ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

MSc. Economics & Business Economics
Specialization in Data Science & Marketing Analytics

Uncovering fine-grained customer sentiment for product and service improvements

Developing a multi-task deep learning framework for extracting business-relevant aspects and associated customer sentiment from customer service calls

December 14, 2022

Author:
K.L. (Karsten) de Neve LL.M.
453427

Supervisor:
Prof. dr. A.C.D. (Bas) Donkers

Second assessor:
dr. F. (Flavius) Frasincar

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Acknowledgements

I would like to express my gratitude for their support and guidance for all those that I have been working throughout the completion of my master thesis. I would like to thank KPN for the opportunity to cooperate on this thesis project, in particular Gianluigi Bardelloni, my KPN supervisor and AI Lead at KPN and Niek de Win my team manager.

I would also like to specifically thank my supervisor, professor Bas Donkers, for his valuable feedback and flexibility throughout the process that greatly enhanced the quality of this work.

I am grateful for to all those with whom I have had to the pleasure to work with during my thesis internship at KPN, those at the department of Advanced Analytics and in particular my fellow team members at Data Science Lab, what I have been part of couple of months.

A special thanks to fellow team members Reinier Bekkenutte, Otis Vabolis, Panagiotis Banos, Georgios Vlassopoulos and Ralph van Ierland for their valuable insights and feedback throughout the process. Your expertise and guidance in Data Science have been invaluable, and I am grateful for your support.

Lastly, I want to thank Esmee, my family and friends, for their everlasting trust and support.

Thank you all for your support and collaboration,

Karsten

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ABSA	Aspect-Based Sentiment Analysis
ASC	Aspect Sentiment Classification
ATE	Aspect Term Extraction
ATESC	Aspect Term Extraction and Aspect-Sentiment Classification
APC	Aspect Polarity Classification
BERT	Bidirectional Encoder Representations for Transformers
CDM	Context Dynamic Masking
CDW	Context Dynamic Weighting
DBRD	Dutch Book Reviews Data
LCF	Local Context Focus
NER	Named-Entity Recognition
NLP	Natural language processing
SRD	Semantic Relative Distance
UGC	User-Generated Content

Chapter 1. Introduction

Communication between firms and their consumers is a vital facet of customer services and marketing. KPN receives several millions calls and chats per year from its customers. This communication encompasses useful information. However, it would be inefficient for large service providers, such as KPN, to manually read all transcripts to filter out business-relevant information for service and product improvements.

It is imperative that in order to be competitive in the market, firms must meet their customer expectations. Misfit between customer needs and design inputs has negative implications for customer satisfaction (Aguwa et al., 2012). Several studies showed a strong connection between customer satisfaction and customer loyalty, or its opposite churn.

The sheer volume of these customer interactions makes it by nature difficult to translate all this (customer) feedback data to large-scale insights regarding product and service quality and potential problems in the service chain. A suitable method needs to be found to analyze customer sentiment in their feedback to KPN by associating specific sentiments with different aspects of the product or service that they receive from KPN. Finding a scalable approach that can be explained and evaluated (tested) is of high business value to KPN.

1.1 Research question

In accordance with the problem statement, this research focuses on providing a method to provide (actionable) insights on customer needs and satisfaction from customer service calls with respect to aspects of products and services. This urges the following research question:

"How can opportunities for product and service improvements be identified from customer service calls using a data-driven approach?"

Additionally, it is important to mention that a sound solution for this research problem also depends on business usability. Therefore, this research was conducted in collaboration with a major telecom provider in the Netherlands, Royal KPN. In machine learning research, the focus is usually on providing solutions that are technically beautiful and excellence, judging results on their metric performance. But, what is perhaps even more challenging is to design a method that has managerial relevance and does not lose sight of its end-users from business teams. In terms of business relevance, the ability to generate new managerial insights and interpretability of results are important to create actionable insights. This research therefore focuses on providing a method which can be embedded in existing methods within KPN as best as reasonably possible, while emphasizing that an important objective for the solution is usability for business teams. The following five sub-questions in terms of constructive steps of the research process arise to answer the research question:

I. How can sentiment be extracted from customer service calls?

Sentiment analysis of textual data is widely used in marketing to measure sentiment or emotionality in communication with customers. This can be used as a proxy to measure customer satisfaction. This is traditionally done with a sentiment lexicon that contains words that carry a certain sentiment. This research requires a more novel approach of sentiment analysis due to its specific data domain and given the problem at hand. This is discussed further under sub-question 3.

II. *In what way can business-relevant aspects related to products or services be identified in service calls and at what level of depth is this useful and possible?*

Answering this research question should provide a framework or approach to determine what (type of) aspects of KPN's products and services are important to look at, which in turn should be used to formulate business-relevant aspects.

III. *How can the identified sentiment be linked to the aspects?*

This sub-question should ultimately provide the research method, the previous two sub-questions taken into account. There are several ways to tackle this linkage problem. The research field concerned with finding the sentiments of aspects in a text is called Aspect-based Sentiment Analysis (ABSA).

Besides designing a method for the research question, it is also relevant how the results of this method can subsequently be interpreted and how insights can be generated. Therefore, in addition to the the previous method-oriented sub-questions, two additional managerial or outcome-oriented sub-questions are formulated:

IV. *How does sentiment differ among different aspects and how can this be visualized?*

Depending on how aspects are listed, is it possible to differentiate between sentiment per aspect between products, or to discriminate between sentiment found for different aspects? This has potential to create actionable insights for managers, for example what attributes of products need to be improved.

V. *To what extent can identified sentiment for product or service aspects be tracked over time and what is an appropriate user interface?*

Next to analysis of sentiment among aspects, is it also possible to measure the frequency of sentiment for specific aspects across time? Analyzing frequency and changes in sentiment can be very useful for evaluating product updates or improvements, for example.

1.2 Academic relevance

This study is innovative in its data and methodology. First, the data contains Dutch text. State-of-the-art research in Natural Language Processing (NLP) is primarily conducted on English data. Secondly, the data contains service-related direct interactions with customers, while aspect-based sentiment analysis is predominantly run on product review data or on public sources, such as social media or news articles. More so, aspect-based sentiment analysis is usually performed on famous benchmark datasets, where researchers try to beat the benchmark score with a novel method, without a focus on practical applications. The dominance (and necessity) of these prepared benchmark datasets means that published research on aspect-based sentiment analysis is nearly always done on a few well-known datasets; such as restaurant or laptop reviews or twitter data (Brauwers & Frasincar, 2021). Applying ABSA to an out-of-domain dataset, on 'real world' data - that is unlabeled, uncleaned data - is expensive and therefore scarce.

Because much (labeled) text data is English, in terms of methodology non-English machine learning models are relatively uncommon. Other Dutch aspect-based sentiment analysis research has been done by De Clerq et al. (2016) and De Clerq et al. (2017). Although this research provided valuable labeled Dutch text for aspect-based sentiment classification, its research methods are outdated due to rapid development in NLP. Moreover, this research uses a BERT model that is a state-of-the-art model for ABSA in academia. Its application in business is rare. There is a gap between academia and big tech, e.g. Google and Facebook, and other companies in the field of text analytics. Moreover, the majority of existing studies tackle the problems of aspect term extraction and aspect-sentiment classification independently (Akhtar et al., 2020). However, this research addresses these tasks in a unified multi-task framework. Furthermore, although text analysis and aspect-based sentiment analysis could turn raw text into valuable insights for marketing, it is not applied a lot in academic marketing literature because of the novelty of the method and complexity of NLP for non-data scientists. Recent advances in text analytics and NLP are largely accomplished outside the field of marketing. Berger et al. (2020) therefore point out that there is ample opportunity to apply marketing theory to the domain of text analytics and machine learning. Specifically, the information exchanged between consumers and firms via call centers and chats can be leveraged for interesting insights.

1.3 Business relevance

This research was conducted in cooperation with KPN and has business value for KPN in a few possible ways. Customer service calls reflect inbound communication between customers and the firm which are likely to have critical consequences for the relationship between them (Berger et al., 2020). Moreover, while communicational aspects are of paramount importance for relational measures such as commitment, relationship satisfaction, trust and relationship quality, actual word-level research about B2B communication is very scarce (Berger et al., 2020).

In a more practical sense, when sentiment can be related to characteristics of products, it is possible to understand specific aspects customers like or dislike about a service. Ideally, it would be so precise that KPN is able to take targeted actions to improve product or service features. It would also be possible to compare sentiment across multiple aspects of products and services on aggregate, but also for the same aspects of products, e.g. the price of service X versus Y. Further, it would be possible for KPN to track over time how customer sentiment changes toward specific features of a product or service. Monitoring products and services over time is interesting to see if problems occur for updates or patches or to see if problems exist in the service chain. It could even be possible to do this in real time. An advantage of a data-driven approach is that extracted customer needs and problems can be updated continuously and that a data-driven approach has much higher scalability compared to classic customer interviews or focus groups. Once an initial model is built, it is relatively easy to train the model on new aspects or extend it to another market or domain.

The initial implementation of the model does not necessarily need to fulfill all the above mentioned. If the method has proven its value it can be extended or refined to be able to get more actionable insights. Other potential future endeavors for this model are summed up in Chapter 6.

1.4 Outline

The following chapter is dedicated to creating a theoretical framework for the research problem. In this chapter there will be a focus on marketing literature. Since this work is also heavily method-driven, apart from the marketing angle on the problem the Chapter 2 will also outline aspect-based sentiment analysis. Chapter 3 will extend this with relevant work concerning the NLP machine learning algorithms in order to provide a background to and understand the research methodologies further discussed in Chapter 3. In Chapter 3 the research methodology will be discussed in which primarily sub-questions 1 to 3 will be discussed. In Chapter 4, the data will be described along with some descriptive statistics. In Chapter 5 the model, its results and its implications will be discussed, Chapter 5 is dedicated to discuss sub-questions 4 and 5. Finally, Chapter 6. will provide the conclusion of this work, its limitations and suggestions for future research. This is followed by a bibliography and an appendix.

Chapter 2. Theory & relevant work

This chapter outlines the theoretical framework and relevant work related to product and service improvements in customer service calls. In the first paragraph the concepts of service quality, customer satisfaction and emotions will be discussed. Thereafter, the theory of voice of the customer will be discussed. Paragraph 2.3 discusses the value of text data in marketing. Paragraph 2.4 gives an overview of aspect-based sentiment analysis.

2.1 Key concepts: customer emotions, quality and satisfaction

This paragraph explores the relationships between core concepts within service marketing such as emotional display, customer satisfaction and service quality and customer service calls.

2.1.1 Emotions in service encounters

Sentiment analysis makes it possible to extract some degree of emotionality from language (Balahur et al. 2012). At individual consumer level, sentiment and satisfaction are among the most common measurements in text data (Berger et al 2020). Emotions in service calls are interesting because they have a direct impact on customer behavior such as churning, repurchasing, word-of-mouth effect and complaints. Additionally, emotions measured in service encounters influence the level of satisfaction or dissatisfaction with the service (Zeelenberg et al., 2004). Other research found that customer's displayed emotions in service encounters are linked to their overall assessment of that company; it is strongly related to brand image (Mattila & Enz, 2002). Even in customer interactions with chatbots, emotions play a crucial role. Crolic et al. (2022) find that when customers interact with chatbots in a negative emotional state this will negatively influence customer satisfaction, overall firm evaluation and (future) purchase intentions.

Customer reactions to expectancy violations in service interactions are influenced by their emotional state, particularly anger (Ask & Landström, 2010; Crolic et al, 2022). Anger is an activating emotion which makes it possible to react quickly, to respond to obstacles or retaliate against offending parties. Angry customers are more likely to blame others when performance falls short of expectations (Crolic et al, 2022). Therefore, an angry customer may blame the firm and hurt its reputation by spreading bad word-of-mouth effects. Irrespective of modality, customer anger in customer service interactions is very prevalent. 20% of customer call center contacts involve customers in a negative emotional state, i.e. angry complaining customers (Grandey, Dickter & Sin, 2004). In fact, current research argues that it is likely that during the COVID-19 crisis the prevalence of anger in customer calls was even higher (Shanahan et al., 2020; Smith et al., 2021; Crolic et al., 2022).

Angry customers especially feel the need to divert to a desirable outcome (Roseman, 1984). Therefore, agents could have a mitigating role in service failures because of the existence of emotional contagion in service encounters (Du et al 2011). Emotional contagion means that in service failure and recovery encounters agents' negative emotional displays increase negative emotions for customers, on the contrary expressing positive emotions could decrease customers' negative emotions (Du et al., 2011). However, these interpersonal emotional exchanges are quite complex, it is not so simple that (always) expressing positive emotions would be the ultimate remedy. But there is a role of significance for the agent.

Like customer satisfaction, customer emotions are also associated with perceived service quality. Customer emotions towards a certain service attribute influence the effect this attribute has on perceived service quality. Besides it influences the relative importance of that attribute for perceived quality (Golder et al., 2012). Service providers who monitor consumer emotions during service delivery are more likely to adjust attributes of services and improve experienced attribute quality (Golder et al., 2012). Service quality will be discussed in more detail in section 2.1.3.

Research shows that in general, emotional responses in service encounters partially account for customer satisfaction levels (e.g. Smith & Bolton, 2002). In fact, it is unclear whether positive (negative) emotions can be distinguished (dis)satisfaction according to Bagozzi et al (1999). Satisfaction is neither an emotional category nor a basic emotion. However, it has much in common with positive emotions such as; happiness, joy, delight and enjoyment. In a principal component analysis by Nyer (1997) joy and satisfaction even loaded on the same factor. Therefore, although it is unlikely that there is a single emotional response for customer satisfaction, Bagozzi et al. (1999) argue that in practice it is very difficult to discriminate satisfaction from positive emotions. The difference between sentiment and emotions is further discussed in section 2.4.4.

2.1.2 Customer satisfaction and service calls

Due to the similarity of sentiment and emotions and strong link between emotions and satisfaction; sentiment can also be seen as a proxy for customer satisfaction (Bagozzi et al., 1999). Customer satisfaction relates to positive sentiment while customer dissatisfaction relates to negative sentiment.

Few studies discuss customer satisfaction in relation to user-generated text (Xu et al, 2017). Most studies use surveys to measure customer satisfaction or dissatisfaction. However, these have severe limitations. User-generated content (UGC) includes complaints and experiences, and can provide more accurate insight in customer perceptions because of their open structure (Xu et al., 2017). Customers are not restricted by survey design and express themselves more naturally and do not only discuss preselected items (Büschken & Allenby, 2016).

The theoretical foundation of the concept of customer (dis)satisfaction indirectly used in this research is the expectation-disconfirmation theory which states that before consumption of goods and services consumers compare their expectations with perceived quality of them (Oliver, 1980). Consumer satisfaction arises when the expectation is not greater than the perceived quality. Conversely, earlier work shows that during service encounters 'expectancy violations' harm customer satisfaction (Oliver & Swan, 1989). Expectancy violations occur when service fails to meet expectations and negative disconfirmation arises. Angry customers are more likely to suffer these expectancy violations (Crolic et al., 2022; Ask & Landström, 2010).

In theory, it is evident that customer satisfaction is important for firms to prosper. In fact, this connection has also been studied empirically. At the firm-level satisfaction and returns go hand in hand in the long run. In the short term however they may diverge (Anderson et al., 2014). Firms that achieve high customer satisfaction enjoy superior economic returns. For Sweden firms a one point increase in the customer satisfaction index leads to a $7.48 million increase in Net Present Value over five years for Sweden firms. Considering their net income this amounts to a cumulative 11.5% increase (Anderson et al., 1994). Economic returns from improved customer satisfaction are not immediately realized as they affect future purchasing behaviour and thus future cash flows. Likewise for *Fortune 500* companies a positive association between customer satisfaction and (telecommunications) shareholder value is found (Anderson et al., 2004). Moreover, a study by Gustafsson et al. (2005) found that customer satisfaction has a consistent negative effect on churn in the telecommunications industry. Besides, when satisfaction is measured as an overall evaluation of performance, it is a strong predictor for churn.

In conclusion, the service management literature converges on the fact that customer satisfaction is the result of a customer's perception of the received value, which is the perceived service quality relative to price. Also it argues that the first determinant of *overall customer satisfaction* is perceived quality (Cronin et al, 2000). Therefore, being able to measure customer satisfaction and dissatisfaction toward particular attributes of products and services is very useful for product and service improvements.[1]

2.1.3 Service quality

UGC, such as customer call transcriptions, offers firms major opportunities to receive customer feedback and improve attributes of products and services (Xu et al., 2017). In the satisfaction-profit chain, service quality is crucial to ensure profitable outcomes of service delivery. For most importantly, there is generally a strong positive association between service quality and customer satisfaction (Falk et al., 2010). Quality expectations are determined by experience and may change over time. There is some dynamic and non-linear relationship between quality and satisfaction (Falk et al., 2010). For example, functional attributes of product quality

[1] As explained later, in this research, satisfaction is not measured directly. But the sentiment measured in the customer calls can be seen as a proxy for satisfaction.

(availability, efficiency and privacy) lose their ability to delight customers as the relationship between the firm and customer ages. They become minimal requirements. On the contrary, hedonic quality attributes (design, image, layout, enjoyment) have an increasing effect on long-term customers (Falk et al., 2010).

