a *Positive* or *Negative* sentiment were annotated as *Neutral*. The majority class is then followed by the *Negative* one with

[1] For the cell referring to "# sentences containing actual AC", we consider those that are **not** *Others*, *NA*, *Multiple*

34%. This is also not a surprise, because the feedbacks were written with the intent of improving an aspect of the service that was provided by the company. Therefore, a wide number of the texts contain comments on something that is not good enough yet or needs improvement and, hence, in the customer's opinion, is on the negative spectrum of the experience. Fewer examples are labelled with *Positive* sentiment, while the poor percentage of *Mix* shows that expressing multiple polarities within the same sentence was not a common pattern. This last point also entails that the system is going to have less chances to encounter problematic cases and, consequently, have better performances.

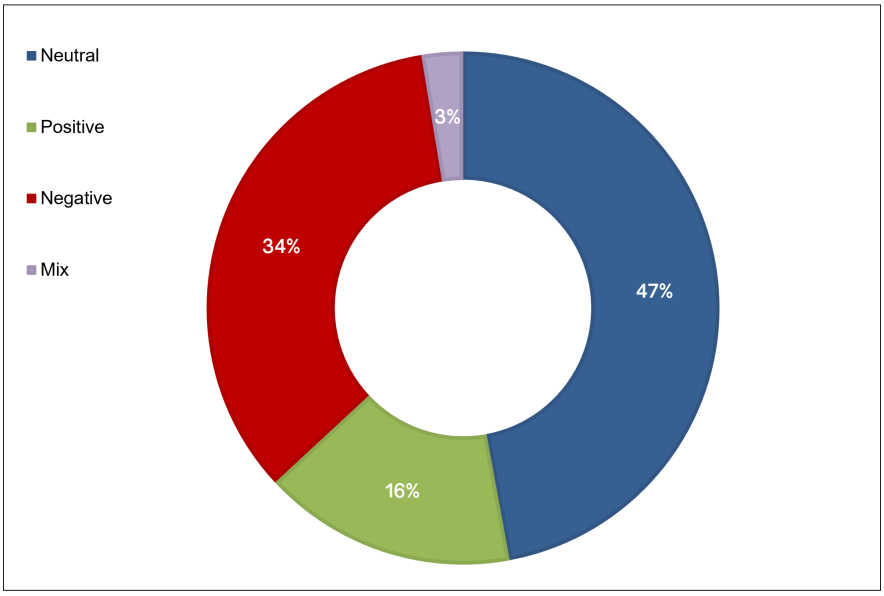

Figure 3.1: Sentiment Polarity labels distribution over the annotated dataset

On the AC side, in Figure 3.2, we see an evidently unbalanced distribution of the classes. In fact, we notice a great predominance of the non-class *NA* over the rest, being representative of close to the 40% of the data. The reason behind this high percentage can be found in the fact that most of the feedbacks contain anecdotal sentences which describe facts, without expressing explicit judgements towards an aspect. Among the categories that have a number of occurrences between 250 and 100 we find *Staff*, *Service*, *Aircraft equipment*, *Food*, *Company*, *Travel*, *Multiple*, *Information*, and *Luggage*. The presence of the *Multiple* class within this group shows a different pattern from the SP classification task. In fact, while the presence of multiple polarities in the same sentence was rather uncommon, in the AC spectrum, we see the opposite tendency. In fact, although the number of occurrences is not extremely high (121), it shows a marked superiority over classes that were expected to be more recurrent (e.g. *Price*). It is still to be explored whether this phenomenon would also be representative of an annotated dataset with a wider number of examples. For what concerns this work, it is clear that the systems will have a higher chance to encounter a problematic case than a supposedly easier one (e.g. *Luggage*). In the group of categories that occurred less than 100 times we find categories that were not completely representative of this small set (e.g. *Price*). Future work might also take into consideration the option to merge this small classes with related ones that are more extensive in numbers. Finally, follow-

ing the guidelines, the Other class was included to cover those categories that might have been left out while designing the task. However, the extremely low number of occurrences proves that all the representative classes for this domain were successfully taken into consideration and the use of this label becomes rather superfluous.

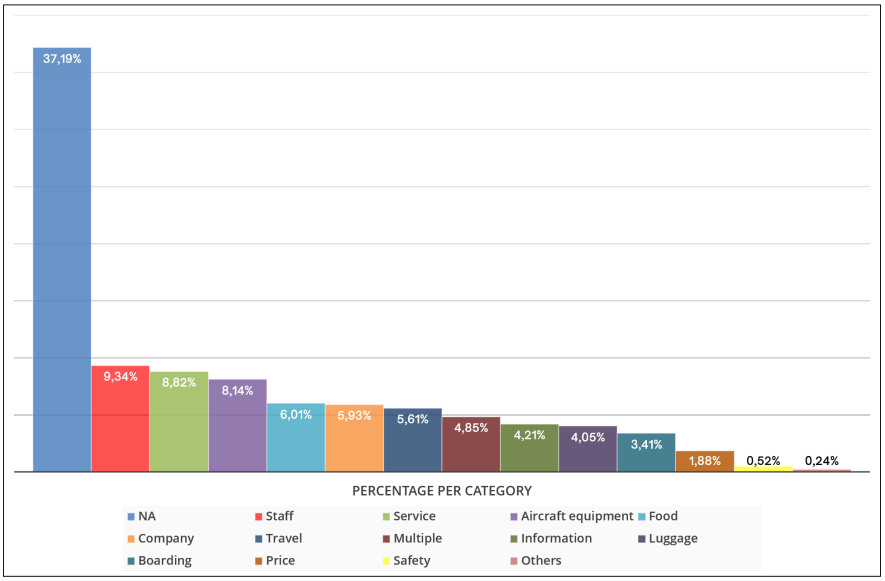

Figure 3.2: Aspect Category labels distribution over the annotated dataset

Aspect	Occurrences
NA (no aspect)	923
Staff	233
Service	220
Aircraft equipment	203
Food	150
Company	148
Travel	140
Multiple	121
Information	105
Luggage	101
Boarding	85
Price	47
Safety	13
Others	6

Table 3.8: Number of occurrences of each Aspect Category label.

The dataset was ultimately splitted into 3 parts: 80% for the training set, 10% for the validation set used in the training process as a tool for preliminary evaluation, and another 10% for the test set of unseen data.

Chapter 4

Classification Method

This chapter contains the information about Transfer Learning, the experimental approach that was used for the task. It describes the reasoning behind the choice of this technique, the architecture of the language models that were fine-tuned (BERT and RoBERTa) and the systems setup for classification.

4.1 Transfer learning and finetuning

Traditional machine learning methods have proven that, although great results are achievable through them, the circumstances and the premises related to their use entail restricting boundaries. The assumption that training and test sets share the same feature space and distribution results in the limited application of the models to their source task and domain (Pan and Yang (2010)). The use of these models in different circumstances from the original one for which they were trained for, results in poor performances. This usually leads to the necessity of collecting new suitable data and use it to (re)train a completely different system from the ground up. In real life scenarios, these steps are costly to carry out in terms of time and effort. In fact, to achieve positive results, traditional machine learning systems rely on plentiful amounts of hand-labelled data, the availability of which is usually scarce and not balanced between tasks or domains. Furthermore, even when the ideal resources for retraining are attainable, a reasonable ratio between costs and benefits still needs to be taken into account. Transfer Learning (**TL**) is a technique that allows an enhanced reusability of pre-trained systems, with a consequent reduction of the expenses, by overcoming the previously mentioned restrictions present in traditional machine learning methods. This becomes possible thanks to the concept of transferring the knowledge that a model acquired during the training process to new tasks and domains. Transfer Learning gained increasing popularity over the past years by researchers in the field of Natural Language Processing and it has been successfully applied on variety of tasks.

The simplest way of describing Transfer Learning would be as "a means to extract knowledge from a source setting and apply it to a different target setting" (Ruder et al. (2019)). Figure 4.1 is a good visual representation of the concept taken from Pan and Yang (2010).

A more formal and detailed definition was provided in the survey of Pan and Yang

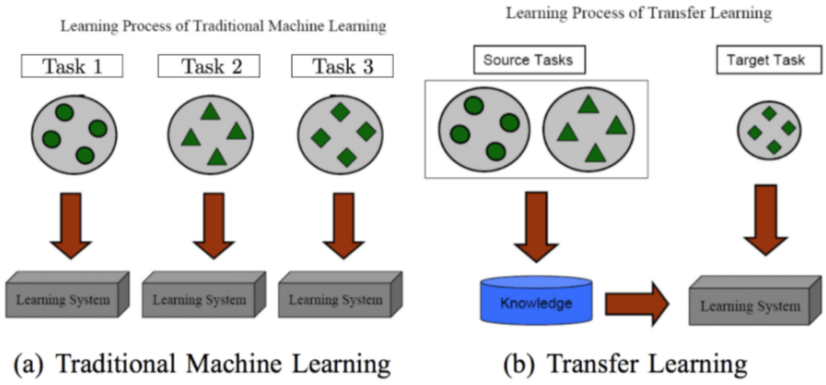

Figure 4.1: Transfer Learning concept in comparison to Traditional Machine Learning. Source: Pan and Yang (2010)

(2010). Here, a domain is defined as $\mathcal{D}$ and consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = x_1, \cdots, x_n \in \mathcal{X}$. Taking, for example, a document classification task that uses a binary representation of the feature terms, is the vector space of the document, x_i is the ith feature vector corresponding to a document, and X is a training sample.

Given a domain, $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a task $\mathcal{T}$ consists of two different components: a label space $\mathcal{Y}$ and a conditional probability distribution $P(Y|X)$ that is learned from the training samples consisting of pairs $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. In the document classification task example previously mentioned, $\mathcal{Y}$ is the set of all labels (e.g. True, False) and y_i is one of the binary labels (True or False).

Assuming all the above, given a source domain $\mathcal{D}_S$ and a relative source task $\mathcal{T}_S$, and given a target domain $\mathcal{T}_T$ and a target task $\mathcal{T}_T$, the goal of TL is to help the system in learning the conditional probability distribution $P(Y_T|X_T)$ in $\mathcal{D}_T$ through the knowledge acquired from the $\mathcal{D}_S$ and $\mathcal{T}_S$ where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

Generally, four different scenarios with related conditions are presented when applying Transfer Learning, as described in Pan and Yang (2010). These can be summarized as follows and easily described with the documents example of the paper:

a) $\mathcal{X}_S \neq \mathcal{X}_T \rightarrow$ The source and target domains do not share the same vector spaces; for example, the documents are written in two distinct languages. This is referred to as cross-lingual adaptation.

b) $P(X_S) \neq P(X_T) \rightarrow$ The marginal probability distributions between the source domain and target domain differ; for example, the topics between the documents are not the same. Dealing with a case like this is referred to as domain adaption.

c) $\mathcal{Y}_S \neq \mathcal{Y}_T \rightarrow$ The label spaces in the two tasks are not the same; for example, in the source task, documents are classified with different labels.

d) $P(Y_S|X_S) \neq P(Y_T|X_T) \rightarrow$ Source task and target task have distinct conditional probability distributions; for example, source and target documents are not balanced in the distribution of their classes.

In Computer Vision, TL has been successfully applied for more than 10 years by re-training supervised learning models on ImageNet, a huge dataset containing 1.2 million of labelled pictures (Deng et al. (2009)).

For what concerns Natural Language Processing, the application of TL methods is more recent. In this area, three distinct factors play a major role when applying Transfer Learning (Ruder et al. (2019)). In fact, it should be considered:

a) if the source and target settings are sharing the same task;

b) what the characteristics of source and target domains are;

c) whether the tasks are learned simultaneously or sequentially.