Measuring service quality is thus a key factor in (e-)commerce success and driver of customer satisfaction. Historically, collecting customers' opinions has been difficult. However, the rise of UGC enables this (Palese & Usai, 2018). A well-known measurement for service quality is SERVQUAL (Parasuraman et al, 1988). It is based on five dimensions: reliability (ability to perform promised service accurately), responsiveness (willingness to help customers quickly), tangibles (physical facilities equipment and staff), assurance (knowledge and courtesy of staff) and empathy (company's attention to customers) (Palese & Usai, 2018). However SERVQUAL is difficult to leverage on text data. Instead of looking at latent dimensions it is more interesting to look at the perception minus expectation measurement for different characteristics of the product. Xu et al. (2017) e.g. examine customer dissatisfaction by analyzing customer perception in user-generated text. With sentiment analysis perceived service quality may be measured, via the indication of customer perception minus expectation.

There is strong empirical support that improving service quality leads to improved customer satisfaction and favorable behavioral intentions. However, improving service quality is only beneficial for profits if companies do so in a cost-effective manner. This makes the effect of service quality on profits complex (Zeithaml et al., 1996). Besides the link with profit, perceived service quality is related to emotional display in the service encounters (e.g. Ask & Landström, 2010; Oliver 1980). In perception emotions play a role. Customer contact employees are found to be able to influence customers' perception of service quality (Hartline & Ferrell, 1996).

2.2 Value and use of text data in marketing

80 to 95% of company data is unstructured, and mostly text data (Gandomi & Haider, 2015). The business value of automated textual analysis for marketing insights is therefore evident (Berger et al., 2020). The previous paragraph has shown that (emotionality in) service calls contain insights about customer satisfaction and perceived service quality.

Text signals something about its producer, e.g. customers in customer service calls. It might reveal insights about the individual in general as well as how they are feeling or may be thinking at the moment (states). Besides, it provides insight into the person's attitude towards other objects (Berger et al., 2020). Beyond single actors, text can also be aggregated to study larger segments of customers. More generally, text is shaped by its contexts (e.g. time period). Aggregated texts give insight in these contexts. Analyzing changes over time can provide relevant insights into aspects such as attitudes towards industries of products (Berger et al., 2020). Text can provide insights that are not easily obtainable through other methods, such as how customers feel about a brand, whether they like a new product or what attributes are relevant for decision making (Berger et al, 2020; Lee & Bradlow, 2011, Netzler et al, 2012).

Context affects text in three ways, through: (i) technical constraints and social norms, (ii) shared knowledge of the parties involved and (iii) prior history. Text from customer service calls are interactive, informal and contain short statements and responses relative to written reviews. Second, the relationship between the parties affects what is said; conversations in a customer service call may be driven by getting monetary compensation. Lastly, history may affect the text's content (Berger et al., 2020).

Text data can be used both for prediction as for understanding. Using text for understanding often involves examining only a small number of textual features or aspects that link to e.g. underlying psychological processes. For example: why do customers use certain types of emotional language when talking to customer service? Challenges for using text for understanding are drawing causal inferences from observational data and interpreting relationships with textual features (Berger et al., 2020).

2.2.1 Entity extraction

The most commonly used method for text analysis in academia and practice is entity extraction. Entity extraction extracts the meaning of one word or phrases in a text. Named-entity recognition (NER) is a form of entity extraction that involves extracting words and phrases that stand for the same referent. A named-entity can be thought of as a proper noun that serves as a name for something or someone (Nadeau & Sekine, 2007; Li et al., 2020). It allows generating a rich set of entities for predictive models or to explore more complex textual expressions such as sentiment or emotion. When no suitable predefined dictionary exists, a hand-crafted dictionary could be used to help define entities based on some rules. It is also possible to use supervised machine learning approaches to define entities; these methods require a hand-coded training data set (Berger et al., 2020).

Entity extraction has two limitations: the dimensionality of the problem (often thousands of unique entities can be extracted) and the interpretation of many entities. When these limitations are too severe, one can use topic modeling; identifying (latent) topics from a text body. However, this approach can be too general for the task at hand and interpretation of the topics can be difficult (Berger et al., 2020). Sometimes marketing researchers are more interested in textual relationships between products, attributes and sentiments, e.g. to identify whether consumers mention a particular problem with a specific product feature. Then identifying textual relationships between extracted entities is necessary: relation extraction. This may be captured by simple word-occurrence, but since this does not always imply a relationship, also with machine learning approaches (Berger et al., 2020).

In conclusion, consumer-firm interactions are a rich area to examine. According to Berger et al. (2020), a promising direction for future research is to use call center data to better understand interpersonal communication and record what drives customer satisfaction, to understand conversational dynamics between customers and agents, to use textual features to predict outcomes such as churn, or to use calls to evaluate the customer-centric strategy by assessing service quality, style or impact on sales.

2.3 Voice of the Customer

A marketing research method related to this research is Voice of the Customer (VOC). The goal of VOC is to extract and organize customer needs in a structured way. The main benefit of capturing and analyzing the voice of the customer is being able to provide input for the product development process. Voice of the customer has shown to increase company competitiveness and avoid product changes that do not satisfy customer expectations (Aguwa et al., 2012). As VOC takes customer satisfaction as a performance indicator, it helps companies reinforce customer orientation. In quality function deployment (QFD) customer wants and needs are linked to design attributes to encourage the combined consideration of marketing and engineering (Griffin & Hauser, 1993). The application of QFD leads to long term benefits such as customer satisfaction, reduced cost and reduced time in 83% of cases. (Griffin, 1992).

Technology management research argues that cooperation and communication between marketing, manufacturing, engineering and R&D leads to more profitable products and more successful new products. Quality function deployment (QFD) addresses the communication problem by ensuring engineering and R&D decisions are driven by VOC. Voice of the Customer (VOC) is a theory that relies on customer needs. Customer needs are descriptions of what customers desire the product or service to satisfy. For a product typically between 200 and 400 needs are identified. These can be categorized in: basic needs, articulated needs and exciting needs. Basic needs are what customers assume the product (feature) does. Articulated needs are what customers want the product (feature) to do. Exciting needs would delight customers if fulfilled (Griffin & Hauser, 1993).

QFD structures the needs hierarchically into strategic, tactical and operational needs. Strategic needs contain five to ten top-level needs used by the product-development team to set the strategic direction for the product. Each primary need (strategic need) can be broken down into three to ten secondary (tactical) needs which provide more detailed information about what the team must do to fulfill the primary need. The tactical need might e.g. specify how the customer judges the primary need. Operational needs (tertiary needs) provide detail so that the team can develop solutions to satisfy the secondary needs. Ultimately the team should prioritize and make those strategic decisions where the cost of satisfying the need balances with the consumer desirability (Griffin & Hauser, 1993). Importance is thus driven by perceived customer needs. Customer needs are in turn partly driven by customers' perception of performance which is based on comparison with competing products in the market. VOC can differ per customer segment, for example in importance of needs. The company should identify measurable aspects (design attributes) of the product or service which influence customer' needs if changed, and ideally they even know what magnitude that influence has (Griffin & Hauser, 1993).

2.3.1 Identifying user needs from text

Many firms rely on voice of the customer analysis to identify opportunities for product development or to select attributes for conjoint analysis to improve service quality. Traditionally, needs are identified through customer interviews and focus groups (Griffin & Hauser, 1995). However, these methods are expensive and time-consuming. Moreover, relying on heuristics (managerial judgement) is susceptible to bias and misses customer needs not fulfilled in the current market. Timoshenko and Hauser (2019) propose a hybrid human-machine method to identify user needs with machine learning and unstructured text data. User-generated content (UCG) such as reviews and social media, is used to source customer needs. A convolutional neural network is used to find sentences that contain informative content. Sentences with similar embeddings are clustered and from each cluster a sentence is sampled for human review and customer need formulation. Customer needs are extracted manually because they need to be embedded in context and this is not yet possible with machine learning methods.

Sourcing customer needs from UGC is at least as competitive compared to traditional interviews. A big plus is that UGC is updated continuously. Additionally, the machine learning approach of VOC analysis brings large time-to-market savings for product/service innovation. Another technique (instead of clustering) to identify customer needs is opinion mining, such as aspect-based sentiment analysis (Bigorra et al., 2020). Bigorra et al. (2020) improve this process by proposing an automated way to validate or update known-customer needs and add new customer needs to their existing database.

2.4 Aspect-based Sentiment Analysis

Sentiment analysis is a widely adopted text mining technique to analyze user generated text. Sentiment analysis or opinion mining, is the computational study of people's opinions, attitudes and emotions towards entities (Pang & Lee, 2008). These entities can represent all sorts of things, for example brands, products or people (Medhat et al., 2014). Sentiment and opinions are related, but slightly distinct. Opinions indicate an individual's view on something. Sentiment indicates an individual's feeling on something. It is thus more related to emotions (see section 2.4.5). In sentiment analysis, a sentiment polarity label is predicted for text (chunks). Sentiment polarity specifies the direction of the sentiment, typically positive, neutral or negative (Brauwers & Frasincar, 2021). Positive sentiments include delight, joy, satisfaction while negative sentiments include anger, fear, guilt, sadness, dissatisfaction and frustration (Balahur et al, 2012). Sentiment analysis detects the implicit expressions of customer's emotions in texts (Xu et al, 2017; Balahur et al, 2012). Sentiment analysis can therefore reveal drivers of consumer satisfaction or dissatisfaction. It is not a direct measure but a proxy for satisfaction.

The granularity of sentiment analysis differs in the (range of) sentiment labels (or scores) assigned, the text-level (documents, paragraphs, sentences) and target it is applied to (Brauwers & Frasincar, 2021).[2] In recent years due to advances in NLP, sentiment analysis has become more fine-grained. Typical sentiment analysis used in marketing studies measures overall sentence sentiment based on word counts techniques of polarity carrying words listed in sentiment dictionaries.

One recently emerged fine-grained type of sentiment analysis is Aspect-level Sentiment Analysis (ABSA). The goal of Aspect-based Sentiment Analysis is to extract aspects and extract their associated sentiment in text documents, typically sentences. ABSA is thus not concerned with overall sentiment, but in aspect-level aspect. ABSA consists of two tasks: aspect-term extraction (ATE) and aspect-sentiment classification (ASC). Current methods differ in which of these tasks (or both) they provide and with which type of algorithm. A third task is sometimes distinguished: the aggregation of aspect-sentiment pairs to be able to present and give an overview of the results (Schouten & Frasincar, 2015). Models that fulfill all tasks are referred to as end-to-end ABSA.

With aspect-sentiment the emotional tendency in text can be extracted better, giving more accurate reference for decision makers (Yang et al., 2021). End-to-end aspect-based sentiment analysis can therefore be used to analyze customer emotions and needs in inbound customer contact calls by extracting different aspects of KPN products or service and their associated sentiment. Consumers express all kinds of customer needs in these calls. Aspects are attributes or components of a product or service. Next two sections will discuss the ATE and ASC subsequently.

2.4.1 Aspect Term Extraction

The aspect term extraction task is a sequence tagging subtask, comparable with Named Entity Recognition. In most current approaches ATE is studied independently, separate from the ASC task (Yang et al, 2021). NER is discussed as well in section 2.2.1. The goal of aspect term extraction is detecting an aspect (KPN product/service feature) in a sentence. Often aspects are distinguished from entities, however sometimes the terms are mixed up. An entity is generally seen as an overall topic or category within the text, while aspects are seen as specific words or phrases in sentences. Sometimes the aspect is regarded as a property of an entity. This research will make use of aspects as nouns that serve as a name for product or service features.

[2] Hence, ABSA has multiple related methodologies: e.g. Entity-based Sentiment Analysis or Target-based Sentiment Analysis.

Main methods for aspect extraction include: frequency-based, syntax-based, supervised machine learning, unsupervised machine learning and hybrid methods (Jin Ding et al, 2018; Frasincar & Schouten, 2015). Supervised machine learning methods arrange the aspect extraction as a labeling problem. Some approaches use a predetermined list of aspects, while others freely discover aspects from the text (Schouten & Frasincar, 2015). For marketing oriented studies Aguwa et al. (2012) propose to use word frequency lists to find keywords in customer comments, these can be manually cleaned by domain experts.

2.4.2 Aspect-Sentiment Classification

The meaning of sentiment is explained at the beginning of paragraph 2.4. Aspect sentiment classification is the task of classifying sentiment polarity associated with specific extracted aspects-terms in a sentence. Sentiment polarity is typically classified according to a set of possible values: positive, neutral and negative. Sentiment classification can be conducted at three text-levels: document-level, sentence-level and aspect-level. Classic lexicon-based sentence sentiment classification uses human rules to generate a 'sentiment dictionary'. The overall sentence sentiment is scored by adding and subtracting all polarity words' sentiment scores. This method is weak in taking linguistic context into account. Often consumers use certain linguistic patterns to express sentiment (Villaroel Ordenes et al., 2017). Moreover, it cannot address sentences with multiple aspects with varying sentiment. And lastly, it's hard to generate a domain specific dictionary. For classification at aspect level {aspect, sentiment} tuples are extracted. This allows for a more clean analysis and can address the weaknesses explained above. It often requires supervised machine learning methods that use many parameters from the data such as natural language processing, labeled text and various classification algorithms to classify sentiment. Sometimes sentiment dictionaries are used as input as well (Jin Ding et al., 2018; Schouten & Frasincar, 2015).

2.4.3 State-of-the-art models

With the rapid development in NLP, attention-based deep learning models recently scored state-of-the-art results for ABSA. These attention-based models include transformers, which dominate the ABSA task and other language processing tasks in terms of predictive performance. This research makes use of attention based models, whose architecture will be explained in the next chapter. While it is common to have a model dedicated to either aspect or sentiment detection, these problems are not independent, so methods that extract both at the same time are more valuable as sentiment information can be used to extract aspects and vice versa (Schouten & Frasincar, 2015; Brauwers & Frasincar, 2021).

2.4.4 Emotion or Sentiment analysis

Brauwers & Frasincar (2021) discuss emotion analysis as a direction the field (of sentiment analysis) can take. Sentiment analysis and emotion analysis are highly related; there is no clear distinction. In sentiment analysis typically peoples feelings towards something are extracted by looking at opinion words. Opinion words indicate someone's view on a specific matter. In sentiment analysis the task is to extract sentiment polarity from text. Polarity refers to the orientation (positive, neutral or negative) of these feelings, i.e. sentiment. Instead of only extracting polarity labels from texts, also a wider range of emotions can be considered; 'anger', 'joy', 'fear' et cetera. This is what would be *emotion analysis* (Brauwers & Frasincar, 2021). Emotion analysis can like sentiment analysis be done at multiple levels of depth. And emotion analysis and sentiment analysis can also be performed jointly. Some papers have done emotion analysis, but this type of analysis is very novel and thus rare.[3] This research does not use a specific-emotion approach. That could in theory deliver more insights, because different specific emotions have different behavioral tendencies (Zeelenberg et al 2004). However, implementing specific emotions in aspect-based sentiment analysis is more time expensive and the call transcriptions are likely not rich enough to support this more complex type of emotion classification. Since for this research the goal is to assign positive, neutral or negative sentiment (polarity) labels towards certain words in the text, it will be referred to as aspect-based *sentiment* analysis.

[3] Topal & Ozsoyoglu, 2016; Suciati & Budi, 2020; Wang et al., 2020.

Chapter 3. Methodology

This chapter provides an overview of the methodology used for aspect-based sentiment analysis in this work. In comparison to the previous chapter, in this chapter aspect-based sentiment analysis will be discussed from a more technical angle. First of all, the BERT architecture will be discussed. Thereafter, the methodology to design a list of business-relevant aspects will be discussed, this is an element of ABSA rarely discussed in other papers. In paragraph 3.3 the neural pipeline designed for this research will be discussed. Last of all, a discussion of the methodology will conclude this chapter.

As discussed in Chapter 2, aspect-based sentiment analysis (ABSA) is a fine-grained task of natural language processing (NLP). With aspect-polarity the emotional tendency in text can be extracted better, giving more accurate reference for decision makers (Yang et al., 2021). ABSA consists of two subtasks: aspect term extraction (ATE) and aspect sentiment classification (ASC) or aspect polarity classification (APC). The purpose of ASC is to predict the exact sentiment of a certain aspect, rather than take the overall sentiment polarity on sentence level or document level. ASC is a classification task. Sentiment polarity is usually classified in threefold: positive, neutral and negative. The purpose of ATE is to extract aspect terms from the text corpus. ATE is sequence tagging subtask, comparable with NER. In most current approaches ATE is studied independently, separate from the ASC task (Yang et al, 2021).

3.1 Bidirectional Encoder Representations from Transformers

BERT models are very capable of performing both ATE and ASC. BERT stands for Bidirectional Encoder Representations from Transformers. In computational linguistics the leading trend is to not build models from scratch, but rather fine-tune for specific tasks on large pre-trained general purpose language models. BERT is the most commonly used pre-trained language model, developed at Google and open-sourced late 2018 (Devlin et al., 2019). Google pre-trained their BERT on a huge text corpus, which enables BERT to understand the (mathematical) fundamentals of language. BERT and BERT-derived models are based on the transformer architecture developed at Google in 2017 (Vaswani et al., 2017). The introduction of BERT is seen as a breakthrough in NLP research; when BERT was introduced it received state-of-the-art results on eleven NLP tasks (Devlin et al 2019). Nowadays many state-of-the-art benchmark results on NLP tasks are still held by (fine-tuned versions of) BERT-family models. This is also true for aspect-based sentiment analysis; for the SemEval challenge tasks that concern ABSA, BERT-family models are consistently ranked among the top scores.[4]

3.1.1 Transformer architecture

To understand the BERT architecture at first its foundational concept, Transformer models, is outlined. Transformers stem from Vaswani et al.'s famous paper 'Attention is all you need'

[4] https://paperswithcode.com/sota/aspect-based-sentiment-analysis-on-semeval

(2017). The transformer architecture is depicted in Figure 3.1. All elements in this overview are intuitively explained below. For a technically deeper level of understanding for these layers, see Vaswani et al. (2017).

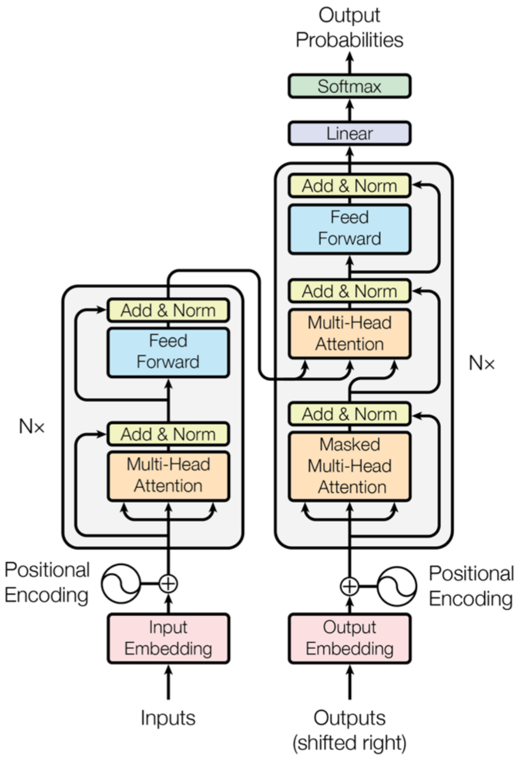


Figure 3.1: High-level overview of the Transformer Architecture, adapted from Vaswani et al. 2017

Transformer models have an encoder-decoder architecture with an encoding block and decoding block. These blocks both consist of multiple stacked encoders respectively decoders. Thinking in terms of a translation application from Dutch to English, the encoder block can be seen as the component that takes a Dutch input sentence and converts it to ‘computer language’: vectors and matrices. The decoder block can then be seen as the component that turns these matrices to an English output sentence. The encoder component will be discussed next. The decoder component is not discussed as BERT only consists of stacked encoders.

All the stacked encoders have an identical structure (but all with different weights). Each of these encoders have two sub-layers: a multi-headed self-attention mechanism and a feed forward neural network. Each encoder takes a list of vectors as input. The bottom encoder takes the word embeddings from the input sequence as input. The self-attention layer processes this list and feeds it to the feed forward network which sends its output to the next encoder. This self-attention mechanism is a key concept for transformers. Self-attention allows the transformer to look at other tokens (words and punctuation) of the input while processing a word. This makes it possible to merge the understanding of other relevant words in the sentence into the word currently being processed. Humans do this as well all the time. For example in the sentence; “A Transformer model, how can I understand that?” computers should learn that ‘that’ refers to ‘Transformer’ and ‘model’. Besides self-attention, another major innovation of Transformers is that it uses positional encodings and therefore does not require sequential input as previous NLP models did. Therefore, some computations lend for parallelization, which strongly increases the model’s speed and scalability (Vaswani et al., 2017). Another strong improvement is that previous NLP models were incapable of remembering longer sentences, often they would forget the first part of the sentence while they were still busy processing the input sentence. Self-attention improves this mechanism strongly because it does not use a single so-called context vector which other models did (Vaswani et al 2017; Affane, 2020a; Weng, 2018; Affane, 2020b; Allamar, 2018, Uszkoreit, 2017).

To summarize, the multi-headed attention layer calculates what parts of the input it should focus on and how relevant each word is to other words; expressed in attention vectors. The attention vectors thus capture contextual relationships between words in a sentence. Multiple attention vectors per word are used, that is why the mechanism is called multi-headed. The final attention vector is the weighted average of the (multi) attention vectors per word, if the attention mechanism is not multi-headed it may weigh the relationship with the word itself too high. The feed-forward net transforms the attention vector to a set of encoded vectors that is processable for the next encoder/decoder block. All words in a sentence are passed into the next block at the same time (Vaswani et al 2017; Allamar 2018, Uszkoreit, 2017).

3.1.2. BERT: Pre-training and fine-tuning

BERT’s model architecture is based on the multi-layer bidirectional transformer encoder; it uses the same stacked encoder layers as transformers, but discards the decoders. BERT uses bidirectional pre-training for language representations and is pre-trained on two core tasks: *masked language modeling* to be able to understand relationships between words and *next sentence prediction* to understand the relationship between sentences. Compared to previous models BERT, exploits deeply bidirectional *unsupervised* language training. While it is pre-trained on plain (unlabeled) text only and the resulting pre-trained representations are both *contextual* and *bidirectional*. Contextual means that the generated representations for every word are based on other words in the sentences; the word ‘bank’ has a different representation in the context of ‘bank account’, than in the context of ‘river bank’. Bidirectional means that BERT can represent the word ‘bank’ on its foregoing as well as next context. This ensures that “I accessed the bank account” and “I accessed the bank of the river” have different representations for the word ‘bank’ (Devlin & Chang, 2019). Devlin et al. (2019) introduced BERT with 12 stacked encoders, 110M parameters and pretrained on 3.3B words.[5] In pre-training it learned what language and linguistic context is, this general ‘understanding’ of language is essential for downstream tasks (Devlin & Chang, 2019; Alammar 2018). Fine-

[5] Hidden-size*: 768 & self-attention heads: 12.
*The hidden size is the number of dimensions of the output vectors

tuning can now be relatively inexpensive compared to pretraining. To solve a downstream task BERT should be fine-tuned to learn this specific task. Only the output parameters are learned from scratch, the model parameters are only slightly fine-tuned (Devlin et al., 2019; Devlin & Chang, 2019). So-called *downstream* tasks could be: machine translation, question answering, sentiment analysis and text summarization. They are 'downstream' because they are not the initial NLP tasks BERT was pretrained on.

Late 2018 multilingual BERT (mBERT) was introduced, which has been trained on the entire wikipedia text for the top 100 Wikipedia languages (Devlin, 2018). Dutch is well represented, being the sixth largest language on Wikipedia.[6] Pires et al. (2019) show that a fine-tuned mBERT performs very well on the CoNLL benchmark for Named Entity Recognition for Dutch language. mBERT is almost as good as on English data, and better compared to German or Spanish data.[7] When mBERT is not fine-tuned on Dutch data, performance drops significantly but is still moderately well.[8] On the Dutch Book Reviews (DBRD) benchmark for Sentiment Analysis mBERT scores 89% accuracy (De Vries et al, 2019). In conclusion, mBERT has a good understanding of (Dutch) language, especially when fine-tuned.

De Vries et al. (2019) introduced the first monolingual Dutch model: BERTje. BERTje is trained on 12Gb of Dutch multi-genre texts (news, wikipedia, reviews) equivalent to 2.4B tokens. Pre-training procedure is slightly altered; next sentence prediction is replaced by sentence order prediction and for masked language modeling (MLM) tokens are not masked entirely random to prevent *too easy* to predict (suffixes of) words. BERTje consistently outperforms equally-sized mBERT's (De Vries et al, 2019). However model size matters too; a larger sized fine-tuned mBERT is able to outperform BERTje.[9] Both on the CoNLL-2002 NER as DBRD Sentiment analysis benchmarks BERTje achieves significantly higher scores than mBERT (De Vries et al. 2019).[10]

The newest Dutch BERT model is RobBERT v2, introduced by Delobelle et al. (2020). It is trained on a larger dutch text corpus, that is 39Gb large with 6.6B words using the RoBERTa training regime (Liu et al., 2019). The RoBERTa pre-training procedure only consists of the Masked Language Modeling task. Besides, the masking pattern is not static, but changes per epoch (Liu et al., 2019). RobBERT v2 outperforms BERTje slightly on the CoNLL NER and DBRD Sentiment Analysis benchmarks (Delobelle et al., 2020).[11]

[6] See: https://meta.wikimedia.org/wiki/List_of_Wikipedias#1_000_000+_articles

[7] CoNLL NER F1 score Dutch: 89.86; English: 90.70; Spanish: 87.18; German: 82.00%.

[8] English finetuned mBERT: F1 score on Dutch: 77.36%; German: 69.74%, Spanish: 73.59%.

[9] See e.g. CoNLL-2002 NER.

[10] 88.3% (CoNLL F1) and 93% (DBRD Acc.) for BERTje and 80.7% and 89.1% for mBERT.

[11] 89.1% (CoNLL F1) and 95.1% (DBRD Acc.) for RobBERTv2 and 88.3% and 93.0% for BERTje.

Since all three models discussed above perform good on Dutch language for NER and (Sentence-based) Sentiment analysis, mBERT, BERTje and RobBERTv2 will be considered as pretrained models in developing a model for Aspect-based Sentiment Analysis. The next paragraph will set forth how a relevant list of aspects can be gathered for the model to learn.

3.2. Exploration and choice of business-relevant aspects

This paragraph answers the sub research question II: “*how to find relevant aspects for aspect-based sentiment analysis?*”. In other research, this element is rarely discussed. However, generating managerial insights requires the model to find sentiment for business-relevant aspects. Therefore a list of relevant aspects must be made. This is not a predefined set of aspects the model is restricted to classify aspects from, but a list used to label aspects and create training data. It should be noted that this task is more art than science and requires quite some manual labor. Therefore, this task is done in a structured manner to ensure repeatability of this research. Aspects are certain words or phrases that carry valuable information. When a word can be considered an aspect depends on the purpose of ABSA. Given the goal of finding product and service improvements, in this case aspects are mostly product or service characteristics. Aspects are chosen using three approaches via: *i.* (domain) expert knowledge, *ii.* a customer journey approach and *iii.* data visualization. The list of aspects can be found in Table 4.5.

The final list of aspects should satisfy the following three conditions. First of all, the list of aspects has to be business-relevant in terms of that it has to be oriented towards a product/service. This is ensured by proper discussion with domain stakeholders. Second, the aspects should be ‘customer oriented’, i.e. mentioned frequently. For example a technical aspect of a modem might be relevant product wise, but if it is not represented in the data, it is not useful in aspect-based sentiment analysis (ABSA). So, ‘customer oriented’ means that product/service aspects are wordings that customers use in service calls. This customer centricity is desirable as it ensures that the aspects represent customer needs. The third requirement entails that the aspect list should be restricted in terms of size. One reason for this has to do with time and resources constraints. The longer the list of aspects, the more labeled sentences are needed to perform ABSA. Another reason is that a list with too many aspects threatens business value as it decreases usability for business (end) users further down the line. For this thesis, no more than 50 aspects are chosen in the final list.

Six service domains within KPN can be distinguished: Mobile, Landline Phone, Internet, TV, Order and (customer) Administration.[12] ABSA can be performed on each of these domains, however each of them requires a different list of aspects. Therefore, to scope the research a specific service domain is chosen: mobile internet (KPN Mobiel). The service domain is chosen based on data availability and the first stage of the exploration of (potential) aspects. To filter

[12] As distinguished by a business team, these exact categories are used as product categories.

for customer calls for the KPN Mobiel domain, the data is filtered on their predicted text 'service topic category'.[13] These service topic categories overlap with the identified service domains.

3.2.1 Approach to find aspects

A three-stage aspect exploration strategy is used to work towards a final list of relevant aspects. First of all, aspects are inferred from a (KPN) domain expert lexicon. This list contains over 700 words from customer service calls that are potentially interesting for customer contact analytics. However not all aspects are relevant in the context of customer needs and product analysis. Together with a second reader, relevant aspects are discussed, picked and categorized for each service domain. Categorizing helped structure the process of finding aspects. This way for each service domain a list of around 60 aspects is created. The choice for the Mobile domain is partially based on these lists. For the *mobile* a list of 51 separate aspects (seed words) is found. For each aspect *(seed word*) multiple synonyms or (very) related words can be found.

The expert-based approach is a great way to create an initial list, however it is susceptible to managerial or expert biases and heuristics. Therefore, the second stage to find aspects is by taking a customer point of view. The customer journey before a service call begins at 'first line customer help'. KPN provides first line customer help via their 'MIJN KPN' app or its website.[14] The help interface of the MIJN KPN app is divided into the six different service paths: Mobile, Landline Phone, Internet, TV, Order and Administration. By scanning through the MIJN KPN app content for KPN Mobiel as if a customer, relevant aspects are found. After all, these pages are dedicated to help the customer solve its needs and provide information. Whenever customers face a problem or need, they will search for information about the issue at hand, trying to solve the problem themselves. They will 'travel' through the MIJN KPN app or website and will probably pick up certain terms and wordings in that content. So based on customer journey through the MIJN KPN app, a second KPN Mobiel list of aspects is created. This list is used to find extra aspects and confirm the aspects found already in the previous aspect exploration stage. Merging the two lists from the domain expert-lexicon and customer journey stage respectively results in a list of 76 aspects, i.e. with the customer journey stage 25 extra relevant aspects to consider for the final list are found.

Finally, a data-driven approach is conducted in order to find relevant aspects and to discard aspects that (almost) do not appear in the data. First, a list is created with the most frequent words from a random sample of 40,000 customer turns from service calls.[15] In order to refrain 'stop words' from dominating the list, stop words are removed using a dictionary with Dutch stop words, along with some manual additions.[16] This way a frequency list of 250 words long

[13] These service topics are predictions from a text classification model developed by KPN.
[14] Content on the app and website are uniform.
[15] A sample is desirable as searching terms in 2.3 million rows is computationally heavy.
[16] Stop words are e.g. *ik, het, de* et cetera.

is generated. All words that occur in this frequency list are considered to add to the aspect list, based on their business relevancy. Second, the word frequency is used to exclude aspects that occur infrequent in the data. Every aspect that occurs in less than one in four hundred sentences (0.25%) is excluded.[17] Because the sample of sentences were on all KPN service domains, a fairly low threshold is taken into account. Aspects that are discarded are e.g. 'uploadsnelheid', 'roaming' or 'simkaartnummer'. Eventually the lists resulting from the three approaches are merged in order to derive a final list of aspects. The final list of aspects is 44 words long and can be consulted in Table 4.5. Synonyms for words in the list are not included in the list.[18] However, diminutives or plural forms for aspects will be labeled and are hence indirectly included.[19]

3.3 End-to-end neural model pipeline

The 'full neural pipeline' model to conduct aspect-based sentiment analysis in this work is based on research by Zeng et al (2019) and Yang et al (2021). Yang (2022) maintains a repository in GitHub in which this research is bundled.[20] This paragraph outlines the used model. The full end-to-end (data and coding) pipeline from data collection to visualization of predictions is shown in Appendix 1.

The pretrained models (tested) in the model pipeline are mBERT, BERTje, KPN BERTje and RobBERT v2, as discussed in section 3.1.2. These pretrained models function as the base models with general (machine) understanding of Dutch language. Yang et al. (2021) designed additional layers on top of the pretrained BERT architecture for the downstream task of aspect-based sentiment analysis. The model architecture is shown in Appendix 2.[21] Yang et al. (2021) named this architecture *LCF-ATESC*, which stands for a *Multi-task model for Local Context Focus Aspect Term Extraction and Aspect Sentiment Classification.*

The software of Yang (2022) is used to train (fine-tune) models on (labeled) KPN domain data and to subsequently make predictions with these fine-tuned models. It is possible as well to use 'checkpoints' (previously trained models) to make predictions. However, the multilingual checkpoint, provided by Yang (2022) is fine-tuned on very few Dutch data on restaurant reviews.[22] More importantly, the KPN customer call data (*target domain)* has a very different distribution than the written review data, and the aspects to be found are very domain specific. Hence, the educated guess is that a domain unspecific fine-tuned model will have significantly

[17] For this the occurrence of the seed words and its synonyms are taken into account.
[18] The exception is 'factuur' and 'facturen', because of an experiment for KPN.
[19] For example: *e-mailtje* for e-mail or *kortingen* for korting.
[20] See: https://github.com/yangheng95/PyABSA
[21] PyABSA v1.16.27 is used, see https://github.com/yangheng95/PyABSA/tree/release. This repo is actively maintained, and newer versions may arise over time.
[22] Only 1 of the 21 datasets is Dutch, <5% of the fine-tuning data.

lower performance than a domain fine-tuned model. Although this is not tested in this specific use case, this follows from previous research as well (Yang et al 2021).[23]

3.3.1 Multi-task Aspect-Term Extraction and Aspect-Sentiment Classification

Many previous models extract aspect-sentiment, given the aspect; they are trained to only extract aspect-sentiment. The model framework used for this research does *both* aspect target extraction (ATE) and aspect-sentiment classification (ASC), sometimes called end-to-end ABSA. Together this joint learning task is called ATESC (Aspect Target Extraction & Sentiment Classification). The benefit of multitask learning is that the model can use information extracted during aspect extraction to predict the sentiments (Brauwers & Frasincar, 2021; Schouten & Frasincar, 2015).[24] The LCF-ATESC model from Yang et al. (2021) has state-of-the-art performance on the ATE and ASC task the SemEval-2014 (task 4) English Restaurant and Laptop datasets and four Chinese review datasets.

The network architecture for LCF-ATESC-BERT is depicted in Appendix 2.[25] The network consists of two units: a local context generator and a global context generator. Two independent BERTs are used. One pretrained BERT layer models for local context and the other for global context. Global context refers to the ‘information extracted’ from all tokens in the sentence. In other words, all source words are used to calculate (global) attention (vectors). Local attention is calculated by using a subset of the source words. Local context in this model refers to ‘information extracted’ from all words nearby the aspect term’s position in the sentence (Yang et al., 2021). Local context will be further illustrated in the next section.