Sequential Transfer Learning, which is addressed in the last point, is the type that has brought the most successful results in NLP, with language modeling as main method. Here, as a key difference with the previously mentioned supervised approaches applied to Computer Vision, the pre-training step is usually self-supervised, which means that it does not require human-annotated labels. In fact, a common pre-training objective is to guess the next word in a sentence which has as sole requirement a large amount of text data. An example of self-supervised source task is to predict the value of random masked words, a process that is similar to a cloze task which requires to fill in the blanks in a text. In this way, the structural and semantic knowledge is extracted from the text and it is ready to be transferred to a new target objective.

The transferring process related to language modeling is usually called fine-tuning, which consists in removing the last layer of the language model architecture, and replace it with a new one that returns a classification output related to the target task. In the case of masked language modeling as a source task and Spam classification (Spam, Not Spam) as target task, the linguistic knowledge is retrieved through the pre-training step and the final layer of the model is substituted with a randomly initialized binary classifier, which outputs the prediction labels using the previously acquired information. Studies have also shown that the more similar are the source and target task/domain, the better the performance of the model (Phang et al. (2019)). The language models that in last years have received the highest success are based on the famous Transformer architecture (Vaswani et al. (2017)).

The Transformer neural network architecture was initially created to solve the task of language translation, when BiLSTM systems were usually preferred until that moment. However, in an BiLSTM model, words are inputted and generated sequentially, which is a process that takes a significant number of steps and, consequently, time. Furthermore, the true meaning of words is hardly learned. The reason behind this is that by learning left-to-right and right-to-left contexts separately, and then concatenating them, the true context of the text is slightly lost. On the other hand, the Transformer archi-

tecture addresses these matters by, firstly, processing the words simultaneously, making the model faster than the mentioned NN. Secondly, the contextual learning power is enhanced by the creation of an embedding representation of the words that preserves meaning through the use of the so-called "self-attention mechanism", and structural information with positional embeddings. If we take the example of a translation between L1 and L2, the encoder layer takes the words of L1 and generates embeddings for each word simultaneously. These embeddings are vectors that encapsulate the meaning of the word, where similar words share the same vector space. The decoder takes these embeddings from the encoder, together with the previously generated words of the translated sentence in L2, and uses them to generate a translation one word at a time until the end of sentence is reached. Therefore, in this RNN inspired architecture, the encoder is in charge of understanding the structure and semantic knowledge of L1, while the decoder takes care of learning the linguistic relations between the two languages to generate a new sequence in L2. These Transformer's components have eventually been used to develop new language model architectures. In fact, by stacking a series of decoder layers, the team of OpenAI created the GPT model (Radford et al. (2017)), while by stacking encoders, Google developed BERT (Devlin et al. (2019)). The latter became famous for its successful fine-tuned application on various NLP tasks.

4.1.1 BERT

BERT stands for Bidirectional Encoder Representation from Transformer and it is the Transformer-based model that has received the highest attention by researchers since its release. This model, as the name suggests, was designed for learning deep bidirectional representations from unlabeled text through multi self-attention heads for each layer that capture both left and right context. This design makes it possible to fine-tune the pre-trained BERT with only one extra output layer to provide state-of-the-art models for a variety of tasks, such as summarization and question answering, without requiring significant task-specific architectural changes (Devlin et al. (2019)). BERT was developed over two simultaneous pre-training tasks: Masked Language Modeling and Next Sentence Prediction. When the model is finetuned, the output layers of these source tasks are replaced by the new target task classifier.

Architecture

BERT comes in a "base" and a "large" version. The base version consists of 12 encoder layers, each of them having 768 dimensions and 12 attention heads, accounting for a total of 110M parameters. BERT large, instead, is an extended version consisting of 24 encoders, 1024 dimensions and 16 self-attention heads, for a total of 340M parameters. The bidirectional use of the self-attention mechanism inspired by the Transformer, gave to this model the advantage of capturing both left and right contexts, as opposed to the left side constraint of OpenAI GPT. For the purpose of this project, BERT base is the version that was used.

Input representation

In order to process the input, BERT needs to first create an embedding representation of a single sentence or pair of sentences depending on the task. Figure 4.2 taken from BERT's original paper (Devlin et al. (2019)) provides a good visual representation of how the input is processed by the model.

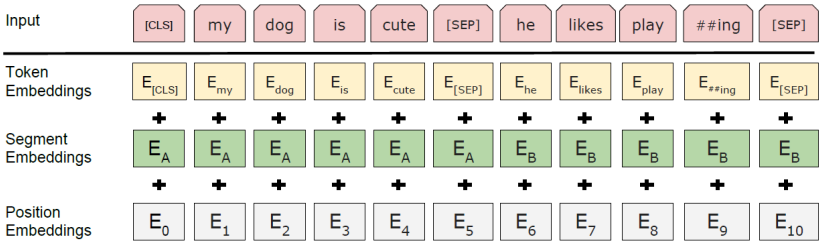

Figure 4.2: Input representation in BERT. Source: Devlin et al. (2019)

Each sentence, or better "sequence", is encoded as a series of tokens using the WordPiece algorithm (Wu et al. (2016)). This is a technique for subword tokenization that consists of breaking words into smaller units when a correspondence in its vocabulary is not found. Those units can go up to the character level, hence, creating an embedding representation for any word present in the input, even the ones that the model has never seen. It works in an extremely similar way to the Byte Pair Encoding (BPE) (Sennrich et al. (2016)) where the words are splitted into sequences of characters that contain a special token for beginning or end of the sequence. Then, the most frequently occurring character n-grams are merged until a predetermined vocabulary size is reached. As opposed to this, WordPiece considers the probability instead of the frequency when merging a pair of characters. The vocabulary of BERT consists of 30K tokens, among which also words containing common typos are included. Among the tokens that belong to each sequence, there are also some special tokens. The [CLS] is the first token of any sequence and contains also the information for the output of the classification task. On the other hand, the [SEP] token is the last token of a sequence and serves as a delimiter between sequences. The [PAD] token is used to assign the same length to each input sequence. As a rule of thumb, before inputting the data into the model, the maximum sequence length is verified in order to assign as many [PAD] tokens are necessary for shorter sequences. For instance, if a sentence contains 6 tokens and another contains 4, they will be both "padded" to 6 and two [PAD] tokens will be assigned to the shorter sentence. Finally, the [MASK] token is only assigned during pre-training and it is used to mask a random word in a sequence.

Apart from assigning the special tokens, BERT also creates two extra embedding representations during as preprocessing step. One is the Segment Embedding, which defines whether a token related to a sentence or to another. Lastly, a Positional Embedding is assigned to each token in order to preserve structural information of the sequence. This is further important because the elements in the sequence are fed simultaneously to the model, and not sequentially, as mentioned in the previous sections for the BiLSTM.

Pretraining Tasks and Data

The pre-training tasks of BERT consist of Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). The first is similar to cloze task where gaps in a text need to be filled in through the surrounding context. It consists in randomly masking one token within a sequence with the special [MASK]. In this process 15% of the tokens contained in the data are masked. The masking is carried out to force the prediction of the word in the context and avoid the problem that occurs in bidirectional systems where a word is capable of seeing itself. In order to avoid a mismatch between pretraining and finetuning when the [MASK] token is not appearing in the second step, the model replaces 15% of the data with a mask 80% of the times, with another random token 10% of the times , and with the original token the rest of the times.

For what concerns the Next Sentence Prediction, the system is trained to carry out a binary classification and recognize if Sentence A is following Sentence B. In this process, Sentence B is actually following Sentence A 50% of the times and being labelled with *IsNext*. IN the remaining 50%, a random sentence is selected and labelled as *NotNext*. This pretraining task is particularly useful the system needs to be fine-tuned for downstream tasks like Question Answering, where a pair of sentences is involved.

Following these pre-training steps, the relatedness between tokens within the sequence are learned through MLM, while long-distance dependencies are captured by NSP.

The text data that was used for the pre-training of BERT consisted of a total of 16GB from the combination of English Wikipedia articles and Book Corpus (Zhu et al. (2015)).

4.1.2 RoBERTa

A recent study conducted by the Facebook AI team in collaboration with the University of Washington claimed that BERT was undertrained (Liu et al. (2019b)). Starting with this assumption, they developed RoBERTa, a Robustly Optimized BERT Pretraining Approach, by experimenting with the pre-training setup of Goolgle's popular language model. The main adjustments focused on the removal on the NSP task, training on bigger batch sizes and longer sequences, using Dynamic Masking, and using a larger dataset.

According to (Liu et al. (2019b)), the NSP was not enhancing the understanding of long-distance dependencies. In fact, contrary to what was affirmed in Devlin et al. (2019), several experiments showed that by removing the NSP in the pre-training step, equal or improved performances were achievable on downstream tasks. Also training using bigger batch sizes proved to improve the learning power of the system. In the specific, the 1M steps with a size of 256 sequences were replaced with 125K steps of 2K sequences, leading to a higher perplexity score of the MLM task. In addition, a Dynamic Masking method, where the masked tokens change in each epoch, was found

to be slightly better than the original Static Masking of BERT. Finally, a bigger dataset was used for the pre-training of RoBERTa. It consisted of the 16GB that were already used in BERT, combined with:

- CC-NEWS: 63 million English crawled news articles (76 GB);
- OPENWEBTEXT: web content extracted from the URLs shared on Reddit (38 GB);
- STORIES: subset of Common Crawl data (31 GB).

Eventually, the final dataset consisted of a total 160GB of English text, with an extended vocabulary of 50K BPE encoded subwords.

The work of Liu et al. (2019b) showed that RoBERTa outperformed BERT over a number of benchmarks, among which the General Language Understanding Evaluation (GLUE) benchmark (Wang et al. (2019)), the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. (2016)) and the ReAding Comprehension from Examinations (RACE) (Lai et al. (2017)).

4.2 Baseline

A simple majority baseline inspired to the one used for the SemEval 2015 (Pontiki et al. (2015)) was also created for evaluation purposes. This is a system that always assigns the most frequent class in the data to each test sample. If the task is Aspect Category Detection, the system will always predict *NA*, while for Sentiment Polarity classification it will assign *Neutral*. Given the unbalanced distribution of the labels seen in the Data Section, it becomes reasonable to verify if the system is simply over-assigning the majority class or actually using the linguistic knowledge to predict the correct class.

4.3 Systems Setup

Designing a balanced and efficient setup is not an easy task when working with Transformer-based models. In fact, the high expenses in terms of time for training and effort in the interpretation of early results with consequent adjustments of parameters, play a major role in the achievement of a successful outcome. The challenges that are involved in a real life scenario, like in this specific project, where limited time availability for development and the low computational power provided, which did not include the use of any GPU, brought to the choice of a practical, but still efficient experimental plan.

The experiments, in fact, involved the use of the same setup for both BERT and RoBERTA, in order to have a more straightforward comparison between the two, and obtain an insight into the performance of each of them under the same conditions. Then, a series of experiments with different setups would follow for the best performing model, in order to verify if improved results are achievable.

Furthermore, the same setup was used to train two separate classifiers, one for each task (ACD and SP). This choice followed the approach that has also been used in related work on Aspect Based Sentiment Analysis.