The aspect extractor identifies aspect terms based on the global context features only. It is not only predicted whether a term is an aspect but also which aspect it is. The aspect-sentiment polarity extractor identifies associated aspect-sentiment for every aspect based on both the global context features and the local context features. These are combined by the feature interactive layer that concatenates the features. During the model fine-tuning on labeled data, each independent BERT (layer) is fine-tuned independently according to the joint loss function (Yang et al., 2021). The joint loss function is the sum of the ATE loss and the ASC loss (Yang, 2022).

Instead of using two independent BERTs for the *fast*-lcf-atesc models one BERT model is used to generate both global and local context features (it remains a multi-task network). This speeds up fine-tuning without significant loss in performance (Yang, 2022).

[23] This is referred to as zero-shot performance, which means that the model is tested on data it has never learned in training, i.e. the source domain and target domain differ.

[24] The SemEval 2014 Task 4 datasets are (amongst others) benchmark datasets for ABSA.

[25] Full (technical) details can be found in Yang et al. (2021) & Zeng et al. (2019).

3.3.2 Local Context Focus

Yang et al (2021) deploy a 'local context focus' (LCF) mechanism in their model. The global and local approach to attention was proposed and set out by Huong et al. (2015). It was found that local attention calculated by using a window of words around the target(s), enabled significant improvement in neural translation tasks (Weng, 2018; Luong et al., 2015; Huong et al., 2015). The local context focus mechanism for aspect-based sentiment analysis was successfully introduced by Zeng et al. (2019). Their particular LCF-BERT model gained state-of-the-art performance on three aspect-sentiment classification benchmarks.[26] LCF design assumes that local context of aspects contain more significant information apart from global context. LCF is thought to be effective because of the characteristics of natural language. The context word used to judge an aspect is likely to be near the aspect to be commented on (Zeng et al., 2019). Appendix 3 depicts the local context focus architecture.

To determine which context word belongs to the local context of an aspect, Zeng et al (2019) use the concept of Semantic-Relative Distance (SRD).[27] SRD is used to determine whether a context word belongs to the local context. The idea of LCF is to preserve the information of context words within a certain semantic relative distance range. The formula for SRD is stated below.

$$SRD_i = |i - P_a| - \left\lfloor \frac{m}{2} \right\rfloor \tag{3.1}$$

Where, SRD$_i$ refers to the semantic relative distance of contextual word towards the aspect.
i refers to the position of the contextual word token,
P$_a$ refers to the position of the centralized aspect term and
m refers to sequence length.

All context words within a certain SRD threshold (a hyperparameter) are regarded as local context. Zeng et al. (2019) designed three mechanisms to pay more attention to words in the local context of the aspect: Context Features Dynamic Mask (*CDM)*, Context Features Dynamic Weighted (*CDW)* and a *Fusion*. The CDM layer masks out representations of contextual words outside the SRD threshold. CDW weakens the features of less-semantic-relative contextual words. Fusion does both and takes the average of the two methods.[28]

3.4 Discussion of methodology

This section discusses performance, interpretability and (labour) expenses for the used methodology both for the manual compiled list of aspects as for the neural model.

The list of gathered aspects can be thought of as a sort of knowledge base. Knowledge bases are defined as stores of information with underlying sets of rules, relations and assumptions

[26] Improvements of 2-4 percent points on the accuracy and F1 scores.
[27] A certain score for the distance between the context word and aspect's position.
[28] For the full (technical) details of LCF, see Zeng et al. (2019).

that computers can draw upon (Brauwers & Frasincar, 2021). The advantage of knowledge-based methods is interpretability; it is easy to identify the information used that produces models output. The downside is that construction of knowledge bases can take considerable time and that they should be updated frequently. While the type of neural model should be able to discover more interesting aspects on its own, its aspect recognition ability is likely to be strongly dependent on the training procedure, and more particularly the labeling of aspects. If new products are launched most likely extra training data is required in order to consistently find these terms in the call text. The interpretability of the aspect list is good as long as it is clear to stakeholders what its purpose and (entry) requirements are. This holds partially for the aspect term extraction task as this depends strongly on the aspect list.

The strong advantage of the neural ATESC-BERT model is that it is capable of providing fine-grained sentiment information per aspect. It does not relate found aspects to sentence sentiment, but can really find (local) associated sentiment. The aspect-sentiment classification (ASC) task is however believed to be a complex task because sentiment is often product dependent. A battery that lasts long is positive, however a malfunction that lasts long is negative. People express themselves differently depending on the context. Next to that, sarcasm or idioms remain hard to understand. The aspect-sentiment classification task might be enhanced by exploiting a knowledge base as well (Brauwers & Frasincar, 2021). For example a list of specific words per aspect that carry a certain sentiment could be created and feeded as extra features to the model.[29] This might improve interpretability for ASC as well.

The major downside is that this neural model for ABSA is a supervised method, where a significant number of labeled sentences are required to train the model (as holds for most existing studies). Such labels, however, are expensive and difficult to obtain, especially for a new domain or new language (Brauwers & Frasincar, 2021).[30]

In practice the majority of (academic) aspect-based sentiment analysis studies use the same (labeled) benchmark data (e.g. SemEval data: Pontiki et al. 2014; Pontiki et al. 2015; Pontiki et al. 2016). In business context with unseen data from another distribution, the resulting models from those papers regardless of their technical ingenuity are not useful, unless the company produces a fine-grained labeled dataset to train the model. Domain specific fine-tuning is believed to be very important (Chen & Wan, 2022; Yang et al., 2021; Ramponi & Plank, 2020). And previously trained ABSA models cannot be re-used.

The labeled data problem has recently been addressed by unsupervised domain adaptation, where cross-domain performance of a model can be optimized by transferring knowledge from the labeled source domain (e.g. Dutch restaurant reviews) to the target domain (e.g. KPN

[29] In a sense, the aspect extraction task is also steered with the relevant aspect list to be labeled.

[30] For each sentence every target aspect must be labeled together with its sentiment polarity.

Customer calls) (Ramponi & Plank, 2020). However, since the target domain data is unlabeled, the problem still arises that many aspect terms are strongly related to the specific domain. For cross-domain models it is a huge challenge to find business-relevant aspects. Apart from that, domain transfer learning methods have too many complex modules and require expensive multistage pre-processing (Chen & Wan, 2022). In conclusion, domain adaptation is not (yet) a robust method to avoid labeling of data to extract business-relevant results.

A general downside is that the ATESC BERT is a huge model, pretrained on huge amounts of data, which needs to be fine-tuned on a small dataset with a few hundred to a thousand examples to be able to perform ABSA. Some researchers believe this can hurt model reliability or can hurt performance (Rogers et al., 2020). This is a general critique on using BERT models for downstream NLP tasks.

In conclusion, the proposed method has potential to deliver good performance, but requires time and skill in setting up the model. Because an aspect list has to be created, data has to be labeled and the model needs to be fine-tuned on the domain.

Chapter 4. Data

Methodology in machine learning research is often strongly dependent on data preparation. The data chapter is on purpose separated from the methodology chapter. Data wrangling, researching and finding the aspects as well as applying these to real data, i.e. the labeling thereoff, play a decisive role in aspect based sentiment analysis.

4.1 Type of data

For this research, service-oriented data will be used upon which aspect-based sentiment analysis will be performed. The data encompasses text transcripts of customer calls towards KPN. The text is service-oriented as it involves customers reaching out to KPN with service or product-related questions. This contains opportunities in the dimensions of service quality and delivery, i.e. technical errors or customers calling about their subscription respectively. Originally customer calls are spoken text, a third party converts this speech to text using a specifically designed algorithm for this purpose. All transcriptions are depersonalized as well. In consequence, the resulting text contains more noise than written text. This research is not concerned with nor has any influence on the speech-to-text conversion.

In terms of type of customers it should be mentioned that the calls used in this research are SME B2B calls.[31] The reason to choose for the B2B market is that KPN simply tests and experiments their new data-driven methods and models on their SME business market, predominantly because it is smaller in size; in terms of service revenue the ratio between the business market and consumer market is 20:80.[32] In the SME B2B market there are already tenthousands of customer service interactions per month. More specifically this research focuses on the *Kleinzakelijk* (KZ) customer segment. This includes businesses with up to six employees of which many self-employed persons, these types of business customers are very similar to consumer market customers. This segment is chosen because the algorithm to filter on product groups is only available for this customer segment. All customer service interactions used are in Dutch language, in data production there is no differentiation made between Dutch and (regional) languages such as Frisian, Limburgs or Low Saxon.

4.2 Data collection and handling

The flow of data collection and processing is shown simplified in Figure 4.1. First of all, data is extracted from ElasticSearch, which provides data storage solutions for KPN. Specifically the data will be focussed on customer interactions about the *mobile phone* service domain of KPN; *KPN Mobiel.* This service scope is intentionally taken to narrow the scope of the research for methodological reasons, as KPN provides a very broad range of service areas. The extracted

[31] *Small & Medium Enterprises and **Business-to-Business.

[32] KPN Integrated Annual Report 2021, p. 101.

data contain transcripted *customer contact calls* of the most recent 10 months, from customer contact calls extracted data ranges from October 2021 to July 2022.[33]

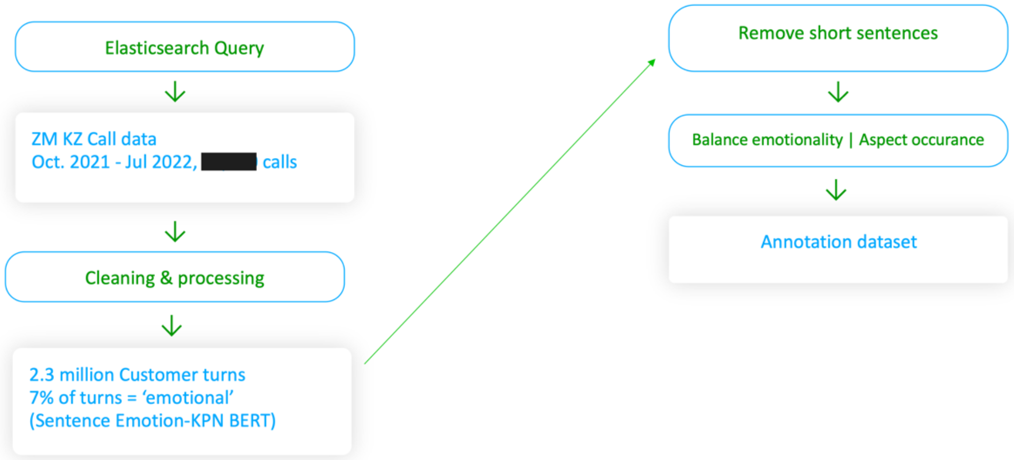

Elasticsearch Query
ZM KZ Call data
Oct. 2021 - Jul 2022, [redacted] calls
Cleaning & processing
2.3 million Customer turns
7% of turns = 'emotional'
(Sentence Emotion-KPN BERT)
Remove short sentences
Balance emotionality | Aspect occurance
Annotation dataset

Figure 4.1: *Flow of Data Extraction and Processing*

Between 50,000 and 100,000 calls are extracted within the given time-range for the *Kleinzakelijk* customer segment and *KPN Mobiel* product segment.[34] Then, the call conversations are split into turns (sentences) in which the customer is speaking. After all, customer opinionated text is desired, agent turns are left out. In about half of cases, the full customer call text would contain too many tokens for BERT models to properly train and predict. The turns are cleaned from timestamps. Predicted turn-sentiment information is kept in a separate column but removed from the text. 844 NAs in the data are removed (~0.03%). Further, unusually long turns, longer than 100 words are removed, this affects 7,075 turns (~0.30%) and can be considered outliers.

No syntactic or semantic changes are deliberately made; the data should reflect the real spoken text as closely as possible. The result is 2.3 million turns from speech-to-text without any interpunction. Table 4.1 gives a quick overview of the data and some statistics regarding sentiment and length can be found in Table 4.2 and 4.3.

4.2.1 Creation of labeling dataset

The following stage is to work towards the annotation dataset. As discussed before, for every sentence the model should be able to perform two tasks: i) extracting business-relevant aspect terms and ii) extract aspect-sentiment for all aspect terms in the sentence. For every sentence therefore, all relevant aspect terms and corresponding sentiment have to be labeled. In order to make the labeling process more efficient the labeled dataset should include (only) sentences with aspects from the list of business-relevant aspects (Chapter 3.2). Table 4.5 shows these aspect terms along with their frequencies on a sample of the data.

[33] The research does not take customer-chatbot texts into account.

[34] Exact number of calls cannot be provided as contact volumes are not officially published publicly.

For the annotation dataset it is valuable to include sentences with multiple aspects. That way more aspects can be labeled without increasing the number of sentences to be annotated. It is also interesting to include sentences which are more ‘emotional’. The reality is that the majority of sentences are neutral, this does not necessarily mean that all aspect-sentiments are also neutral (on the contrary), however by including emotional sentences the probability to find more negative and positive aspect-sentiment is expected to be higher. That is why a stratified sampling method to create the annotation dataset is utilized.

The annotation sample is made according to three (random) sample filters. Figure 4.2 illustrates this process. First, ‘sentimental’ turns are sampled that contain one aspect (*1B*) or more aspects (*1A*). These sentences are relatively ‘long’ and have negative or positive sentence-sentiment, as predicted by a sentence-level emotion BERT. For every unique aspect at least 35 ‘emotional’ examples are sampled.[35] Second, sentences are sampled from neutral sentences (also containing aspects), these sentences are somewhat shorter. Now a minimum of 20 examples are sampled for every aspect. Lastly, 150 random turns are sampled, these are much shorter and do not contain aspects necessarily. Within the boundaries of these criteria, sampling is random. This filtering should ensure a variety of sentences with enough different aspects and hopefully different aspect-sentiments. The annotation dataset contains minimal 55 example turns for every aspect.

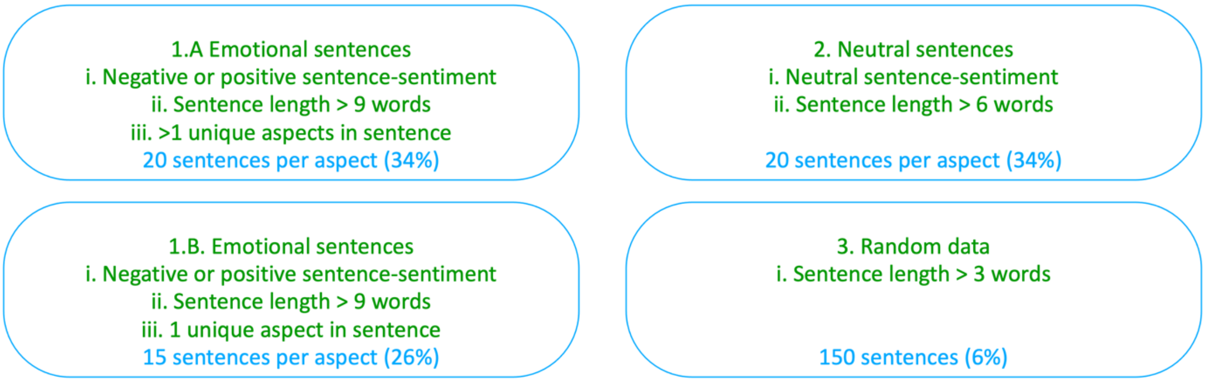

1.A Emotional sentences
i. Negative or positive sentence-sentiment
ii. Sentence length > 9 words
iii. >1 unique aspects in sentence
20 sentences per aspect (34%)
2. Neutral sentences
i. Neutral sentence-sentiment
ii. Sentence length > 6 words
20 sentences per aspect (34%)
1.B. Emotional sentences
i. Negative or positive sentence-sentiment
ii. Sentence length > 9 words
iii. 1 unique aspect in sentence
15 sentences per aspect (26%)
3. Random data
i. Sentence length > 3 words
150 sentences (6%)

Figure 4.2: *Sampling method to create the annotation dataset*

4.2.2 Labeling data

The annotation dataset contains 2,570 unique sentences with 44 unique aspects. For every sentence in this data, every occurring aspect must be labeled and their accompanying sentiment. Over 4,500 aspects in different contexts and their accompanying sentiment are labeled this way. For labeling a specific labeling tool is used, although other tools could likely be used as well.[36] The output is a .txt file with per line the sentence (aspect T marked), the aspect and the sentiment. If multiple aspects occur, the {sentence, aspect, sentiment} output is repeated for as many aspects occur. This output is coded to IOB-format, the required model

[35] There might be more, as a specific aspect can also occur in sentences sampled for other aspects.
[36] Labeling tool used: https://github.com/yangheng95/ABSADatasets/tree/v2.0/DPT

input (see: Yang et al., 2021). IOB-type formats are commonly used as input for aspect extraction models. The labeling task is done together with a call domain expert at KPN.

4.3. Representation of data

Now that the data collection and development of the annotation data set is described, the data overview is given in table 4.1. Some descriptive statistics on sentence length, sentiment and aspects in the raw data and labeling dataset are discussed in this paragraph.

Table 4.1.: *Overview of the dataset in frequencies*

Data Type	Frequency
9 months of KZ Mobiel product calls	50,000 - 100,000 calls
… in customer turns	2,273,372 customer turns
Cleaned data	2,256,216 customer turns
Labeled data	2,570 customer turns

The sentence length of the raw call data is visualized in table 4.2. Almost a third of the data contains three or less words. For annotation and prediction efficiency, very short turns are discarded in the selection of annotation data and later on in prediction. It is trivial that these very short sentences are unlikely to contain business-relevant aspects or information. Typical examples of these short sentences are "*Ja, ja, ja*" or "*Hallo, goedemiddag*".

Table 4.2: *Relative frequency of sentence length in the full data*

Sentence wordcount	Relative frequency
< 4 words	31% of the data
< 5	36%
< 6	41%
< 10	57%
> 70 words	1.14% of the data
> 80	0.7%
> 90	0.5%
> 100	0.3%
Total	100%

Table 4.3: *Relative frequency of the number of aspects in a sentence in the full data*

Aspects in sentence	Aspects in sentence
None	77%
1	15%
2	5%
3	2%
4	1%
5 or more	1%
Total	100%

Overall turn-level sentiment in the full data is illustrated in Table 4.4a. 7% of the sentences are predicted 'emotional', with almost as many positive as negative sentiment. The rest is overall neutral. Table 4.4b shows the labeled *aspect-sentiment* terms in the labeled data. 18% of the annotated aspects are 'emotional'; with much more negative emotional sentiment than positive. As customers set forth problems they experience in the calls, it is rather plausible that negative sentiment is dominant. More emotionality now is likely found because of the sampling of emotional sentences and the extraction of more granular sentiment information the hypothesis that aspect-sentiment is more fine-grained than sentence sentiment.

Table 4.4: *The frequency of turn-sentiment in all data and annotated aspect-sentiment in the annotation data*

Table 4.4a: *Predicted sentiment for turns in the full data*

Turn Sentiment	Customer Turns	Relative frequency
Neutral	2,116,816	93%
Positive	69,730	3.1%
Negative	87,024	3.8%

Table 4.4b: *Annotated aspect-sentiment in the labeling data*

Aspect-Sentiment	Relative frequency
Neutral	81.8%
Positive	3.2%
Negative	15.0%

Table 4.3 shows how many times aspects occur in sentences in the raw data. 24% of the data contains at least one aspect. The majority of sentences contain one aspect, however 10% of the sentences contain multiple aspects. In the annotation data sentences with multiple aspects are far more common because of the stratified sampling method (Figure 4.2). Table 4.5 shows the list of all aspects or aspects and their frequency in the data. Why and how these aspects are chosen, is defined in the previous chapter, the methodology.

***Table 4.5**: All Business-relevant Aspects and their frequency in sampled raw data ±*

Aspect	Count	Aspect	Count	Aspect	Count
kpn	109,896	bestelling	11,757	prepaid	3,690
abonnement	90,662	korting	11,613	bereik	3,394
bellen	57,709	account	8,131	prijs	3,244
internet	45,625	service	7,663	kapot	3,090
simkaart	40,574	netwerk	7,649	verbruik	2,985
factuur	38,045	buitenland	7,591	pincode	2,762
toestel	36,604	inloggen	6,897	overstap	2,282
contract	25,483	opzeggen	6,155	mobiele telefoon	2,179
mail	23,197	mijn kpn	5,527	klantenservice	1,664
telefoonnummer	19,664	provider	5,278	sim only	1,524
facturen	14,572	wachtwoord	4,757	internetverbinding	1,408
wifi	12,694	voicemail	4,494	signaal	1,408
kosten	12,665	ziggo	4,413	ip	1,266
kleinzakelijk	12,513	hussel	4,071	internetsnelheid	378
bundel	12,469	storing	4,001		

± The aspect frequency is found with regular expressions.

Chapter 5. Results

This chapter will discuss the results for the deep-learning model built for aspect-based sentiment analysis. The chapter reviews both the ‘technical’ results as well as the business relevant results or insights. The first paragraph regards the outcomes of the training procedure, herewith both hyperparameters as metrics will be discussed. It will be debated which model configuration is best and why. The next paragraph will look at the model in action, with some real examples. Ultimately, paragraph 5.3 shows results on newly generated customer call data to test the models’ ability to produce business relevant insights (for KPN) in alignment with the research question: ‘*how can opportunities for product and service improvements be identified from customer service calls using a data-driven approach?’* More specifically this chapter answers to the last two sub-questions: *(iv) How does sentiment differ among different aspects and how can this be visualized? And (v) To what extent can identified sentiment for product/service aspects be tracked over time and what is an appropriate user interface?* The results are discussed in paragraph 5.4.

5.1 Training & Performance

There are a lot of decisions to be made with regard to arranging the training process of deep learning models such as BERT-family models. In order to train the model labeled data is needed, as discussed in Chapter 4.2. In line with machine learning standards this labeled data is divided into three sets: training data (80%), validation data (10%) and test data (10%). The model trains on training data, it learns the distribution, patterns and ‘logic’ of the data. Based on the training data the model alters its weights of the parameters. After each training epoch the model validates what it has learned on (withheld) validation data. A training epoch corresponds to one flow of all data through the model; more about epochs can be learned in the next section. Validation data is used to track the model’s performance after every epoch, this is convenient because if there is no gain it could be decided to early stop the algorithm. After the training procedure is completed, the model is tested on test data in order to assess its performance.

For this research more than 750 different model configurations were trained. Since the performance gains with regard to the hyperparameters are not linear it is necessary to do an extensive search over parameters and model performance. With ‘not linear’ it is meant that, for example, increasing the learning rate will not per se increase or decrease the performance of the model because performance is not linear in learning rate, but also because the learning rate interacts with other hyperparameters. Certain combinations of parameters might work well together, but if you change one of the parameters model performance might worsen. Because of the large number of possible combinations, it is not obvious that the global maximum performance will be found, however it gives a good grasp on what the possibilities are in terms of good results. The next section will discuss the hyperparameters involved in the model, the section thereafter will discuss the metrics to track (dual task) performance.

5.1.1. Hyperparameter fine-tuning

Beforehand, all important and adjustable parameters are collected and examined. It is a matter of trial and error to find out what (combinations) of parameters work best. However, testing the model one fine-tuning configuration at a time is really time-expensive. Therefore, a hyper grid search was developed to efficiently tune the model using KPN's computation resources.[37] Even with 4x core NVIDIA GPUs training time expands quadratically for every new configuration setting added to the grid (eventually leading to out-of-cache problems). Therefore the search grids coded were not too large, and after every run the relations between the parameters (and performance) were manually evaluated. By logging every training configuration and associated performance and its metrics a training history database was built with all training information and performance. In total 742 different configurations are trained (and tested). The best 30 performing configurations from this database are displayed in Appendix 5. When the train history database contained a few hundred logs, a random forest model was estimated to learn how parameters interact with test metrics (performance). It did not generate new insights, however it did confirm certain thoughts about parameters. This algorithm shall not be reported on further.

In total there were 16 hyperparameters that the model is tuned on. These hyperparameters are stated in Appendix 4. Several important hyperparameters will be further discussed below: the pretrained model, the ATESC architecture, the seed, the batch size, learning rate, number of epochs, LCF, SRD-window and the dropout.

I. Pretrained model:

For this research four pretrained models were downloaded to train on: (i) multilingual BERT, (ii) BERTje, (iii) KPN BERTje and (iv) RobBERTv2. Multilingual BERT is not tested extensively. In theory the Dutch models should outperform the multilingual model, empirical research has also confirmed this (De Vries et al, 2019). When this research's first exploratory results also confirmed this, those were reasons enough to drop mBERT. The other three pretrained models were tested extensively. Their performance did not vary significantly. Nonetheless, RobBERTv2 tended to perform better on ATE, however worse on ASC. KPN BERTje's had some configurations that performed best on the ASC task, however too much at the expense of ATE. BERTje's best configured models performed best overall in the end. Appendix 5 states the quantile performances ATE and ASC scores for the three pretrained models. Note that ultimately a model that performs well on both tasks is desired.

II. Number of epochs & Batch size:

The number of epochs determines how many times the training data flows through the model, upon which it changes its training weights. Enough epochs are needed to give the model time

[37] While it is possible to run a BERT model on a CPU, proper fine-tuning BERT requires a GPU.

to learn, in that sense the amount of required epochs depends on the complexity of the task to be learned. However, too many epochs hurt the model's generalization capabilities as it will lead to overfitting on the training data. Devlin et al (2019) recommend to use two to four epochs in the fine-tuning procedures. Four epochs without early stopping are chosen as there are two tasks to be learned, and especially the aspect-sentiment categorization task is difficult and more so due to imbalanced classes. Finally, in training procedures there were no clear benefits found of using more epochs, while it does strongly increase training time. That is why mostly four epochs were chosen to train the models.

It is not evident how batch size affects model performance. Too large batch sizes lead to poor generalization, too small batch sizes lead to slower convergence. Computational constraints also play parts, batch sizes of 32 or higher led to more problems. The optimal batch size is not tested extensively (16 or 32 was default), only for the top models also smaller batch sizes were tested; that resulted in better test performance.

III. Seed:

The seed is actually one of the most determining parameters. That is because BERT models are inherently non-robust mostly due to randomness in optimization. The seed affects two random components in BERT models: first of all, the seed determines the random initialization of stochastic gradient descent (the iterative algorithm that forms the basis of neural network used to minimize the training loss). Second, the seed influences the dropout.

In this research one seed number is picked and for that seed the hyperparameters are optimized. Another strategy is to tune the seed as well and cherry pick the seed giving the best test performance. It is even known that some seeds, known as bad seeds, lead to bad initializations and bad outcomes. In a sense, the seed is also a hyperparameter to be tuned. The idea is that with the best training seed, the resulting fine-tuned model also performs well on new unseen data, given that the test set is large enough and a true representation of the real world data. Initially the seed was not tuned in this research, however the best performing three models were also trained on ten different seeds to see how robustly the models perform on different seeds and to see which seed they perform best on.

IV. Dropout probability:

The dropout parameter determines the ratio of weights of the feed forward networks and attention given to certain words that drop out. This helps reduce overfitting because it forces the network to find multiple 'learning paths' to reduce (training) loss function. The direct consequence is that the model that way learns to generalize better, i.e. it does not depend on certain specific paths. Another way of looking at dropout is by comparing it as a model ensemble technique, deactivating random neurons during training essentially results in multiple different networks that are ensembled during the evaluation phase (Kravets, 2021).

The final chosen dropout is with 0.4 quite high. This will prolong the convergence rate of the model, however it will improve generalizability (Yang et al., 2021).

V. Model architecture, SRD, LCF mechanism and Learning rate:

The tried models share the same model architecture, however *fast-lcf-atepc* models use one bert model for global context and local context feature generation, while *lcf-atepc* models use two BERTs. The (*Fast-)lcfs* models use syntax-based SRD, which could potentially substantially improve performance for contextualized models for aspect extraction and aspect sentiment classification and make models more domain-independent (Phan & Ogunbona, 2020).[38] The fast-lcfs variant is chosen because it is faster and allegedly does not really differ in performance.[39] No clear performance differences were found in the fine-tuning procedure. A faster model is moreover convenient for usage (predictions) later on. The SRD threshold parameter affects the identification of the local context of a targeted aspect, based on the position of words. If the SRD threshold is equal to 3, for example, every contextual word with a SRD of less than the threshold is regarded as the local context (Zeng et al., 2019). Larger SRD will thus widen the local context. Best performing models have a SRD threshold of 5. The LCF mechanism parameter subsequently determines how to treat words outside the local context. With context dynamic masking (CDM) the words are entirely masked (discarded), with context dynamic weighting (CDW) the words are more modestly weighed down. The fusion approach merges both methods. CDM and CDW worked best for Chinese language while fusion worked best for English data (Yang et al., 2021). Since Dutch is far more similar to English, 'fusion' is chosen for most training configurations. Finally, the learning rate determines how quickly the model converges, it affects the step size of the stochastic gradient descent algorithm. Lower learning rates might lead to slower convergence towards the minimum of the loss function. Most high performing models had a learning rate of 1e-05 or 2e-05.

Table 5.1 contains the hyperparameter settings for the chosen model. Note that parameters with an asterisk were not found to be of importance. The optimizer is not changed.

Table 5.1: *Fine-tuned configurations of the chosen final model*

Hyperparameter	Value	Hyperparameter	Value
Pretrained BERT	BERTje	Seed	52
Model architecture	Fast-LCFs-ATESC	Batch size	4
SRD (threshold)	5	Dynamic Truncation*	True
LCF mechanism	Fusion	Max. Sequence length	100
Number of Epochs	4	Syntax-based SRD*	False
Learning rate	1e-05	BERT SPC*	False
Dropout	0.4	SRD alignment*	True
L2 regularization*	1e-05	Optimizer	'Adamw'

[38] Syntax: *''the arrangement of words and phrases to create well-formed sentences in a language''.*

[39] Yang (2022), https://github.com/yangheng95/PyABSA/discussions

5.1.2. On the performance metrics

This section explains which metrics are used and what they implicate. As the model performs two tasks this is slightly more complex than usual. The performance metrics used are Macro F1 score for Aspect Term Extraction (ATE), the macro F1 score for Aspect-Sentiment Classification (ASC) and the accuracy for Aspect-Sentiment Classification.

The macro F1 score is the (arithmetic) mean of the F1 score for every (predicted) class. For ATE this is the average score over all 44 aspects and for ASC the average score over the three sentiment classes. The F1 score in turn is the result of the following formula:

$$F = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \tag{5.1}$$

$$\begin{aligned} \text{Precision } &= \frac{TP}{TP + FP} \\ \text{Recall } &= \frac{TP}{TP + FN} \end{aligned} \tag{5.2}$$

Where, True Positives (TP): number of aspect terms predicted correctly.
False Positives (FP): number of non-aspects predicted as aspects.
False Negatives (FN): number of aspect terms classified as non-aspect terms.

The F1 score is the harmonic mean of the precision and of the recall. The precision is the rate of correct predictions per class over all true items for that class - how many retrieved items are relevant? And the recall is the rate of correct predictions per class over all predicted items being that class - how many relevant items are predicted? Only the Macro F1 scores are used and necessary to evaluate the training/fine-tuning process. This had also to do with the ease of extraction; it would have been necessary to back-engineer a lot in Yang’s (2022) software to also extract e.g. precision and recall or Area under the Curve (AUC).

A perfect macro F1 score would mean that all F1 scores for all classes should be 100%. Looking at ASC F1; if one of the three classes would always be predicted wrong, the maximum score would be 66.7%. Because of the sentiment class imbalance, the F1 score deviates quite a bit from the accuracy; sparse classes are inherently more difficult to predict for models. It is suspected that this makes the F1 score among different fine-tuned models volatile. Volatility is assumed to arise because the amount of positive classes in the validation and test data are small, 18 and 16 classes only respectively. This is too few holdout data to draw a stable conclusion which fine-tuned model is best, at least with respect to ASC. Therefore, the best model's performance on training and validation data is shown as well and to mitigate this issue somewhat the top models are tested on multiple seeds.

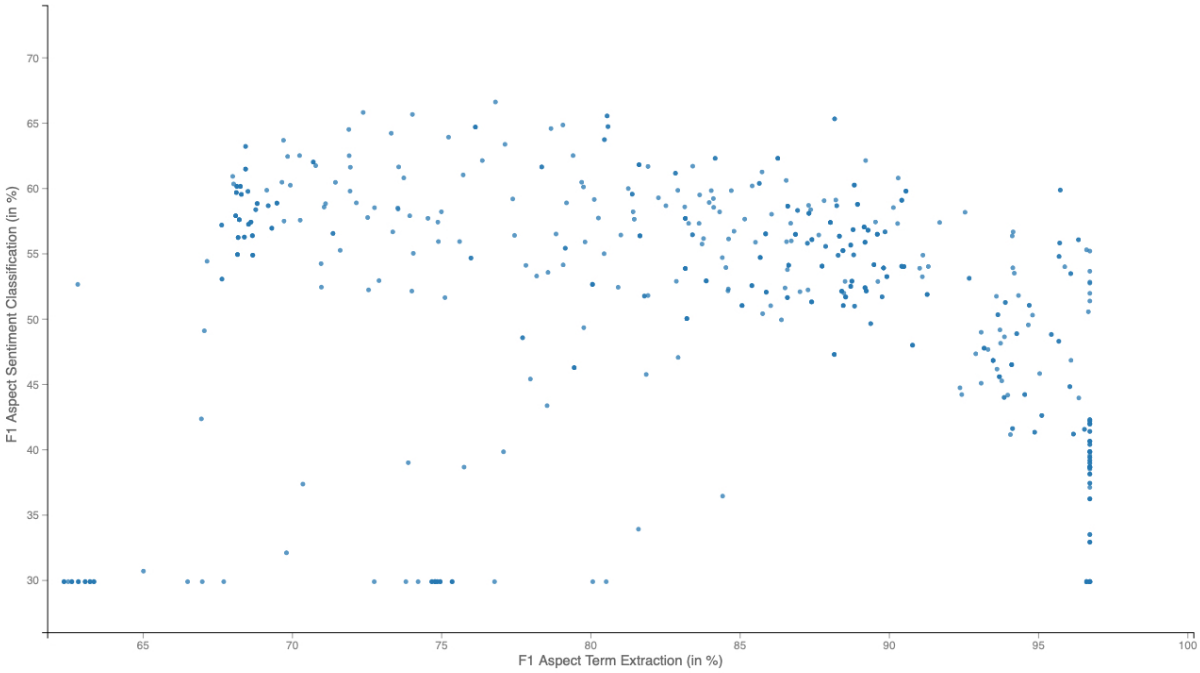


Figure 5.1: *Scatterplot of Test data performance of fine-tuned models for ASC and ATE*

5.1.3 Choosing the right model

Hyperparameter tuning was found to be of importance as it greatly increased performance compared to the initial runs. Appendix 6 includes a table with the 'best' 20 configurations. In total 742 configurations were fine-tuned and tested (Figure 5.1). What model is best is not trivial. First of all, because the test data contains too few positive aspect-sentiment classes. Second, the fact that the models perform two tasks makes deciding even more complex. The models that are best on one metric do not perform best on other metrics as well. This can be seen in Figure 5.1 which shows the test performance on both the ASC task and the ATE task for different fine-tuned models. A few outliers are excluded in the plot. It is shown that there is generally a downward slope between the scatters. Note that the graph is not equally scaled.