BERT and RoBERTa were fine-tuned for the sequence classification task using their popular HuggingFace (`https://angel.co/company/hugging-face`) implementation of the models [1]. The setup consisted of:

- 4 Epochs
- 63 Batch size
- 64 Sequence length (padding)
- 2e-5 Learning rate
- AdamW Optimizer

The data was splitted into three sets: 80% for training (training set), 10% for validation during training (validation/development set), and 10% for testing on unseen examples (test set). As a preprocessing step, the systems read in the sentences, use the built in tokenizer to create a WordPiece or BPE representation of the elements in the sequence, and pad them to an equal length of 64. Augmenting the number of the sequence length would be an appropriate solution when the maximum length of the input is not known, as it would happen in real life scenarios if the system was used as a product. However, padding to longer sequences also involves that the fine-tuning might take up to twice the usual time (e.g. RoBERTa: 64 seq. len., 50min train.; 124 seq. len., 2h train.). Therefore, for practical reasons the padding was kept just above the maximum length of the sequences contained in the dataset. After the padding, the labels are encoded and the batches are created. When the preprocessing is completed, the systems are ready for the fine-tuning step. Here the models learn the correlation between the input features of a sequence and the assigned label. In this way, each token is assigned to a vector space with other tokens sharing the same context, and an association is learned between this space and a class. The connection created between tokens and labels is returned as a probability for the sequence to be related to a certain label depending on the task (AC label for ACD and polarity for SP). The acquired knowledge is then verified against a validation set in each epoch, in order to minimize the error and adjust the weights until an efficient performance is achieved at the end of the training. Using the right parameters in this case is of fundamental importance to avoid both underfitting and overfitting. Underfitting happens when the system is not able to learn enough information for the classification, resulting in poor performances. Overfitting is the opposite case, where the system learns perfectly the correlation between the features of the validation set, but it is not able to generalize and make correct predictions on new unseen examples. Avoiding both situations and find the most balanced setup becomes, for this reason, a priority in order to achieve successful results.
Given the company policy involving the careful treatment of private data of the client,

[1] The code was adopted and modified for this task from the work of George Mihaila: `https://gmihaila.medium.com/fine-tune-transformers-in-pytorch-using-transformers-57b40450635`

it was not possible to run the systems on virtual machines like in Google Colaboratory (`https://colab.research.google.com/`), which provides access to GPUs and TPUs. This is usually the most common method used for running Transformer-based models considering the computational power needed. Instead, the experiments considered in this work only involved the use of a 1.60-2.11GHz Intel® Core™ i5-10210U CPU and 8GB of RAM. Therefore, the training time involved was proportional to this power, with an average of 1.30h.

The fine tuning process for BERT showed the following trend, also displayed in Figure 4.3. For ACD, both training and validation loss progressively decreased, while the accuracy was experiencing an average improvement of 10% for each epoch. This shows that the system was able to ameliorate its knowledge through minimization of the error and adjusting properly through validation. However, by analyzing the values of the last epoch, the system did not manage to reach a loss close to 0 and, as a consequence, the validation accuracy scored a rather low value. Augmenting the number of epochs could bring overall improvements of the loss, although overfitting could be also possible considering that the threshold recommended by Devlin et al. (2019) would be passed (2-4). For SP classification, Figure 4.4, although the trend was similar, the system proved to be better at minimizing the loss, with values closer to 0, and a higher accuracy.

The RoBERTa system showed an overall superiority during the training process (Figure 4.5). As showed by the figures, training and validation loss achieved an ideal equality of values for ACD, after experiencing a constant decrease. This proves that the parameters were appropriately tuned for the enhancement of the learning power of the model. The suspicious higher score of the validation accuracy compared to the training one might be, instead, explained by the scarcity of training examples combined to unbalanced distribution of the labels. This is usually a common outcome when test sets of this size are employed.

For the SP classification (Figure 4.6), impressive results are achieved, where again the training and validation loss are extremely similar in the last epoch. Furthermore, the slightly higher loss in the validation implies that this setup was ideal for balancing the learning power by both avoiding underfitting and overfitting. Finally, given the 0.81 validation accuracy of the 4th epoch, high performances on the testing process became foreseeable.

By comparing the training performances of the systems, it is evident that the chosen setup was suitable for both BERT and RoBERTa. Moreover, having less classes for the SP task seems to be helping the systems in minimizing the error, as the chances of making a correct prediction are higher (1/3 for SP, compared to 1/13 for ACD).

Given the overall superiority of the RoBERTa model, some attempts were made on the parameters tuning of this system in order to test if improved results were feasible. The new experimental setup (we will refer to this model as **RoBERTa Rec** from this point on) followed the recommendations for fine-tuning contained in the original paper (Liu et al. (2019b)). It consisted in lowering the batch size from 63 to 32 and increasing the learning rate from 2e-5 to 3e-5. This was also motivated by the fact that

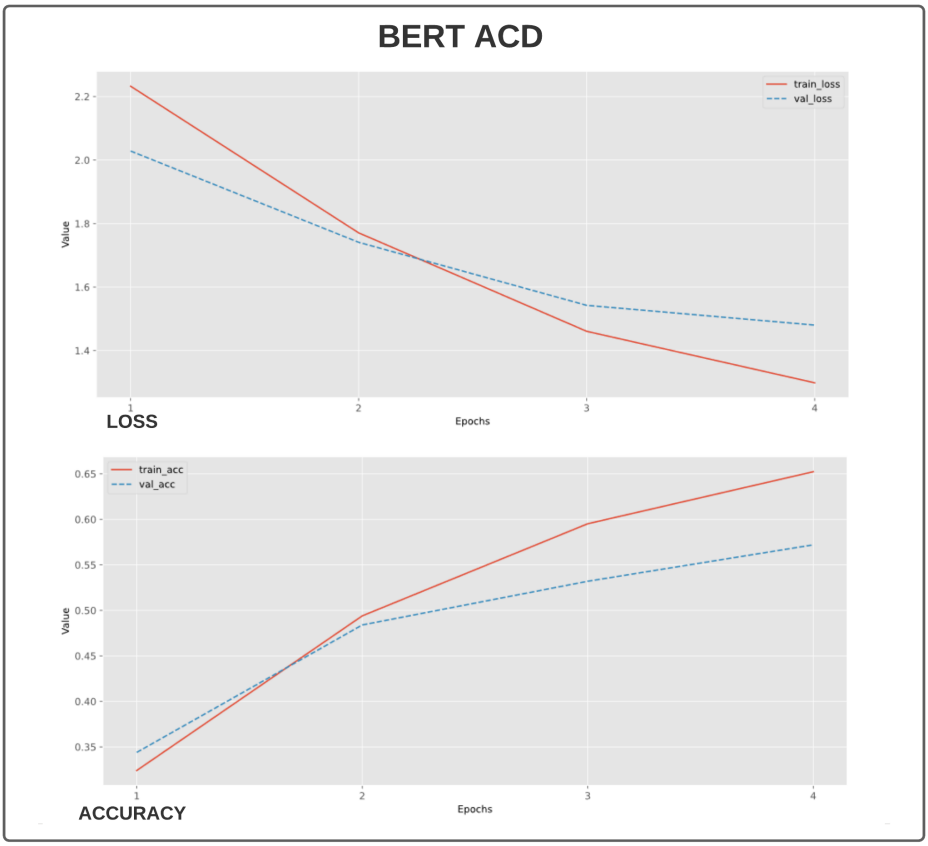

Figure 4.3: Loss and Accuracy trend during finituning for BERT on the Aspect Category Detection task.

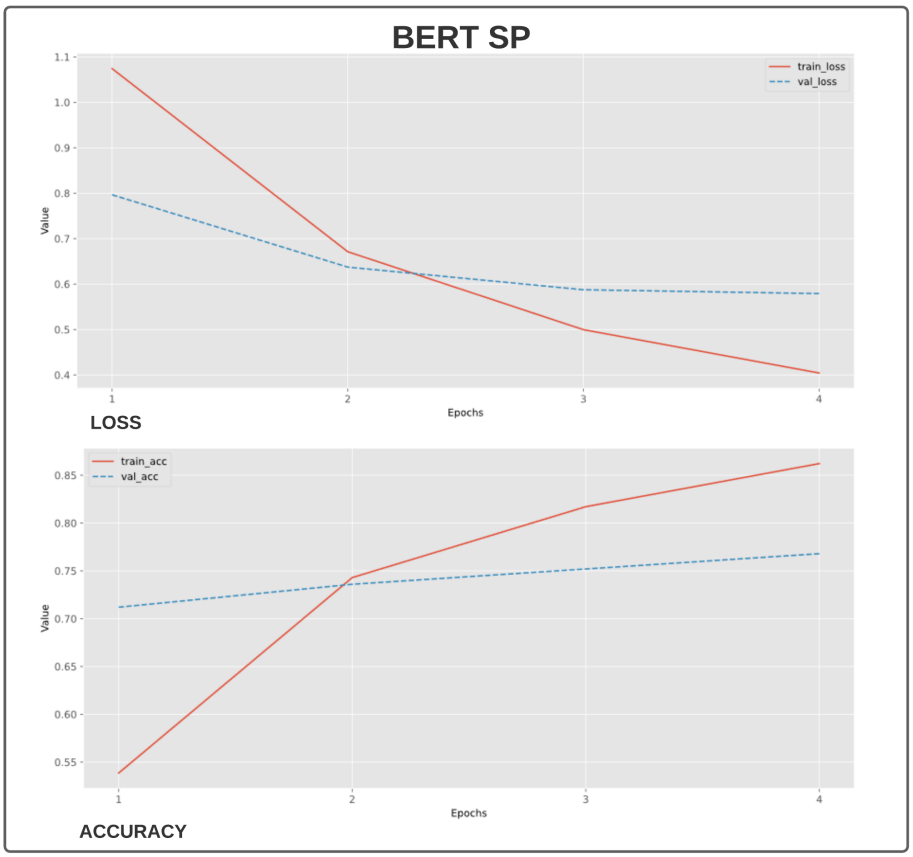

Figure 4.4: Loss and Accuracy trend during finituning for BERT on the Sentiment Polarity task.

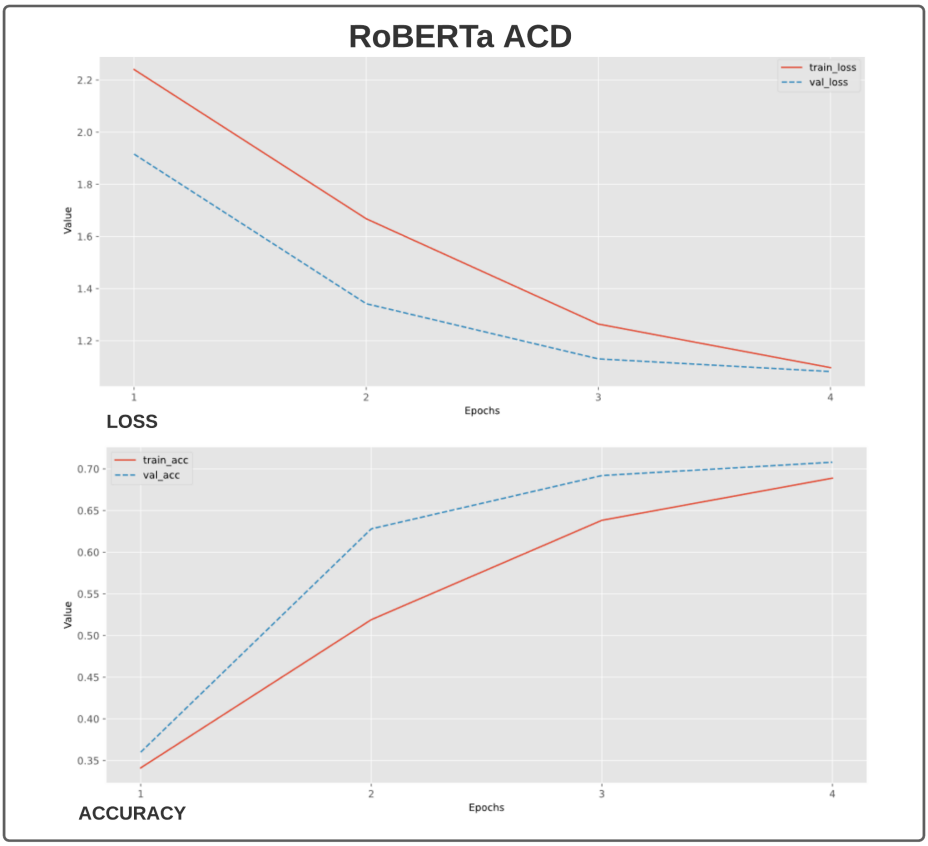

Figure 4.5: Loss and Accuracy trend during finituning for RoBERTa on the Aspect Category Detection task.