For this research it is therefore decided to choose the 'best overall' models based on their sum of F1 ATE, F1 ASC and Acc ASC on test data. This might be a bit biased towards favoring Aspect-Sentiment Classification. However, since the models score generally very good on ATE there are diminishing marginal returns; a one percentage point increase in the ASC F1 score from 52% to 53% is more interesting than increasing ATE F1 from 90% to 91%. The ASC Accuracy is also included in the summed measure, because it is less volatile than the F1 score.

The top five models are tested on different batch sizes and different seeds. A model that performs well on multiple seeds is more robust, as explained in section 5.1.1. This results in three 'best of the best' models. The predictions for these models are manually checked on a small sample of test sentences. The best performing model on the metrics also checked out best on the manual check (see next paragraph), reasons enough to pick this model.

The final model performs best on Aspect Term Extraction with a Test-ATE F1 over 86%. This is generally a good performance, however some fine-tuned RobBERTv2 models have ATE F1's of 95%. The model performs relatively very well on ASC F1 over 62%. It is amongst the highest performing models here, quite some other fine-tuning configurations had ASC F1's in the range of 50-55%, the best configuration had a 66% ASC F1. The model's accuracy is quite good as well, the best configurations' accuracy was 85%. These single metric comparisons are merely meant as a benchmark, because the best configurations based on ASC F1 were poor in ATE F1 (70-75%), top ATE performing configurations lack in ASC (F1: 40-55%), and the top accuracy configuration lacked in ATE F1 and ASC F1. The chosen final model performance metrics are shown in table 5.2. The chosen fine-tuned model scores relatively high on all three test metrics, as shown in Figure 5.1.

Table 5.2: *Performance of the best model on different metrics and data*

Data	ASC Accuracy (%)	ASC F1 (%)	ATE F1 (%)
Train	83.37	58.59	91.23
Validation	81.27	55.38	87.16
Test	82.60	62.33	86.25

Besides generally very good performance in all test metrics, the model is also quite stable in performance on different seeds. The performance of this model on different seeds is shown in Appendix 7. This is important to make sure this model configuration performance was not just an outlier. If a model performs well on multiple seeds it can generate good results with multiple paths, this increases the chance that the model will also work on other unseen data. The model also has a high dropout rate of 0.4, which increases generalizability too.

In table 5.2 apart from the test metrics also training and validation metrics are shown. Training scores are slightly higher for ASC accuracy and ATE F1. This could be some overfitting, especially for ATE, however it is not peculiar that training performance is higher than test performance. ASC F1 is actually lower, but that is most likely a coincidence due to the class imbalance (see Appendix 8 for the class distribution among modeling datasets). Due to the small sample of positive classes in the validation and test data, it might be that the test classes are easier to predict. This law of small numbers could also account for the difference in ASC performance between validation and test data.

5.2 Model inference

Because numbers simply do not tell the whole story for language tasks apart from the test metrics the model is also manually evaluated on sample data. NLP tasks are hard to evaluate based on metrics only. This paragraph therefore provides six prediction examples to highlight caveats and possibilities. These are not only based on the examples shown below. But also learnings from the manual inspection of the model on a large sample of turns.

Example 1: dat was mijn tegen maar die is al 2 jaar geleden opgeheven en ik wil nu ben ik al 1 half jaar mee bezig om mijn <abonnement:Negative> wat ik dus het bij jullie als <kleinzakelijk:Negative> om te zetten naar gewoon 1 prive abonnement

–

Example 2: over hoor top top top top het moet conflicten te vermijden de kabel d r uit de de kabel van de . . . uitgetrokken hij zegt nu zwak of geen <signaal:Negative> is dat voldoende of heeft u weg die tv nodig

–

Example 3: ja het is alleen het in en dan heb ik het ook niet over <wifi:Neutral> op moment dat ik mijn . . . ga gebruiken dan dan staat de teller lopen

The majority of turns are relatively short and if they contain aspects, contain one (or two) aspects, mostly with neutral aspect-sentiment.[40] Above are some typical examples shown. Example 1 seems like a good result, the customer struggles to change the KZ subscription to a regular subscription. Example 2 is too, while the customer says *'top'* (positive term) four times in a row, the 'zwak of geen' indicates that 'signaal' should be negative. This might be a nice example of local context focus in practice. Example 3 is another typical sentence, it is not clear what is going on here (partially due to data quality); neutral seems the right choice.

*[41]

*Example 4: dat is niet waar want dat is jullie bedrijf nu 1 ik zit daar 1 half jaar lang in mijn . . . en het 3 hetzelfde probleem zat ligt niet mijn <internet:Neutral> thuis denk ik dan maar ik ben al helemaal klaar mee want ik heb slecht <bereik:Positive> en ik ben . . . ik ben klant en ik mijn <abonnement:Positive> is bijna afgelopen en <overstappen:Positive> van ik want anders ik ben helemaal klaar met de <service:Neutral> van jullie

–

*Example 5: ja dankje zei dat mijn <abonnement:Neutral> . . . bij <kpn:Neutral> zakelijk direct omgezet naar 1 mkb blijkbaar met de migratie heeft <kpn:Neutral> allerlei dingen fout gedaan waardoor geen <facturen:Negative> binnen zijn gekomen helaas dus ook niet betaald zijn en nu kan ik niet <bellen:Negative> naar buiten van de week zou dat opgelost worden en zo ik kunnen <bellen:Negative> maar dat duurt nog 2 dagen ik word het 1 beetje zat snap je het gaat om . . .

The model is completely lost in example 3, this is also a long and polarized turn with five aspects (what less than 1% of the sentences have). Bereik and service clearly should have been negative, abonnement and overstappen rather neutral. In example 4 the predictions are better, maybe the second kpn should've been negative. One caveat here seems that the model tends to give nearby aspect terms the same sentiment. Another caveat might be that it is hard to switch from negative sentiment to positive sentiment.This would possibly be even worse without local context focus in the model. In general, the longer the turns are, and the more differentiated sentiment among aspects is, the harder it is for the model.

*Example 6: het <modem:Neutral> ja sorry ja nee geen <internet:Neutral> dit is in 1 vakantiehuis en ik daar pak staat een computer en pak ik dus van <4g:Neutral> pak ik via die <sim only:Neutral> pak ik data af kun jij helpen ik weet weet niet hoe

This last example highlights the model's ability to find new aspects, both *modem* and *4G* were not labeled terms (at least structurally). It is extraordinary that the model is able to extract these 'extra' aspects, they do represent the core meaning of the sentence and could be business relevant.

[40] See table E & F, par. 4.3.

[41] * Printed text is a truthful exemplification of the real one due to confidentiality for example 4,5 & 6.

Examples 4,5 and 6 are not representative for all turns, they are longer and more complex than usual. More generally, the lessons learned from the manual evaluation of a batch of thirty test turns are: first, sentences with different aspect-sentiments within the sentence are harder to predict. In addition, the model tends to avoid predicting multiple different aspect-sentiments per turn, however it is capable of doing so. This held for other fine-tuned models as well. Sentences with uniform aspect-sentiment seem easier to predict. Neutral aspect-sentiment is the easiest to predict. Negative aspect-sentiment seems to be the most difficult to predict (sensitivity). However the recall of the model for negative sentiment is likely high, it did not predict often negative sentiment where it should be neutral or positive. Typically, the model predicts neutral sentiment for wrong predictions. Another general finding is that true sentiment is sometimes vague and even difficult to judge for humans let alone for models, besides this has to do with poor data quality and interpretation.[42] As can be seen in the examples, words are often repeated or fallen away.

5.3 Towards relevant findings for business stakeholders

In-text predictions are a nice feature that enables users to zoom in on the model's predictions to better understand them. However, the model's ability to generate insights from large scale data, unravels the true power of fine-grained sentiment information. To highlight the big text data insight possibilities some graphs are included to answer sub-questions *(iv)* and *(v)* regarding visualization of sentiment among aspects and aspect-sentiment over time. These figures are originating from a dashboard built for business end-users. This dashboard is not included in this paper. In the dashboard it is possible to filter on aspects, time-range, selected products and other descriptive categories on calls.

To create an insights dashboard, new service calls data is extracted from elastic search, in the time range from Oct. 2021 until Oct. 2022. The model can run predictions directly on the texts derived from customer service calls. The model's predictions are processed such that columns for all extracted aspects and their (predicted) sentiment are attained. To create the final 'dashboard data', the obtained data frame is subsequently 'exploded' on these columns, so that every aspect and corresponding aspect-sentiment have their own data row. Using the counts of aspects and aspect-sentiments, all kinds of visualizations can be created. This aggregation of aspect-sentiment pairs is sometimes referred to as the third task of ABSA, necessary to present the results.

5.3.1. Visualizing sentiment per aspect

First, sub-question (iv) is discussed: *How does sentiment differ among different aspects and how can this be visualized?* Figure 5.2 shows a visualization possibility to compare sentiment among aspects. Note that not all aspects are shown in the graph. This figure is derived from a

[42] Note that these 'findings' are supplementary to the test metrics and are not based on a large enough sample. The decision what model is best is mostly decided by the test metrics.

non-representative sample of data, because its purpose is only to showcase a visualization possibility. Because of the sensitivity of the data some aspects are masked as <Aspect X>.

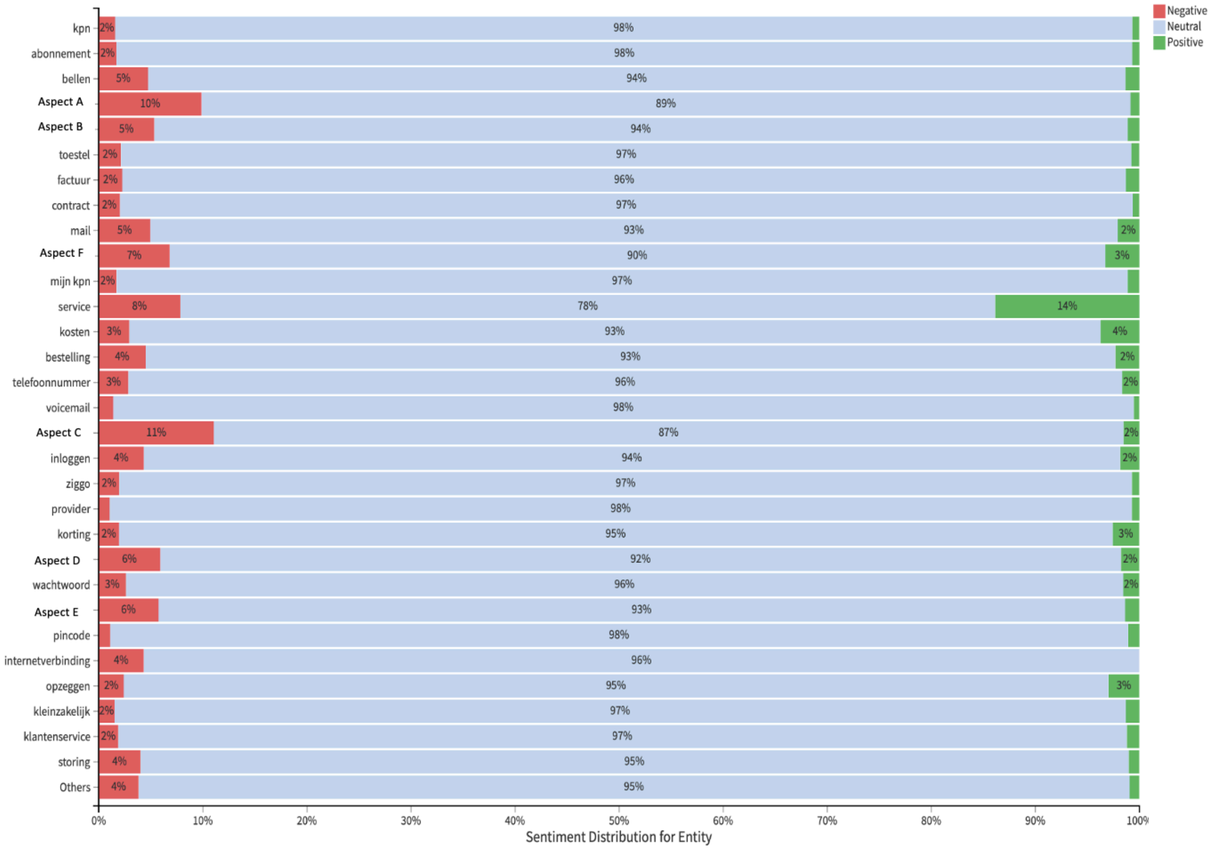


Figure 5.2: *The Aspect-Sentiment Breakdown: how do customers feel about KPN Mobiel features?*

Figure 5.2 shows what particular aspects of KPN Mobiel products customers like and dislike, or at least how they express their feelings towards it. In this figure it is for examples shown that customers mention *Aspect A* and *Aspect C* relatively often in a negative way (see e.g. the benchmark 'Others'). On the other hand, customers do not seem to have many complaints about the 'Mijn KPN app'. Only 2% of the mentioned aspects hereabout are in negative context. The aspect 'service' really jumps out in positivity. However, many aspect terms are negative, it is thus more interesting to compare the amount of negativity to other aspects. This type of analysis can be used to make product/service improvements, by trying to improve underperforming aspects. More opportunities for product and service improvements can be identified by focusing on fewer aspects, and e.g. visualizing the barplot for a certain month. In this regard it is also interesting to look at other variables, e.g. by combining the aspect-sentiment with selected products, customer problems, follow-up calls or call topic.

5.3.2. Visualizing aspect-sentiment over time

Next sub-question (v) is discussed: *'to what extent can sentiment for product/service aspects be tracked over time and what is the appropriate user-interface for this'.* Figure 5.3 and 5.4 depicted below are not representative visualization of the data as their purpose for this work is to highlight visualization and insight extraction possibilities rather than creating insights. Some labels are masked due to confidentiality of company data.

In Figure 6 the aspect-sentiment is shown over time for two features of KPN Mobiel: ‘*service’* and ‘*Aspect G*’. It allows for close tracking of the positive and neutral sentiment for one specific aspect over time. Figure 6 shows that *Aspect G* has an increase in negative sentiment in august. Such change of sentiment is interesting to link to certain corporate events or product or service changes. Figure 6 could also be shown on a relative frequency scale to be able to better compare changes in positive and negative sentiment, since customers mention most aspects structurally less in a positive manner. This is not visualized.

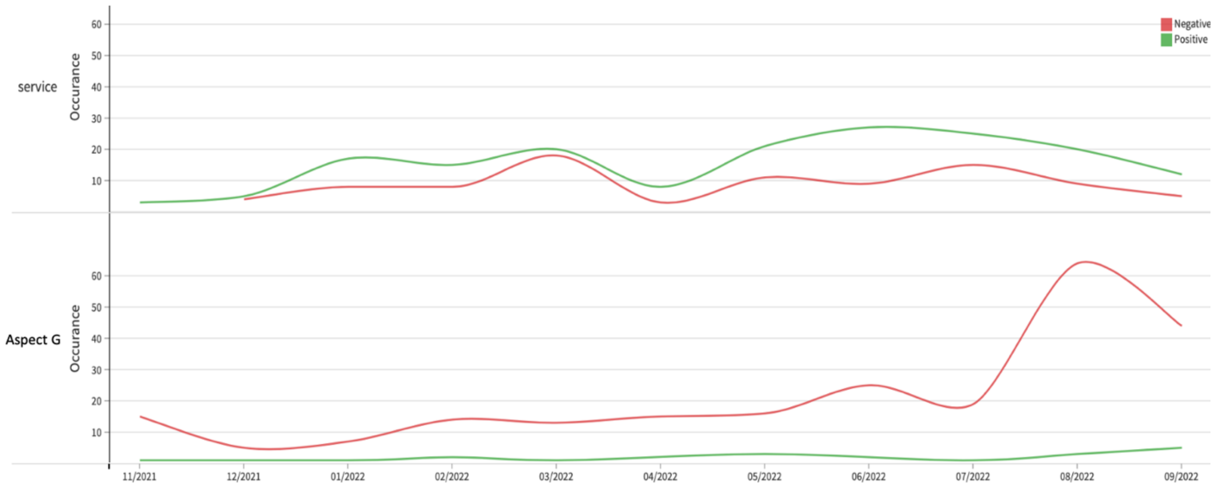


Figure 5.3: *The Aspect-Sentiment over time: how do positive and negative customer sentiment per aspect develop over time?*

Figure 5.4 also allows to track aspect-sentiment per aspect over time, however with the use of the aspect-sentiment score. The aspect-sentiment score is a numerical measure derived from numerical aspect-sentiment where each positive aspect-sentiment is counted as 1, each negative as -1 and each neutral as 0. By taking the average score per quarter for each aspect Figure 5.4 is created. The benefit of this score is that it makes it easier to compare multiple aspects in one graph at the same time compared to Figure 5.3. It can be seen that the aspects follow different trends, it is not directly clear why. To increase the information density, time can also be expressed in smaller units such as months or days.