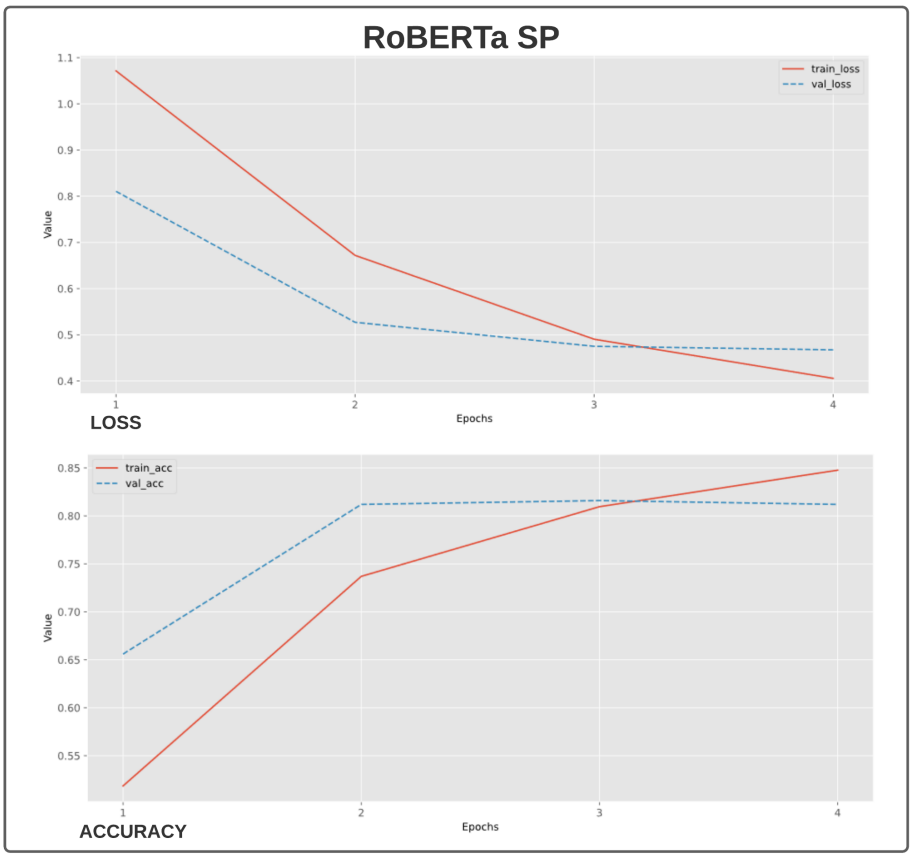

Figure 4.6: Loss and Accuracy trend during finituning for RoBERTa on the Sentiment Polarity task.

keeping both high learning rate and batch size might have brought some overfitting. Therefore, as a rule of thumb, it is usually recommended to lower one when increasing the other to keep the setup balanced. On the paper, it was also recommended to use up to 10 epochs for training, but several experiments showed that the system stopped learning after the 4th and started overfitting, therefore, the number of epochs remained unchanged. Against my expectation, the new settings did not bring sensitive improvement on the initial results (Figure 4.7 and 4.8), proving that the first setup was probably the most efficient and balanced for the task at hand.

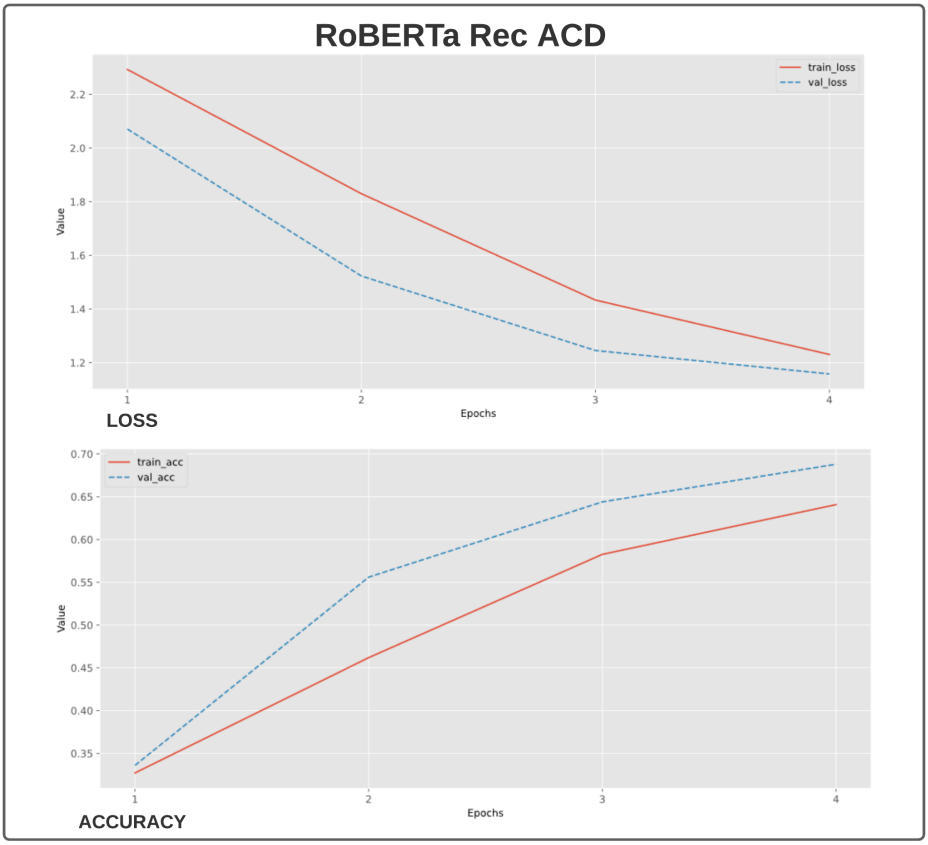

Figure 4.7: Loss and Accuracy trend during finituning for RoBERTa Rec on the Aspect Category Detection task.

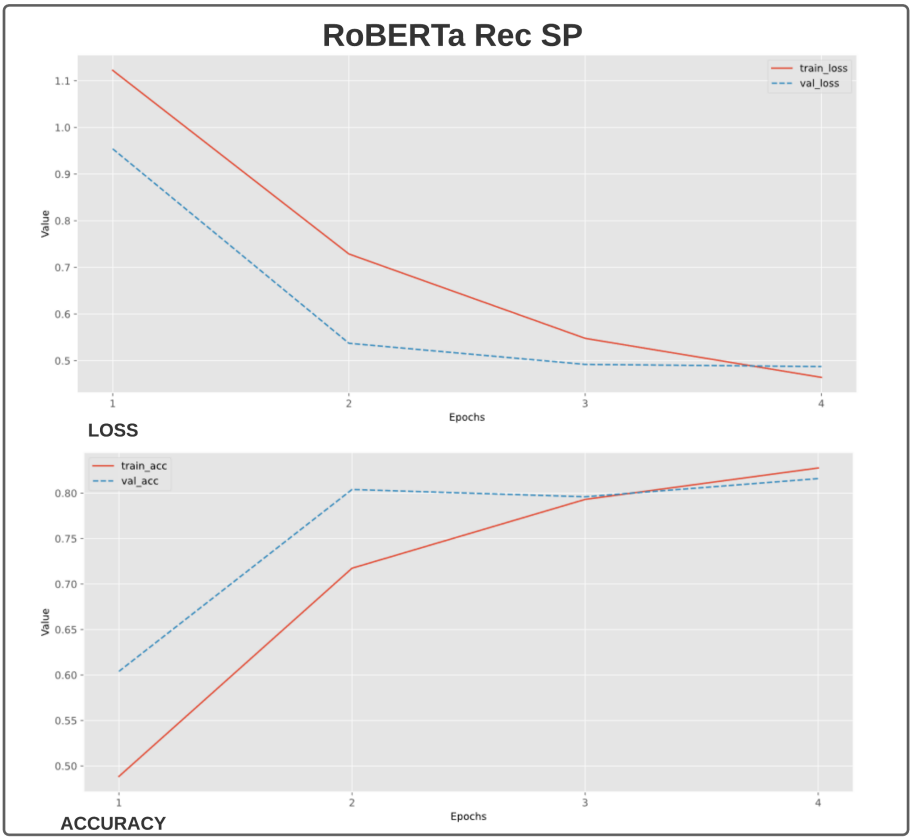

Figure 4.8: Loss and Accuracy trend during finituning for RoBERTa Rec on the Sentiment Polarity task.

Chapter 5

Results & Analysis

This chapter is going to focus on results analysis and evaluation. First, an overview on the annotation outcome will be provided. Secondly, a comparative evaluation of the systems will follow. Finally, a detailed description of the performance of the best system will be discussed through examples of correct and mistaken predictions.

5.1 Annotation Results

Being the systems trained and evaluated on human annotated data, an analysis of the annotations becomes of fundamental importance, as this will also determine the performances achieved in the experiments. As already shown in Chapter 3, the computation of the inter annotator agreement demonstrated that A1 and A2 shared a more similar understanding of the annotation guidelines by scoring a Cohen's Kappa of 0.81 for ACD (Table 3.5), and 0.90 for SP (Table 3.6). This becomes also evident when looking at the confusion matrix in Figure 5.1.

On the 50 sentences considered for this step, the cases of disagreement are rare and not entirely representative of a defined pattern. This proves that the annotation guidelines were clear enough and appropriate for the task, leading to agreement scores that are comparable to the ones of related work (Pontiki et al. (2014)). On the other hand, the annotators might have probably struggled in determining the difference between an implicit judgement and an anecdotal statement, when assigning the non-class *NA* for ACD.

If the analysis is also extended to the third annotator, who scored a lower agreement with the other two (Table 3.5 and 3.6), a small percentage of confusion regarding *NA*, *Boarding*, and *Travel* classes is displayed (Figure 5.2). Nevertheless, given the rather occasional occurrence of disagreement cases, the high agreement on most of the classes, and the exceptionally high agreement between the two annotators (A1 and A2) that eventually annotated the largest number of sentences (1000 for each, against 495 of A3), a promising performance of the models becomes foreseeable.

As displayed in the pair-wise comparison between the annotators on SP of Figure 5.3, the annotators demonstrated a clear understanding of the task by achieving Cohen's Kappa scores above 0.85 (Table 3.6). These results show a definition of the guidelines

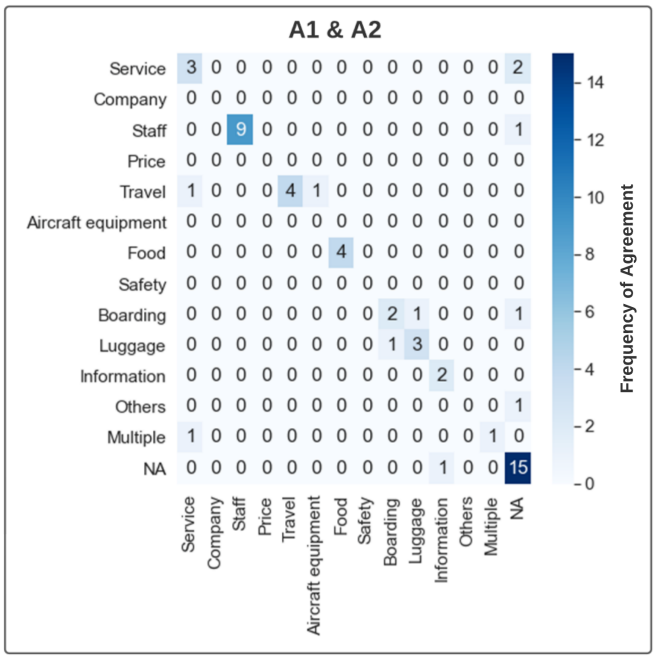

Figure 5.1: Confusion matrix concerning the agreement between A1 and A2 on the Aspect Category task

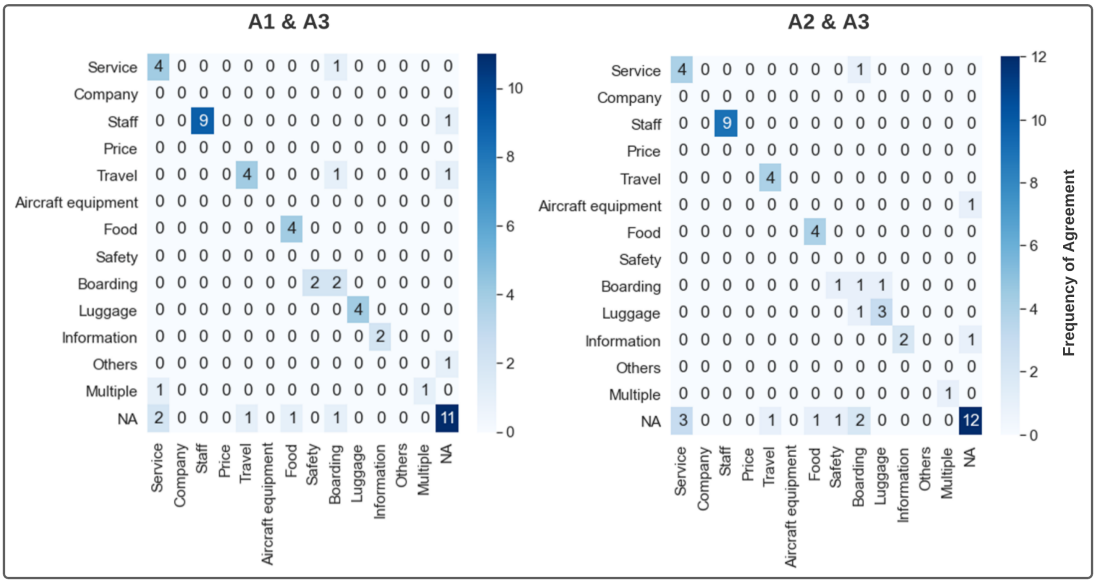

Figure 5.2: Confusion matrix concerning the agreement of the first two annotators with A3 on the Aspect Category task

that left little room for interpretation and clear understanding of the task. Furthermore, the ratio between the number of classes and the number of samples proved to be more balanced compared to ACD. This led to a more straightforward interpretation of the results and opened up to new considerations for future work, where the same distribution might be taken into consideration for improved insight into the agreement on ACD. In fact, given the tight time availability for this project, a more extensive annotation process would have been problematic and not entirely feasible. For this reason, remains to be explored the option of calculating the agreement on a wider number of samples, with a focus on a more balanced distribution of the classes and the use of a set containing at least more than 3 occurrences for each category. This might provide a fine-grained insight into the weak spots of the guidelines and an accurate representation of problematic cases or overlapping classes.

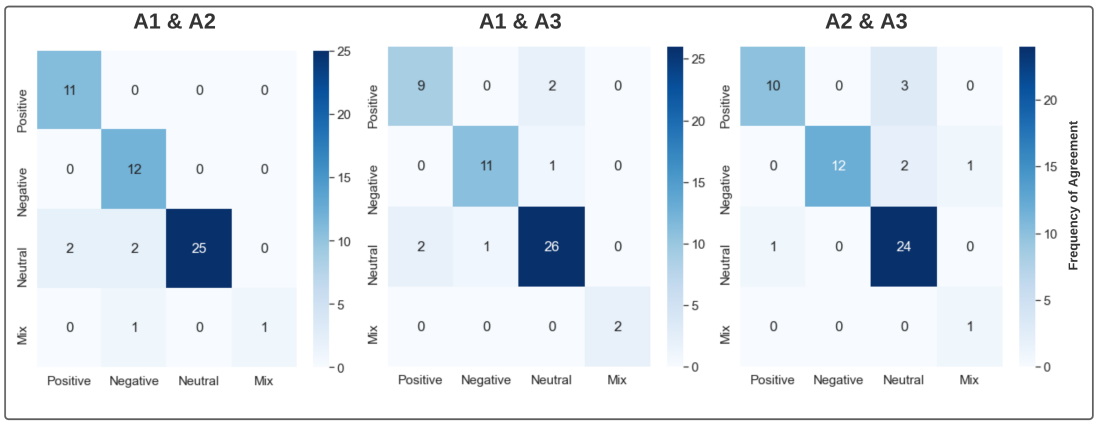

Figure 5.3: Confusion matrix concerning the pair-wise agreement between the three annotators on the Sentiment Polarity task

5.2 Classification Evaluation

The evaluation of the classification was conducted on a test set containing 249 examples, representing 10% of the original dataset. For the random selection of the samples, the scikit-learn data-split tool (`https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`) was used. In fact, the models have been evaluated based on Precision (p), Recall (r), F1-score ($f1$), and Accuracy ($acc.$). Following what was established in the SemEval 2014 (Pontiki et al. (2014)) and the standard of the great majority of related work, the comparison between systems focuses on $f1$ for ACD and $acc.$ for SP. Considering the rather poor support for some classes in the test set, the evaluation will be also based on the *weighted average* instead of the *macro average*, which in this case should provide a more reliable and accurate insight into the actual performances of the systems. The scores were computed with the classification report implementation of scikit-learn (`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html`).

Tables 5.1 and 5.2 show a direct comparison between the Majority Baseline, and the BERT and RoBERTa models with equal setup. Taking into consideration the scarce

precision and high recall on the *NA* class, which resemble the scores achieved by the baseline, it appears that the BERT system has a tendency on over-assigning the major category. The same is not applicable for RoBERTa, where the evaluation shows a clearer understanding of the ACD task with a precision of over 60%. Overall, the results prove a superiority of RoBERTa against BERT and the Majority Baseline on both tasks. This confirms what was affirmed on (Liu et al. (2019b)), stating that the optimization in the pre-training of the original BERT architecture proves to be effective and leading to enhanced language understanding. In fact, while BERT scores 0.47 *f1* on ACD, RoBERTa shows a 20% improvement (0.67 *f1*) and the capability of detecting more classes. For example, RoBERTa was able to detect the *Information*, *Multiple*, and *Boarding* categories, despite their poor support, demonstrating a better understanding of the differences between the classes. Nevertheless, neither of the fine-tuned models was successful in recognizing the minor classes *Price*, *Safety*, and *Others*. The reasons behind this might be the scarce number of training samples labelled with these, making the systems uncapable to learn the necessary information to interpret them, or the poor support of these categories in the test set, which does not contain enough examples to effectively evaluate the performance. In fact, the labels distribution proved to be unbalanced, with the major class (*NA*) occurring 76 times, against the single one of the *Others* class. This implies that the evaluation for some classes might be more accurate compared to others, and opening the discussion for future evaluation with wider test sets.

Baseline	**p**	**r**	**f1**	**#**
NA	0,31	1,00	0,47	76
All the other categories	0,00	0,00	0,00	173
accuracy			**0,31**	0,71
macro avg	0,02	0,07	0,03	249
weighted avg	0,09	0,31	**0,14**	249

Table 5.1: Classification report of the performance achieved by the Majority Baseline on ACD

Moving on to the SP, the evaluation shows a completely different picture compared to ACD. As shown in Tables 5.3 and 5.4, the performance gap between BERT and RoBERTa is smaller, where both successfully overperformed the baseline. The comparison with the latter also shows us a clearer understanding of the task from BERT compared to ACD, by achieving a more balanced ratio between precision and recall on the major class *Neutral* ($p = 0.72$; $r = 0.90$). Furthermore, the use of a test set consisting of a wider support for the *Negative* class, which is a non-representative distribution of the original dataset, demonstrates that BERT was able to achieve the necessary knowledge for tackling the task without over-assigning the major polarity.
Nevertheless, RoBERTa proves once again its superiority by scoring 0.83 accuracy against the 0.78 of BERT, by confirming the previous considerations. It is also interesting to notice the high efficiency, achieved by both the fine-tuned models, in detecting the positive label despite the lower frequency compared to the other classes. Still, neither of the systems was able to correctly identify *Mix*, which brings to the attention the necessity of finding a solution for handling these problematic cases.

BERT	p	r	f1	RoBERTa	p	r	f1	#
Service	0,40	0,11	0,17	**Service**	0,57	0,42	0,48	19
Company	0,50	0,17	0,25	**Company**	0,76	0,72	0,74	18
Staff	0,91	0,65	0,75	**Staff**	0,86	0,97	0,91	31
Price	0,00	0,00	0,00	**Price**	0,00	0,00	0,00	5
Travel	0,62	0,40	0,48	**Travel**	0,63	0,50	0,56	20
Aircraft equipment	0,62	0,69	0,65	**Aircraft equipment**	0,72	0,88	0,79	26
Food	0,73	0,80	0,76	**Food**	0,75	0,90	0,82	10
Safety	0,00	0,00	0,00	**Safety**	0,00	0,00	0,00	3
Boarding	0,00	0,00	0,00	**Boarding**	1,00	0,22	0,36	9
Luggage	0,75	0,25	0,38	**Luggage**	0,90	0,75	0,82	12
Information	0,00	0,00	0,00	**Information**	1,00	0,13	0,22	8
Others	0,00	0,00	0,00	**Others**	0,00	0,00	0,00	1
Multiple	0,00	0,00	0,00	**Multiple**	1,00	0,09	0,17	11
NA	0,47	0,97	0,63	**NA**	0,66	0,95	0,78	76
accuracy			0,55	**accuracy**			0,71	249
macro avg	0,36	0,29	0,29	**macro avg**	0,63	0,47	0,48	249
weighted avg	0,50	0,55	**0,47**	**weighted avg**	0,72	0,71	**0,67**	249

Table 5.2: Classification report comparing the performances achieved by BERT and RoBERTa on ACD

Baseline	p	r	f1	#
Neutral	0,40	1,00	0,57	99
All the other polarities	0,00	0,00	0,00	150
accuracy			**0,40**	249
macro avg	0,10	0,25	0,14	249
weighted avg	0,16	0,40	**0,23**	249

Table 5.3: Classification report of the performance achieved by the Majority Baseline on SP

BERT	p	r	f1	RoBERTa	p	r	f1	#
Positive	0,86	0,81	0,83	**Positive**	0,79	0,92	0,85	37
Negative	0,83	0,73	0,78	**Negative**	0,91	0,77	0,83	103
Neutral	0,72	0,90	0,80	**Neutral**	0,78	0,94	0,85	99
Mix	0,00	0,00	0,00	**Mix**	0,00	0,00	0,00	10
accuracy			**0,78**	**accuracy**			**0,83**	249
macro avg	0,60	0,61	0,60	**macro avg**	0,62	0,66	0,63	249
weighted avg	0,76	0,78	0,76	**weighted avg**	0,80	0,83	0,81	249

Table 5.4: Classification report comparing the performances achieved by BERT and RoBERTa on SP

Confirming what was already showed during the fine-tuning step, the recommended setup for Roberta (RoBERTa Rec) reports no sensitive improvements on the achieved scores (Figure 5.5 & 5.6). In fact, despite the extremely similar accuracy achieved during training, Robert Rec demonstrates slightly poorer performances compared to the original setup on both tasks. The tables below provide the two classification reports for this system.