For both figures it holds that insights can be customized using filters and extra variables. Therefore, an interactive dashboard is a more appropriate interface than single graphs. A sample of a dashboard is shown in Appendix 9. Given the large amount of aspect categories, filters are indispensable. For many visualizations it is necessary to filter aspects to prevent information overload. Besides, it provides the opportunity to visualize aspect-sentiment over different time units, e.g. days of the month or even hours of the day. Moreover, it would be possible to combine fine-grained sentiment information with other business-related variables such as the products customers talk about, call reasons or whether calls are repeat calls or especially the churn calls. All this is not visualized due to the confidentiality of the data.

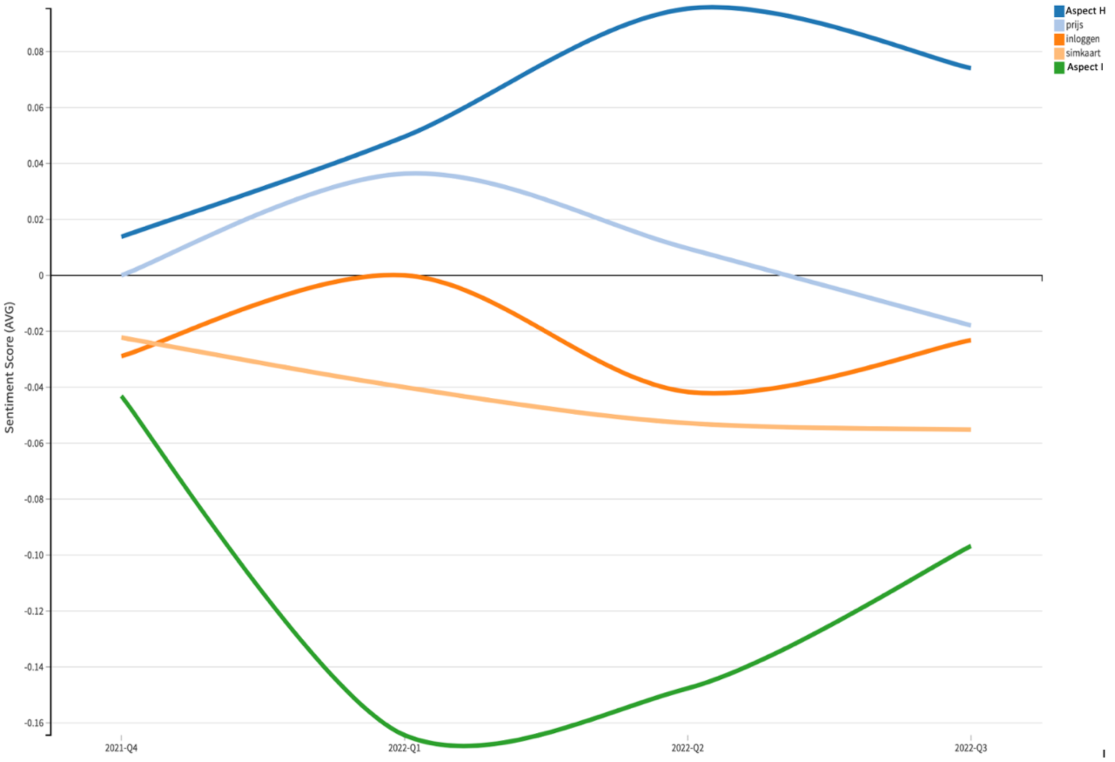


Figure 5.4: *The Aspect-Sentiment Score over time: how does average customer sentiment per aspect change over time?*

5.4 Discussion of results

This chapter sub-questions 3, 4 and 5 were discussed. Using a multi-task learning BERT model for Aspect Term Extraction and Aspect-Sentiment Classification (ATESC), customer sentiment is successfully linked to business relevant aspects. Results can be visualized both on a text-level as on a big data-level as shown in paragraph 5.2 and 5.3 respectively.

Given the (dual) task complexity the fine-tuned model performs adequately on a modest amount of labeled data. As discussed in paragraph 5.1, the training procedure can be stabilized by combating the sparseness of positive aspect-sentiment. The easiest way to do this is by combining the validation and test data, and therefore hyperparameter tune based on 20% of the data. This is not tested. Another way is to oversample the negative and positive classes in the training data. This would require advanced oversample techniques such as creating synthetic sentences with positive aspect-sentiment. Simply replicating cases is not likely to lead to improvements because the model can already ‘simulate’ this replication itself, namely by altering weights so that more attention is paid to improving positive classes compared to the others.

Next to the data imbalance, a second point of discussion is why the domain-adapted pretrained KPN BERTje underperforms compared to BERTje. Two possible reasons might be, first, that the KPN BERTje for this specific data was an outdated version. Therefore, drift can lead to worse results. This refers to the fact that the way people speak and the things discussed in the calls change over time. The second point lies in the noise that has been discovered in the tokenization during pretraining of the language model.[43] This issue is of course not present in the original BERTje that has been academically published, extensively documented and peer-reviewed. An up-to-date correctly trained pretrained domain-adapted model is expected to increase model performance, because the model is pretrained on a data distribution (type of sentences, words that occur) that is the same or very similar to the data distribution on which the downstream tasks are executed.

The model has consistently higher F1 scores for ATE compared to ASC. This has most likely to do with task complexity. Aspect-sentiment classification could be more complex for the algorithm because many aspect-sentiments are aspect-dependent. The context words that carry a certain polarity differ per aspect. Besides, sentiment classification is inherently more subjective than aspect term extraction. It may also be that the model suffers less from poor data quality for the ATE task than for the ASC task. However, the model can be steered to valuing ASC somewhat higher than ATE by changing the losses. The model performs two tasks with and therefore uses an ATE loss and APC loss, these losses are summed in training. By changing the model training architecture code and applying a weight (e.g. 1.5) to ASC loss in the summed loss the model will value ASC above ATE. This is not tested.

To improve model performance in general, (training) data input is very important. As for now input data quality can be improved in the short term by adding interpunction to the input data. Apart from that, a newer speech-to-text algorithm would likely make the data less noisy. A manual evaluation suggests that data quality is lacking to correctly learn and predict aspect-sentiment in sophisticated sentences.

In addition, this chapter has discussed the limitation of too few positive aspect-sentiment cases in the holdout data, which likely leads to ASC F1 volatility, making it hard to draw a stable conclusion about what fine-tuned model is best. This issue can be mitigated by enlarging the dataset that parameters are tuned on, either by merging the validation and test data. Or by enlarging the test set (more annotated data). This ensures that the model is tested on more (absolute) minority examples positive and negative aspect-sentiment, which model configurations that perform good on the test data on coincidence. In general, more labeled data is expected to improve outcomes because it also gives more training data.

[43] This issue is being addressed by KPN.

Paragraph 5.3 has shown that aspect-sentiment can be visualized successfully by aspect and over time. It has the potential to provide business-relevant insights for product and service improvements. Aspect-sentiment has not been further analyzed with respect to product changes or competitors' marketing actions, due to confidentiality of company data. The downside of the dashboard is the information overload that can easily arise with three sentiment categories and 44 aspects. The sentiment score is already a way to avoid this, but it loses some nuance. Another solution to address the overload are dashboard filters (for aspects). Another limitation is that the model is now a little too creative in finding new aspects. Either a model should be chosen with higher ATE Recall. Or ‘newly’ predicted aspects that are very infrequent, e.g. ‘abonnementsnummer’ should be discarded. However, infrequent aspects that could be regarded synonyms of aspects that are more frequent could be merged into one aspect, for example ‘simkaart’, ‘simkaartje’ en ‘simkaartjes’.

Chapter 6. Conclusion

This chapter concludes this work with the findings and discussion thereof. Paragraph 6.1. answers the research question. Paragraph 6.2. discusses the results and limitations of this paper. Paragraph 6.3 proposes research directions for future work and recommendations.

6.1 Conclusion

This thesis researched the use of aspect-based sentiment analysis on customer service calls for product and service improvements. The research question central to this work is:

How can opportunities for product and service improvements be identified from customer service calls using a data-driven approach?

This question has been researched by means of five sub-questions discussed in various chapters of this work. In Chapter 2 it is discussed why extracting sentiment from customer service calls is valuable for a firm's marketing. Previous research has shown that sentiment is strongly connected to customer satisfaction, and that inbound customer call text contains unstrained real-time customer needs. A data-driven approach is found useful to identify customer needs on large-scale data. Sentiment may be extracted from customer service calls using various methods. In this thesis, sentiment is extracted with a state-of-the-art method: Bidirectional Encoder Representations for Transformers (BERT).

Chapter 3 discusses how *business-relevant* aspects can be found, an element that is found to be neglected in almost all research on Aspect-based Sentiment Analysis, which is still for the overwhelming majority focused on technical model performance and not extraction of relevant insights. A three-stage strategy was used to create a relevant, concise list of aspects. The three stages are: using a domain expert's lexicon, doing a customer-oriented search and looking at word frequency. It has found that the choice of aspects is essential to train a model that delivers opportunities to improve products and services.

Chapter 3 outlines how sentiment in customer text can be identified and linked to the relevant aspects. In this thesis a *multi-task model for Local Context Focus Aspect Term Extraction and Aspect Sentiment Classification BERT (LCF-ATESC)* is used for Aspect- based Sentiment Analysis. The model architecture is originating from Yang et al. (2021) and is capable of both identifying aspects and associated sentiment within a unified framework. The pretrained BERT model that worked best was found to be BERTje (De Vries et al, 2019).

In Chapter 4 an overview is given of the used customer service call data from a large telecom company in the Netherlands. Sentiment research on speech-originated texts was previously scarce or even non-existing, more so for the Dutch language.

Chapter 5 discusses model fine-tuning. Fine-tuning was found to be of great importance in terms of model performance. The model scores especially well for aspect term extraction (ATE). It scores consistently lower for the aspect-sentiment classification (ASC) task. Possibly this is due to the fact that this task is more complex than ATE and, in addition, the aspect-sentiment class imbalance makes drawing a stable conclusion about ASC difficult.

The model predictions are as well interpreted in terms of business usefulness. Sentiment between aspects can be visualized in-text for individual sentences, and large-scale aggregated in graphs. Chapter 5 shows some visualizations that show sentiment differences across aspects and aspect-sentiment (changes) tracked over time. It is found that the model is potentially able to generate managerial insights from service calls that were previously not possible, at least not on such a large scale. Although, likely appropriate filtering is needed to give specific recommendations for product or service improvements.

In conclusion, this research has shown that by building an end-to-end pipeline for (i) finding business-relevant aspects for KPN Mobiel products, (ii) extracting these aspects from customer service calls, (iii) extracting associated aspect-sentiment and (iv) aggregating the aspect-sentiment pairs in an appropriate user interface, valuable and customizable insights related to product and service improvements for the firm can be found.

6.2 Discussion and limitations

In this paragraph the results from this work are discussed from a technical, academic and managerial angle. First, a limitation from all three perspectives is that the model depends on the quality of labeling, this is an important technical and academic limitation. If non-relevant aspects are labeled, the model will learn to extract these. Even more challenge lies with labeling aspect-sentiment which is to an extent subjective, especially with noisy data. A related managerial limitation is that implementation of this model requires uniform labeling guidelines or at least agreements among different annotators. This research used two annotators, hence one annotator may have a strong influence on what the model learned. A second managerial and technical limitation is the quantity of data to be labeled. More labeled data would likely improve the model learning. However, labeling is labour expensive.

A third limitation, from an academic perspective is the volatile ASC F1 score. This makes it hard to draw a stable conclusion about the optimal fine-tuned model. It is hypothesized that this volatility is caused by too infrequent occurrence of positive aspect-sentiments in the training and validation data. Suggestions to combat this issue are made in Chapter 5. Another possible explanation is that a bug in the ASC training design has caused this issue.

A fourth limitation is that the multi-task model has unbalanced scoring for its ATE and ASC task respectively, as shown in Chapter 5. The first explanation might be that ASC inherently is a more complex task. Even for the human annotators the aspect-polarity choice of aspects is not always trivial, especially because of the noisy data. Noisy data is in contrast not an issue in annotating aspects. The second explanation brings up the (academic) discussion between model performance for multi-task or single-task models. Research suggests that joint learning tasks might increase performance for both tasks, because learned features from one task can be transferred to the other, and so over iterations it can help determine the optimal weights. It has been found that tasks might cooperate but also compete (Standley et al, 2020). This research found indications that there is a trade-off between ATE and ASC performance, especially for top performing models in one of the tasks. One explanation for this phenomenon might be that the tasks differ too much from each other. More research would be required to indicate whether this is true for ABSA models in general.

The data quality brings up a fifth limitation from a managerial and technical perspective. The phone call noise, the text depersonalization (anonymization) algorithm, the speech-to-text algorithm and lack of interpunction (within the turns) decrease the data quality. This results in more neutral labels in annotation and difficult prediction circumstances for the model.

Finally, from a technical perspective limitations can be made on the model training to improve model performance. Such as tuning the seed and batch size parameters from the beginning. And implementing an early stopping mechanism in the training design.

Finally, an academic limitation is that the number of annotated examples per aspect for each sentiment class are unknown. An error-analysis of the model would gather information on the performance for extraction of every single aspect and for every associated sentiment for that aspect. This information can be used to strategically add more labeled data.

6.3 Future research and work

This paragraph will discuss future directions for research and development from an academic and managerial angle. First, for this research the training, validation and test data are all on the same distribution of data distribution. From an academic perspective it is valuable to research the generalizability of the model by testing it on external out-of-domain data with a different distribution; how well does the model perform on Dutch ABSA benchmark data? If the model performs decently on external data, then it is more robust and its performance over (three years) time or on similar data is expected to be better as well. Related to this from a managerial perspective it would be interesting to apply the model on customer feedback data from questionnaires or customer forums. This data differs from service calls, it is written and more to the point. Within KPN, business stakeholders have actually requested to apply the model to this feedback data.

A third suggestion for future research is model explainability. How does the model make predictions and what (context) words does the model base its predictions on? These types of findings can be used to improve the model and overcome biases or undesired results. From a managerial perspective, the model should be embedded in an AI governance framework.

A fourth direction for future research is to conduct a type of event study by relating aspect-sentiment information to marketing or operational changes of the firm (KPN) or of competitors. From an academic perspective this would show the external validity of the model. From a managerial perspective, KPN could use these outcomes for marketing purposes. After having the required opt-in from customers conform GDPR. In a meeting with business stakeholders, certain aspect-sentiment changes were actually already linked to real service changes within the company. Due to restrictions of confidentiality of company information these types of analyses however, are not pursued in this work.

A fifth direction for future research lies creating modules to make the model more business relevant. First, if the order of the turns are kept or chained per conversation, sentiment changes could be tracked over the call. A call might start off negative, however during the call the agent is able to change the aspect-sentiment to neutral or even positive. This could generate useful insights into service delivery and customer satisfaction. Second, a sentiment search tool module could be integrated in the pipeline. A search tool could enable end-users to dive deeper in the results and read exactly what customers say on the matter. Third, an active learning (feedback) tool can be designed which allows in-text predictions to be manually evaluated by end-users who can confirm or correct the predictions. These corrections can then be fed (periodically) to the model to improve its performance.

Sixhtly, there is ample room for future work concerning the aspects. The list of aspects is limited intentionally, every extra aspect required more labeling. The list of aspects can be enlarged. First, by broadening the aspect-product category from KPN Mobiel to for example KPN Internet at home or KPN TV. Second, more specific aspects can be listed. To overcome the limitation of too generic aspects to make service improvements. Third, aspects can also be listed as proxies for customer issues or service actions. Currently, aspects are mainly as attributes of a product category (KPN Mobiel). It could even be possible to not use single words as aspects but for example triples and find their associated (total) sentiment. Triples consist of a subject predicate and object format, e.g. <customer, buys, iphone>.

In conclusion, the currently planned work on the model within KPN is summarized. The model currently functions as a prototype. The created dashboard with predictions will be published as an information source. Both the customer feedback team and customer contact team expressed interest in further developing the model. The built infrastructure will be used to expand, evaluate and improve on. Plans are made to run the model written customer feedback data. My given recommendations for development are that domain experts should be closely involved in model development. Collaboration of the data scientist and domain experts ensures the best actionable insights regarding product and service improvements.

7. References

Affane, R. (2020a), Decoding NLP Attention Mechanisms; Towards Transformers Overview and Intuition (Blog post), *Data from the Trenches Series january 2020*, from: https://medium.com/data-from-the-trenches/decoding-nlp-attention-mechanisms-38f108929ab7

Affane, R. (2020b), Dissecting the Transformer; A thorough review of the transformer inner mechanisms (Blog post), *Data from the Trenches Series february 2020*, from: https://medium.com/data-from-the-trenches/the-transformer-the-model-c921e43574e3

Akhtar, M. S., Garg, T., & Ekbal, A. (2020). Multi-task learning for aspect term extraction and aspect sentiment classification. *Neurocomputing*, *398*, 247-256.

Alammar, J. (2018), 'Illustrated BERT, ELMo & co; How NLP cracked transfer learning' (Blog post), December 3th 2018, from: https://jalammar.github.io/illustrated-bert/

Alammar, J. (2018), 'Illustrated Transformer' (Blog post), June 27th 2018, from: https://jalammar.github.io/illustrated-transformer/

Anderson, E. W., & Sullivan, M. W. (1993). The antecedents and consequences of customer satisfaction for firms. *Marketing science, 12*(2), 125-143.