Although it performs better than the baseline and BERT, the recommended setup was not successful in enhancing the learning capabilities of the model. On the contrary, some categories for ACD that were identified using the original setup, failed to be recognized with the new one (e.g. Information). Nevertheless, the reasonable use of the major class, displayed by the precision and recall scores, shows a non-applicable tendency for this system in over-assigning the label with the highest frequency. Finally, the F1-score of 0.63 confirms once again the superiority of RoBERTa over BERT, and on the SP side the evaluation shows an almost equal performance with the other RoBERTa model (0.82 *acc.*), even if slightly inferior compared to the previous one.

RoBERTa Rec	**p**	**r**	**f1**	**#**
Service	0,50	0,47	0,49	19
Company	0,74	0,78	0,76	18
Staff	0,85	0,94	0,89	31
Price	0,00	0,00	0,00	5
Travel	0,55	0,30	0,39	20
Aircraft equipment	0,79	0,88	0,84	26
Food	0,82	0,90	0,86	10
Safety	0,00	0,00	0,00	3
Boarding	0,67	0,22	0,33	9
Luggage	0,88	0,58	0,70	12
Information	0,00	0,00	0,00	8
Others	0,00	0,00	0,00	1
Multiple	0,00	0,00	0,00	11
NA	0,61	0,93	0,74	76
accuracy			0,68	249
macro avg	0,46	0,43	0,43	249
weighted avg	0,61	0,68	**0,63**	249

Table 5.5: Classification report of the performance achieved by RoBERTa Rec on ACD

RoBERTa Rec	**p**	**r**	**f1**	**#**
Positive	0,80	0,89	0,85	37
Negative	0,87	0,78	0,82	103
Neutral	0,78	0,91	0,84	99
Mix	0,00	0,00	0,00	10
accuracy			**0,82**	249
macro avg	0,61	0,64	0,63	249
weighted avg	0,79	0,82	0,80	249

Table 5.6: Classification report of the performance achieved by RoBERTa Rec on SP

To summarize, the evaluation conducted on the test set proved that the RoBERTa system adopting the original setup was successful in overperforming the Majority Baseline, BERT, and RoBERTa Rec. It confirmed the already achieved superiority during the validation and demonstrated high efficiency for both the tasks. A further analysis of its results will follow.

5.3 Results Analysis

The results that are analyzed in this section are presented and discussed through the use of two means. First, a confusion matrix in the form of a heatmap displays the correspondence between gold label and prediction for the Aspect Category Detection, followed by one for the Sentiment Polarity classification. Then, a discussion over some representative examples of the performance of the system is provided. The analyzed results strictly regard the best performing model, hence, the RoBERTa one using the original setup. This section contains references to the annotation guidelines that can be found in the corresponding Appendix, where the details about the categorization of the data are described.

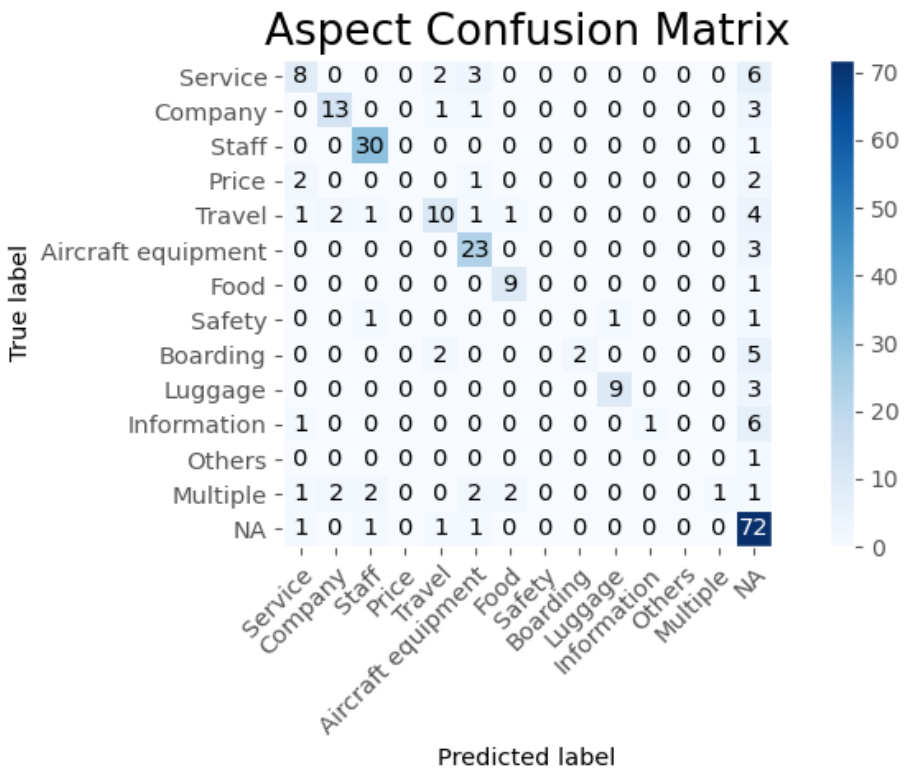

Figure 5.4: Confusion matrix on the correspondence between gold label and prediction for ACD of the RoBERTa system

As Figure 5.4 related to ACD shows, RoBERTa performs an almost perfect classification for some classes and struggles with others. The highest efficiency is shown in the prediction of classes that had a clearer definition in the annotation guidelines and are characterized by easy patterns to capture. An example is the *Food* category, which is usually recognizable by mentions of any kind of food or drinks, and related elements, like verbs or adjectives, that occur in the same context. In this case, if the system sees the combination of *tasty* and *sandwich*, it will not have any particular doubts in assigning the food related label.
The same consideration is not applicable when the system encounters the class *Service*,

where it shows to have had a difficult time. As already showed in Chapter 3, *Service* represents the third most occurring class, characterizing 8.82% of the entire dataset. Given the high frequency of this label, the system should be theoretically better at identifying *Service* than *Food*, which represents 6.01% of the data. Instead, the figure shows that RoBERTa misclassified this category 11 times over 19, demonstrating that having an easily recognizable pattern might contribute more than having a high number of training samples. The following example (1) is a clear representation of described phenomenon. Here, in fact, the system fails in understanding that the element receiving the judgement is the poor availability of selection for the seats, which falls under the *Service* category, and not the long-haul flight. This could be happening because, during training, an association between the word *time*, *flight*, and an adjective, might have been created and linked to the *Travel* category, where delays or comfort of the flight are taken into consideration. Given the lower frequency of the predicted class (5.61%) compared to the gold one, this example serves a further proof that an easily recognizable pattern might have a more impactful influence on the performance than a higher number of training samples.

1. "*This time I was also unable to select a seat which was annoying for the long haul flight that I was on.*"
GOLD ASPECT: *Service* **PREDICTION:** *Travel*

Another kind of issue that is encountered during the classification is the ambiguity of some classes. In the example below (2), this phenomenon is illustrated, as the sentences that talk about online services, like WhatsApp in this case, are usually labelled with *Service*. For this reason, even though the element that is being qualified as "very good" is the efficiency in providing information, it is understandable that a confusion with the channel used for the communication might occur. Therefore, considering the overlap between categories in cases like the one described, future work should evaluate the possibility of refining the annotation guidelines in order to avoid doubts generated by non-perfectly differentiated classes.

2. "*Used WhatsApp and received info - very good*"
GOLD ASPECT: *Information* **PREDICTION:** *Service*

One of the most evident type of mistakes regards the sentences containing more than one category, which are labelled with *Multiple*. As described in the Annotation Guidelines (Appendix A), the *Multiple* label was used when more than one AC was present within the same sentence. The presence of this class leads to some issues especially for Language Model based systems. In fact, the correct association between multiple elements and related opinions within the same sequence, becomes a challenging task when syntactical features, usually restricted to traditional Machine Learning, are not implemented. This is a phenomenon that was already observed in related work, where the limitations of language models over complex linguistic tasks (as ACD is) were pointed out (Liu et al. (2019a)). A concrete representation of this is the example below (3), within which the presence of both the *Boarding* and *Food* categories is not captured, and just one of the two is recognized. What the output shows is that the system was not able to make a successful association between the occurrence of two categories and the label *Multiple*, outputting only partially correct predictions, where just

one of the present classes is identified. To tackle this problematic case, an experimentation with sub-phrase level annotation might be a viable solution. This would involve an extra preprocessing step consisting in splitting the sequences into sub-phrases, and the identification of just one category per sequence. The fact that these sentences are already labelled with *Multiple* will make them easier to be isolated and addressed in future research.

3. "*Boarding is often a mess... and the meals really are not that good!*"
GOLD ASPECT: *Multiple* **PREDICTION:** *Food*

For what concerns the classes with not significant frequency, among which *Price*, *Safety*, and *Boarding*, it still needs to be assessed whether more examples should be annotated and provided to the system in order to correctly recognize their characteristics, or simply merging them to bigger classes. On this concern, avoiding the latter might be preferable in some cases, as the level of information provided by a fine-grained category brings more detailed insight in the customer experience. Refining the annotation guidelines with more easily recognizable patterns (as done with the *Staff* category) might also enhance the categorization of these feedbacks. In addition, removing the *Others* label, considering its rather rare frequency and questionable importance, might bring a clearer layout of the categories and reduce the noise in the evaluation.

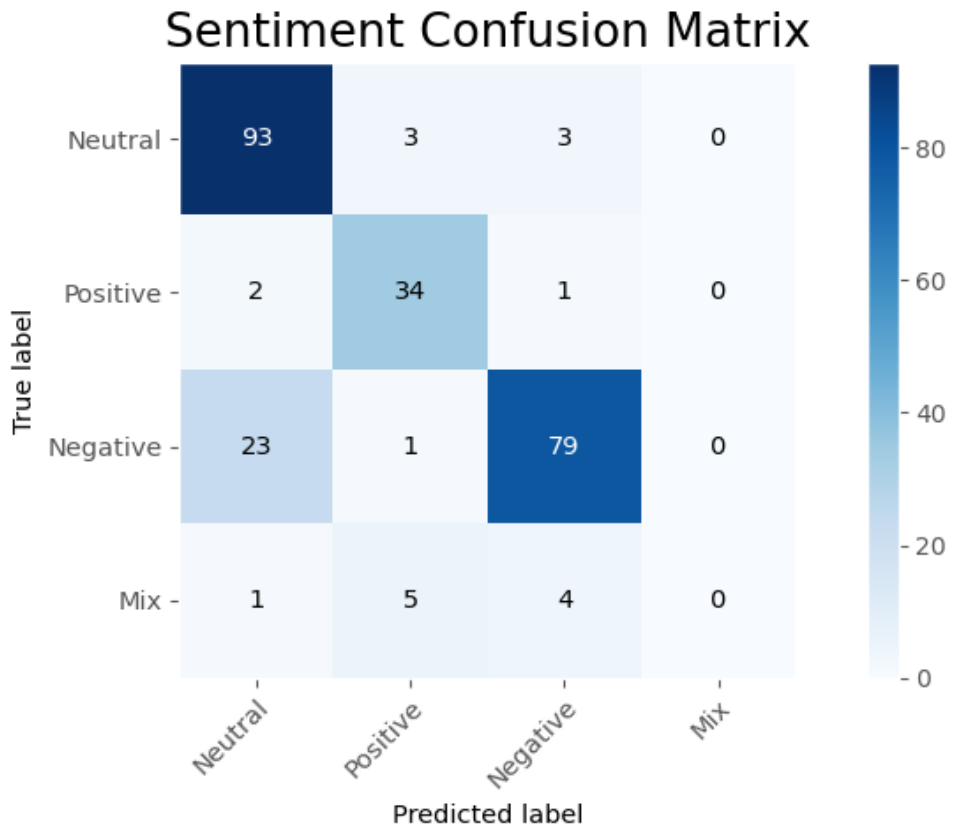

Figure 5.5: Confusion matrix on the correspondence between gold label and prediction for SP of the RoBERTa system

Moving on to the SP classification, the system shows a clearer understanding of the task, as previously mentioned. The related confusion matrix (Figure 5.5) shows that no particular doubts were encountered by the model, except for the *Mix* class. In fact, a great performance is evident in the prediction of the *Neutral* and *Positive* polarities, which were almost always correctly identified. This result gains further importance when considering that the *Positive* class is representing just the 16% of the dataset, being the smallest after *Mix*. This proves that the reflections made over the greater impact of well defined patterns might also apply for the sentiment classification. Nevertheless, an analysis of the mistakes of the SP classifier might provide a better overview on what could still be improved. A representative example of the challenging cases that were encountered, is the one that follows (4).

4. "*Pilot is the Main Man.*"
GOLD SENTIMENT: *Positive* **PREDICTION:** *Neutral*

This sentence contains one of the typical characteristics of user-generated text: colloquialism. In fact, here the customer is expressing a personal appreciation for the pilot by defining him "the Main Man". This shows a use of the adjective *main* as a positive connotation that would be not contemplated in the standard use of the English language. The also rather uncommon frequency of this expression, makes it more difficult for the system to correctly assigning the polarity label, which was identified as *Positive* by the human annotator. The reason behind this might be that the system probably never learned the use of the adjective *main* as a positive quality in the training step, instead, making the word occupy a vector space which is closer to features that are associated with *Neutral*. For future research, it might be worth verifying the frequency of these cases and, consequently, evaluating their importance in the dataset. If their occurrence is not rare, it might be interesting to provide more examples containing colloquialisms to the system in order to improve the classification of sentences characterized by these expressions. In alternative, a comparative analysis between RoBERTa and a language model trained over tweets, in which the use of unstandardized language is more common, could bring some extra insights.

5. "*My outbound flight on [competitor company] comfort econmy was a much more pleasant experience.*"
GOLD SENTIMENT: *Negative* **PREDICTION:** *Positive*

On this last example (5), the misclassification regards the interpretation of the text. In fact, the sentence above is expressing an implicit negative feedback towards the airline company by expressing a comparison with a competitor company. In this comparison, the competition is receiving an appreciation, which the system was not able to recognize. The association between the positive connotations expressed in the sentence and a competitor company is easy to interpret for humans, because of the knowledge of the world that they share, but not yet for a language model (Zellers et al. (2019)). This problematic case opens to new discussions on whether providing more examples (if available and frequent) labelled as *Negative*, within which positive connotations associated with competitors names would improve the performance, or it would worsen it creating a misleading interpretation of the features and the related polarities. This last point underlines the legitimacy of the debate regarding the effective efficiency

regarding language understanding of these new state-of-the-art models in NLP technology.

Chapter 6

Conclusion & Discussion

6.1 Summary of the research

This research focused on the extraction of sentiment opinion related to certain aspects of a flight from customer feedbacks. The goal behind the retrievement of this information was the in-depth insight into the passengers satisfaction, which could serve as a key tool for an efficient improvement of the service of the company that commissioned this work. The main approach consisted of two steps: first, data annotation, where three annotators manually labelled 2495 sentences with Aspect Category and Sentiment Polarity; second, design and implementation of fine-tuned systems, based on BERT and RoBERTa models, for the classification of categories and polarities separately. The value of this approach stems from the fact that no previous work was carried out using this method on customer feedback data from this domain.

The comparative evaluation of the systems against a majority baseline demonstrated the superiority of RoBERTa, using a setup consisting of 4 Epochs, 63 Batch size, 64 Sequence length, 2e-5 Learning rate, and AdamW Optimizer. The efficiency of this system was proved by the achievement of 0.67 f1-score on Aspect Category Detection and 0.83 accuracy on Sentiment Polarity classification.

6.1.1 Answer the research question

The relatively high performances showed by the model serve as an answer to the Research Question, which was presented as follows:

Can Transfer Learning, applied to a small set of annotated data, be a solution for domain-specific Aspect Based Sentiment Analysis?

The answer to this matter becomes affirmative, because transferring the linguistic knowledge of a large language model, like RoBERTa, to perform a different classification task proved to be an efficient solution even with small amounts of annotated data. Furthermore, the strength of this approach is underlined by a more sustainable use of the resources, consisting in a time and costs reduction for more extensive annotations, and by the illustration of the potentialities of state-of-the-art tools applied to real life business scenarios. The use of this technique also showed to be preferable for its rela-

tive simplicity of implementation compared to more traditional feature-heavy Machine Learning approaches, which require a higher level of linguistic expertise. Nevertheless, the analysis of the results also opened the discussion for future study and improvements.

6.2 Discussion & Future Directions

Considering the observations made over the outcome of the classification, the system demonstrated a higher degree of confidence in the identification of classes that were characterized by easy patterns to capture. This result is the reflection of the likely presence of fuzzy categories among the ones established for the ACD task, derived by an imperfect definition of the guidelines. This problem originates from the strong relatedness of this task to a certain domain and the complicated reproducibility of approaches due to the limited domain adaptability. In fact, although the example of related work was taken into consideration in the design of the methodology of this project, the definition of the annotation guidelines lacked the accuracy achieved through years of in-depth studies over different domains. This was the case for related work based on the SemEval guidelines concerning laptops, restaurants, and hotels reviews, where the fine-grained categories for those specific tasks provided a high impact on the outcome of the classification. On the other hand, the fact that the services between airlines, and the topics addressed in customer reviews and surveys, vary in a considerable way, a definitive categorization of the data becomes hardly achievable. In fact, while in the reviews (or tweets in the case of Ashi (2019)) the audience is formed by other potential customers, the feedbacks contained in a survey is requested by, and addressed to just the company, with the improvement of a service as a goal. These factors, combined to the restricted time and comparative resources availability for this project, made the achievement of fine-grained categories a challenging objective. Taking these points into consideration, future work should contemplate the elaboration of a refined separation between categories, starting from the ones that generated doubts or were confused for others (e.g. *Service*), and the incorporation of rare labels to more frequent ones (e.g. *Safety*).

The discussion should also be extended to the performance gap between ACD and SP classification. In fact, although the model adopted the same setup for both tasks, the sentiment related one proved to be an easier classification to handle, probably due to a smaller number of classes, which entails a lower probability of mispredictions (1 against 3 for SP, and 1 against 13 for ACD), and to clearer patterns for the identification of each polarity label. Furthermore, considering the higher degree of domain adaptability of the sentiment classification, compared to ACD, the example of years of study conducted in related work acquires more relevance and demonstrates to be a strong source for improvement. Nevertheless, the issues encountered with the feedbacks containing multiple classes, still remain to be addressed for both ACD and SP classification. The fact that these cases are already labelled with *Multiple* or *Mix* will make them easier to be isolated and addressed in future research, where an experimentation with sub-phrase level annotations should be taken into consideration as a viable solution.

Furthermore, it still needs to be established whether using different architectural

setups for each task could lead to more efficient results. It would be interesting to explore the possibilities derived by combination of this option to the use of a more balanced dataset, especially for testing. In fact, using small sets of data brings the problem of in depth evaluation and interpretation of the results. The use of bigger test sets, and trainings on different setups, might bring a better understanding of how the models compare between each other. Therefore, it would be interesting if future work focused on trying setups that were not necessarily recommended in Liu et al. (2019b), as suggested by Ruder et al. (2019). Some practical suggestions include experimenting with higher learning rates (e.g. >5e-5) during the fine-tuning, or applying a lower learning rate for more than 5 epochs for ACD, in order to keep a balanced setup and avoid overfitting. This experiments should be evaluated over a test set that contains at least 1000 samples or, ideally, a set that shows a more balanced distribution of the labels.

More research could also involve the use language models of comparative size to BERT that were pre-trained over user generated data, like tweets, and evaluate whether any improvement on the correct classification of colloquialism is achievable.

Lastly, considering the roots of the company that commissioned this project, future study might also explore the development of systems based on this project for the analysis of feedbacks in Dutch language. The cross language adaptation of these systems appears complicated when using traditional machine learning tools, which rely on heavy linguistic features engineering. However, the continuous development conducted on language modelling provided us not only Dutch-specific version of both BERT and RoBERTa, but also multilingual implementations for high resources languages. For this reason, a comparative analysis of performances between BERTje (de Vries et al. (2019)), RobBERT (Delobelle et al. (2020)), Multilingual BERT (`https://github.com/google-research/bert/blob/master/multilingual.md`), and the English models might bring interesting insights in the reproducibility of results in a multilingual application.

Appendix A

Annotation Guidelines

Annotation Guidelines
for
Aspect Based Sentiment Analysis

Introduction

These annotation guidelines are inspired by the ones used in the SemEval 2014 Task 4 (Pontiki et al., 2014). The purpose of this annotation for Aspect Based Sentiment Analysis is to identify aspects and related sentiment polarity within sentences.

For this task, the customer feedback data provided by the airline KLM will be used. The feedbacks were left by the customers after their flight and contain their considerations about the service provided by the company. Through these feedbacks, it becomes possible to have an insight about the satisfaction of the customers and understand the strengths and weaknesses of the product that the company is offering. In order to retrieve this information, the annotator should try to identify and annotate the following:

- **Aspect Category**

  The category of the aspect discussed in the sentence. The category is selected from one of the following aspect categories:

  a. *Service*
  b. *Company*
  c. *Staff*
  d. *Price*
  e. *Travel*
  f. *Aircraft equipment*
  g. *Food*
  h. *Safety*
  i. *Boarding*
  j. *Luggage*
  k. *Information*
  l. *Others*
  m. *Multiple*
  n. *NA (no aspect)*

- **Sentiment Polarity**

  The sentiment polarity expressed over the aspect of the sentence. The polarities are:

  - *Positive (1)*
  - *Negative (-1)*
  - *Neutral (0)*
  - *Mix (2)*

Figure A.1: pg.1

- **Aspect Term**

  The aspect term is a single or multiword term contained in the sentence that provides the information about the aspect category. In the following example the aspect term is underlined:

  The snacks that were distributed during the flight were really tasty!

  If not explicit, the annotator should type *na* (no aspect) as label.

The table below is a visual representation of how the annotation should look like.

Sentence_ID	Feedback_ID	Sentence	Aspect_Category	Sentiment	Aspect_Term
1	129288273	The snacks that were distributed during the flight were really tasty!	g	1	snacks

In the next section it will be possible to read the detailed information about the modality of annotation for each of the points mentioned above.

The annotation will be carried out on Microsoft Excel. In the annotation process, the sentences must be considered singularly and not in combination with preceding or following ones. Being the data private, in accordance with the security protocol of the company, after the annotation has been completed using Excel, the file must be sent back to the development environment of Underlined and deleted from the private environment.

The aspect category

The category of the aspect discussed in the sentence. It provides the information regarding the category of what is being discussed in the sentence and is receiving a judgment. The category is selected from one of the following aspect categories and should be labelled with the corresponding letter of the alphabet below:

a. *Service :* this includes the general service provided by the company or more specific services like cleanness, disability support, range of products offered, type of flight (economy, business, etc.), website, booking process;

b. *Company :* this includes all the direct mentions to the company or also indirect through comparison with other companies;

Figure A.2: pg.2

c. *Staff :* this includes all the mentions to the cabin crew, pilots or any other employee of the company (on-board and off-board), including customer service staff;

d. *Price :* this includes the mentions to the price of the ticket or the services provided by the company, including food and products sold on board. If a mention to the price of the food, for example, is made within the sentence, the ***d*** category should be preferred over the ***g*** category;

e. *Travel :* this includes the mentions to the flight, delayed or on time departures and arrivals, turbulences and comfort;

f. *Aircraft equipment :* this includes all the parts and equipment of the aircraft (e.g. the seats, the design, architectural choices) and the flight bundle (e.g. movies, headphones);

g. *Food :* this includes all the mentions to food or drinks included in the ticket price or sold on board;

h. *Safety :* this includes all the mentions to the safety equipment, measures and services, like security checks, provided by the company. Also the mentions to the emergency exits should be included here (and not in ***f***). The waiting time at the security checks is not included as it is not related to the safety, but to the boarding process (***i***);

i. *Boarding :* this includes all the mentions to the boarding process, like the waiting time, respect of the priorities, documents check and gates organization. Also the mentions to the waiting time at the airport and at the security checks should be included here;

j. *Luggage :* this includes all the mentions to the handling, storing, tracking, weighting and accidents about luggage;

k. *Information :* this includes all the mentions to the information provided by the company regarding the flight and services and through any channel (e.g. announcement on board or at the gates, website);

l. *Others :* this includes all the mentions to categories that are not present in the list;

m. *Multiple :* this includes all the mentions to multiple categories in a sentence. Obviously, if one of the mentioned categories in the sentence is not the object of any judgement, it should not be considered;

n. *NA (no aspect) :* this should be used when no aspect category is present, like in informative or anecdotal sentences that do not contain a judgment (e.g. "*I was travelling for business.")*

Figure A.3: pg.3

The sentiment polarity

As described in (Pontiki et al., 2014), the sentiment polarity is expressed through attitudes, opinions, evaluations, emotions, or feelings etc. of an opinion holder towards the aspect term within the sentence. The sentiment polarity is selected from one of the following polarities and should be labelled with the corresponding number below:

- *Positive (**1**):* a sentence should be annotated with this label when the aspect term (or multiple aspect terms) within it is described with *positive* or *some-what positive* qualities or expressions.

  *The snacks that were distributed during the flight were really tasty! (**1**)*

  In this example, the aspect term *the snacks* is described with positive characteristics (*really tasty*). Therefore, it should be labelled with ***1***;

- *Negative (**-1**):* a sentence should be annotated with this label when the aspect term (or multiple aspect terms) within it is described with *negative* or *some-what negative* qualities or expressions.

  *This was an older plane so just not quite as nice as some of the others. (**-1**)*

  In this example, the aspect term *older plane* is described with overall negative characteristics (*not quite as nice as*). Therefore, it should be labelled with ***-1***;

- *Neutral (**0**):* a sentence should be annotated with this label when the aspect term within it is not described with explicitly positive or negative qualities or expressions. This label should also be used in informative or anecdotal sentences that do not contain any aspect term.
  - *I was travelling for business. (**0**)*
  - *I think freedom is about choices. (**0**)*

  In this examples, no explicit sentiment is identified. Therefore, they should be labelled with ***0***.

  Note: as already described in (Pontiki et al., 2014), “if a sentence conveys both neutral and negative (or positive) opinions about an aspect category, then the negative (or positive) polarities dominate over the neutral ones”;

Figure A.4: pg.4

- *Mix (**2**):* as described for the *conflict* label in (Pontiki et al., 2014), a sentence should be annotated with this label when the aspect term (or multiple aspect terms) within it is described with both *negative* and *positive* qualities or expressions.

  *The meal was great, but the toilette was disgusting. (**2**)*

  In this example, the aspect terms *the meal* is described with a positive adjective (*great*), while the aspect term *the toilette* is described as negative (*disgusting*). Therefore, the sentence contains multiple polarities and should be labelled with ***2***.

When no aspect term is present, the overall sentiment of the sentence should be annotated.

The aspect term

The aspect term is a single or multiword term contained in the sentence that provides the information about the aspect category. Aspect terms are nouns or (rarely) verbs. If a sentence contains multiple aspect terms, they must also be annotated and separated by "," and a space as in the first example below. If the aspect term contains typos (typical of user generated text), the aspect term should be annotated with the typo, hence, without any correction. If not explicit, the annotator should type ***na*** (not available) as label. In the following examples the aspect terms are underlined:

- *The meal was great, but the toilette was disgusting. (**meal, toilette**)*
- *Overall reasonably priced. (**priced**)*
- *It was not expensive. (**na**)*

In the first example, multiple aspects terms are present in the form of nouns (*meal, toilette*). The second example shows how an aspect term can be identified in verbs through the possibility of rephrasing the sentence (*reasonable price*). In the third example, ***na*** is assigned because, even if the term *price* is inferred by the adjective *expensive*, it is not explicitly mentioned. In informative or anecdotal sentences, the aspect term should be labelled also as ***na*** because no judgment is expressed.

Figure A.5: pg.5

Examples

Sentence_ID	Feedback_ID	Sentence	Aspect_Category	Sentiment	Aspect_Term
#	#	The snacks that were distributed during the flight were really tasty!	g	1	snacks
#	#	The meal was great, but the toilette was disgusting.	m	2	meal, toilette
#	#	It was not expensive.	n	1	na
#	#	I was travelling for business.	n	0	na
#	#	I was upset because my suitcase got lost.	j	-1	suitcase

References

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

Figure A.6: pg.6

Bibliography

M. M. Ashi. Pre-trained Word Embeddings for Arabic Aspect-Based Sentiment Analysis of Airline Tweets. page 11, 2019.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*, June 2017. URL `http://arxiv.org/abs/1607.04606`. arXiv: 1607.04606.

C. Brun, D. Popa, and C. Roux. XRCE: Hybrid Classification for Aspect-based Sentiment Analysis. page 5, 2014.

G. Carenini, R. T. Ng, and E. Zwart. Extracting Knowledge from Evaluative Text. page 8, 2005.

W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim. BERTje: A Dutch BERT Model. *arXiv:1912.09582 [cs]*, Dec. 2019. URL `http://arxiv.org/abs/1912.09582`. arXiv: 1912.09582.

P. Delobelle, T. Winters, and B. Berendt. RobBERT: a Dutch RoBERTa-based Language Model. *arXiv:2001.06286 [cs]*, Sept. 2020. URL `http://arxiv.org/abs/2001.06286`. arXiv: 2001.06286.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. page 8, 2009.

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

M. Fernández-Gavilanes, T. Álvarez López, J. Juncal-Martínez, E. Costa-Montenegro, and F. Javier González-Castaño. Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, 58:57–75, Oct. 2016. ISSN 0957-4174. doi: 10.1016/j.eswa.2016.03.031. URL `https://www.sciencedirect.com/science/article/pii/S0957417416301300`.

J. Firth. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society, Oxford, 1957. reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

A. García-Pablos, M. Cuadros, and G. Rigau. W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis. *Expert Systems with Applications*, 91:127–137, 2018. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2017.08.049. URL `https://www.sciencedirect.com/science/article/pii/S0957417417305961`.

Y. Goldberg. A Primer on Neural Network Models for Natural Language Processing. *arXiv:1510.00726 [cs]*, Oct. 2015. URL http://arxiv.org/abs/1510.00726.

M. Hu and B. Liu. Mining and Summarizing Customer Reviews. page 10, 2014.

D. Jurafsky and J. H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2020. URL https://web.stanford.edu/~jurafsky/slp3/.

S. Kiritchenko, X. Zhu, C. Cherry, and S. Mohammad. NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews. *R F*, page 6, 2014.

G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. RACE: Large-scale ReAding Comprehension Dataset From Examinations. *arXiv:1704.04683 [cs]*, Dec. 2017. URL http://arxiv.org/abs/1704.04683. arXiv: 1704.04683.

N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith. Linguistic Knowledge and Transferability of Contextual Representations. *arXiv:1903.08855 [cs]*, Apr. 2019a. URL http://arxiv.org/abs/1903.08855. arXiv: 1903.08855.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. page 13, 2019b.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, Sept. 2013. URL http://arxiv.org/abs/1301.3781. arXiv: 1301.3781.

G. A. Miller. WordNet: a lexical database for English. 38(11):3, 1995.

S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct. 2010. ISSN 1558-2191. doi: 10.1109/TKDE.2009.191. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

J. Pennington, R. Socher, and C. Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL http://aclweb.org/anthology/D14-1162.

J. Phang, T. Févry, and S. R. Bowman. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. *arXiv:1811.01088 [cs]*, Feb. 2019. URL http://arxiv.org/abs/1811.01088. arXiv: 1811.01088 version: 2.

M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. page 9, 2014.

M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. page 10, 2015.

M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *International Workshop on Semantic Evaluation*, pages 19 – 30, San Diego, United States, Jan. 2016. doi: 10.18653/v1/S16-1002. URL `https://hal.archives-ouvertes.fr/hal-01838537`.

S. Poria, E. Cambria, A. Gelbukh, F. Bisio, and A. Hussain. Sentiment Data Flow Analysis by Means of Dynamic Linguistic Patterns. *IEEE Computational Intelligence Magazine*, 10(4):26–36, Nov. 2015. ISSN 1556-6048. doi: 10.1109/MCI.2015.2471215. Conference Name: IEEE Computational Intelligence Magazine.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving Language Understanding by Generative Pre-Training. page 12, 2017.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv:1606.05250 [cs]*, Oct. 2016. URL `http://arxiv.org/abs/1606.05250`. arXiv: 1606.05250.

S. Ruder, P. Ghaffari, and J. G. Breslin. INSIGHT-1 at SemEval-2016 Task 5: Deep Learning for Multilingual Aspect-based Sentiment Analysis. page 7, 2016.

S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. Transfer Learning in Natural Language Processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-5004. URL `https://aclanthology.org/N19-5004`.

J. Saias. Sentiue: Target and aspect based sentiment analysis in semeval-2015 task 12. Association for Computational Linguistics, 2015.

K. Schouten and F. Frasincar. Survey on Aspect-Level Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830, Mar. 2016. ISSN 1558-2191. doi: 10.1109/TKDE.2015.2485209. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

R. Sennrich, B. Haddow, and A. Birch. Neural Machine Translation of Rare Words with Subword Units. *arXiv:1508.07909 [cs]*, June 2016. URL `http://arxiv.org/abs/1508.07909`. arXiv: 1508.07909.

C. Sun, L. Huang, and X. Qiu. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. *arXiv:1903.09588 [cs]*, Mar. 2019. URL `http://arxiv.org/abs/1903.09588`.

Z. Toh and J. Su. NLANGP: Supervised Machine Learning System for Aspect Category Classification and Opinion Target Extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 496–501, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2083. URL `https://aclanthology.org/S15-2083`.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, Dec. 2017. URL `http://arxiv.org/abs/1706.03762`. arXiv: 1706.03762.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv:1804.07461 [cs]*, Feb. 2019. URL `http://arxiv.org/abs/1804.07461`. arXiv: 1804.07461.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144 [cs]*, Oct. 2016. URL `http://arxiv.org/abs/1609.08144`. arXiv: 1609.08144 version: 2.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a Machine Really Finish Your Sentence? *arXiv:1905.07830 [cs]*, May 2019. URL `http://arxiv.org/abs/1905.07830`. arXiv: 1905.07830.

Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. pages 19–27, 2015. URL `https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html`.