Anderson, E. W., Fornell, C., & Lehmann, D. R. (1994). Customer satisfaction, market share, and profitability: Findings from Sweden. *Journal of marketing, 58*(3), 53-66.

Anderson, E. W., Fornell, C., & Mazvancheryl, S. K. (2004). Customer satisfaction and shareholder value. *Journal of marketing*, *68*(4), 172-185.

Ask, K., & Landström, S. (2010). Why emotions matter: Expectancy violation and affective response mediate the emotional victim effect. *Law and human behavior, 34*(5), 392-401.

Bagozzi, R. P., Gopinath, M., & Nyer, P. U. (1999). The role of emotions in marketing. *Journal of the academy of marketing science, 27*(2), 184-206.

Balahur, A., Hermida, J. M., & Montoyo, A. (2012). Detecting implicit expressions of emotion in text: A comparative analysis. *Decision support systems, 53*(4), 742-753.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing, 84*(1), 1-25.

Bigorra, A. M., Isaksson, O., & Karlberg, M. (2020). Semi-autonomous methodology to validate and update customer needs database through text data analytics. *International Journal of Information Management, 52*, 102073.

Boulding, W., Kalra, A., Staelin, R., & Zeithaml, V. A. (1993). A dynamic process model of service quality: from expectations to behavioral intentions. *Journal of marketing research, 30*(1), 7-27.

Brauwers, G., & Frasincar, F. (2022). A survey on aspect-based sentiment classification. *ACM Computing Surveys*, *55*(4), 1-37.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science, 35*(6), 953-975.

Chen, X., & Wan, X. (2022). A Simple Information-Based Approach to Unsupervised Domain-Adaptive Aspect-Based Sentiment Analysis. *arXiv preprint arXiv:2201.12549*.

Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2022). Blame the Bot: Anthropomorphism and Anger in Customer–Chatbot Interactions. *Journal of Marketing, 86*(1), 132-148.

Cronin Jr, J. J., Brady, M. K., & Hult, G. T. M. (2000). Assessing the effects of quality, value, and customer satisfaction on consumer behavioral intentions in service environments. *Journal of retailing, 76*(2), 193-218.

De Clercq, O., & Hoste, V. (2016). Rude waiter but mouthwatering pastries! An exploratory study into Dutch Aspect-Based Sentiment Analysis. *In 10th International Conference on Language Resources and Evaluation (LREC)* (pp. 2910-2917), ELRA.

De Clercq, O., Lefever, E., Jacobs, G., Carpels, T., & Hoste, V. (2017, September). Towards an integrated pipeline for aspect-based sentiment analysis in various domains. *In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 136-142).

Devlin, J. & Chang, M., 'Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing', *Google AI blog* Nov 2th 2019, https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

J. Devlin, BERT Multilingual, from: https://github.com/google-research/bert/blob/master/multilingual.md

Ding, J., Sun, H., Wang, X., & Liu, X. (2018, June). Entity-level sentiment analysis of issue comments. In *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering* (pp. 7-13).

Du, J., Fan, X., & Feng, T. (2011). Multiple emotional contagions in service encounters. *Journal of the Academy of Marketing Science, 39*(3), 449-466.

Falk, T., Hammerschmidt, M., & Schepers, J. J. (2010). The service quality-satisfaction link revisited: exploring asymmetries and dynamics. *Journal of the Academy of Marketing Science, 38*(3), 288-302.

Golder, P. N., Mitra, D., & Moorman, C. (2012). What is quality? An integrative framework of processes and states. *Journal of marketing, 76*(4), 1-23.

Grandey, A. A., Dickter, D. N., & Sin, H. P. (2004). The customer is not always right: Customer aggression and emotion regulation of service employees. Journal of Organizational Behavior: *The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 25(*3), 397-418.

Griffin, A., & Hauser, J. R. (1992). Patterns of communication among marketing, engineering and manufacturing—A comparison between two new product teams. *Management science, 38*(3), 360-373.

Griffin, A., & Hauser, J. R. (1993). The voice of the customer. *Marketing science, 12*(1), 1-27.

Gustafsson, A., Johnson, M. D., & Roos, I. (2005). The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *Journal of marketing, 69*(4), 210-218.

Hartline, M. D., & Ferrell, O. C. (1996). The management of customer-contact service employees: An empirical investigation. *Journal of marketing, 60*(4), 52-70.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654.*

Jiao, J., & Chen, C. H. (2006). Customer requirement management in product development: a review of research issues. *Concurrent Engineering, 14*(3), 173-185.

Kamil Topal and Gultekin Ozsoyoglu. 2016. Movie review analysis: Emotion analysis of IMDb movie reviews. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016)*, 1170–1176.

Kravets, A. (2021). A Deep Dive into the code of the BERT model, Towards Data Science (blog), from: https://towardsdatascience.com/deep-dive-into-the-code-of-bert-model-9f618472353e

Lee, Thomas Y., and Eric T. Bradlow (2011), “Automated Marketing Research Using Online Customer Reviews,” *Journal of Marketing Research, 48*(5), 881–94.

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50-70.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025.*

Mattila, A. S., & Enz, C. A. (2002). The role of emotions in service encounters. *Journal of Service Research, 4*(4), 268-277.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal, 5*(4), 1093-1113.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3-26.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science, 31* (3), 521–43

Nyer, P. U. (1997). A study of the relationships between cognitive appraisals and consumption emotions. *Journal of the Academy of Marketing Science, 25*(4), 296-304.

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of marketing research*, *17*(4), 460-469.

Oliver, R. L., & Swan, J. E. (1989). Equity and disconfirmation perceptions as influences on merchant and product satisfaction. *Journal of consumer research, 16*(3), 372-383.

Palese, B., & Usai, A. (2018). The relative importance of service quality dimensions in E-commerce experiences. *International Journal of Information Management, 40*, 132-140.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval, 2*(1–2), 1-135.

Parasuraman, A., Zeithaml, V. A., & Berry, L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 16*(1), 12-37.

Phan, M. H., & Ogunbona, P. O. (2020, July). Modelling context and syntactical features for aspect-based sentiment analysis. *In Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3211-3220).

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT?. *arXiv preprint arXiv:1906.01502.*

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jimenez, Zafra, S.M., Eryigit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In: *SemEval-2016. ACL, San Diego, California*, pp. 19–30.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I (2015). SemEval-2015 task 12: Aspect based sentiment analysis. In: *SemEval 2015. ACL, Denver, Colorado*, pp. 486–495.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In: *SemEval 2014. ACL, Dublin, Ireland,* pp. 27–35.

Ramponi, A., & Plank, B. (2020). Neural Unsupervised Domain Adaptation in NLP---A Survey. *arXiv preprint arXiv:2006.00632*.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics, 8*, 842-866.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. *Review of personality & social psychology, 5,* 11-36*.*

Schouten, K., & Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *28*(3), 813-830.

Shanahan, L., Steinhoff, A., Bechtiger, L., Murray, A. L., Nivette, A., Hepp, U., ... & Eisner, M. (2020). Emotional distress in young adults during the COVID-19 pandemic: evidence of risk and resilience from a longitudinal cohort study. *Psychological medicine, 52*(5), 824-833.

Smith, A. K., & Bolton, R. N. (2002). The effect of customers' emotional responses to service failures on their recovery effort evaluations and satisfaction judgments. *Journal of the academy of marketing science, 30*(1), 5-23.

Smith, L. E., Duffy, B., Moxham-Hall, V., Strang, L., Wessely, S., & Rubin, G. J. (2021). Anger and confrontation during the COVID-19 pandemic: a national cross-sectional survey in the UK. *Journal of the Royal Society of Medicine, 114*(2), 77-90.

Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., & Savarese, S. (2020). Which tasks should be learned together in multi-task learning?. *Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119,* 9120-9132.

Suciati A. and Budi, I. (2020). Aspect-based sentiment analysis and emotion detection for code-mixed review. *International Journal of Advanced Computer Science and Applications 11*(9), 179–186.

Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science, 38*(1), 1-20.

Uszkoreit J., Transformer: a novel neural network architecture for language understanding, *Google AI Blog* 31 August 2017, from: https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Villarroel Ordenes, F., Ludwig, S., De Ruyter, K., Grewal, D., & Wetzels, M. (2017). Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media. *Journal of Consumer Research, 43*(6), 875-894.

Wang, Z., Ho, S.B., and Cambria, E.. 2020. Multi-level fine-scaled sentiment sensing with ambivalence handling. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 28*(4), 683.

Weng, L. (2018), Attention? Attention!, *lilianweng.github.io* (Affl.: OpenAI) (Blogpost), 24 june 2018.

Xu, X., Wang, X., Li, Y., & Haghighi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors. *International Journal of information management, 37*(6), 673-683.

Yang, H., Zeng, B., Yang, J., Song, Y., & Xu, R. (2021). A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. *Neurocomputing, 419*, 344-356.

Yang, H. (2022). PyABSA - Open Framework for Aspect-based Sentiment Analysis [Computer software, v1.16.27]. From: https://github.com/yangheng95/PyABSA/tree/release

Zeelenberg, M., & Pieters, R. (2004). Beyond valence in customer dissatisfaction: A review and new findings on behavioral responses to regret and disappointment in failed services. *Journal of business Research, 57*(4), 445-455.

Zeng, B., Yang, H., Xu, R., Zhou, W., & Han, X. (2019). Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences, 9*(16), 3389.

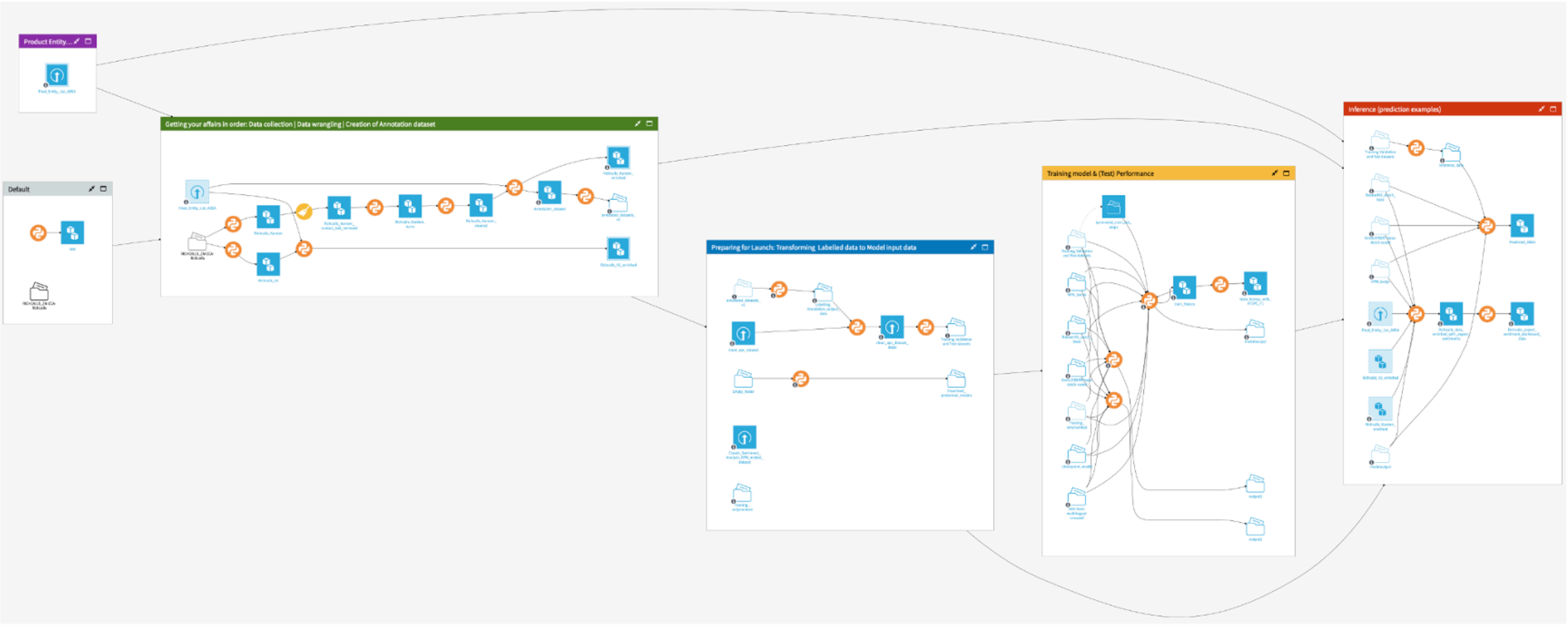
Zhao, T., Du, J., Xu, Z., Li, A., & Guan, Z. (2022). Aspect-Based Sentiment Analysis using Local Context Focus Mechanism with DeBERTa. *arXiv preprint arXiv:2207.02424*.

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2022). A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *arXiv preprint arXiv:2203.01054*.

8. Appendix

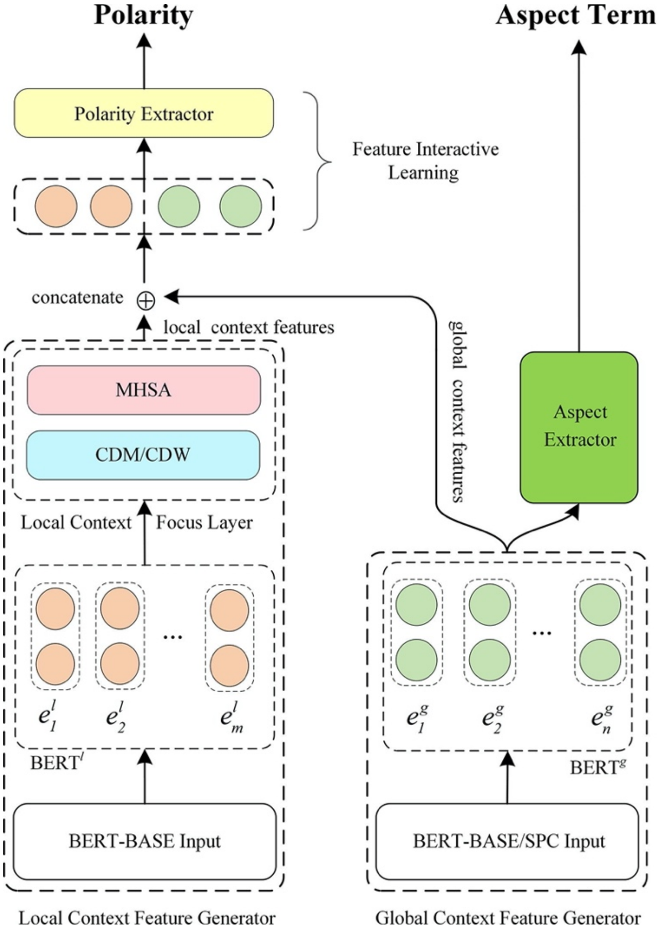
Appendix 1

Appendix 1: *The end-to-end model pipeline visualized in building blocks*


Product Entity...
Default
Getting your affairs in order: Data collection | Data wrangling | Creation of Annotation dataset
Preparing for Launch: Transforming Labelled data to Model input data
Training model & (Test) Performance
Inference (prediction examples)

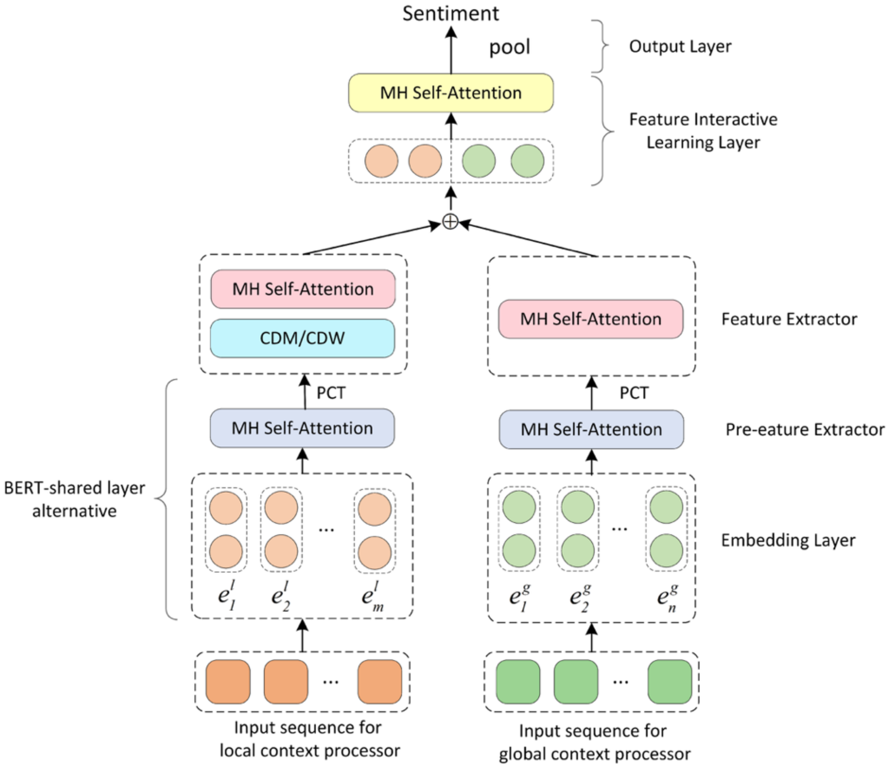
The first (purple) phase includes the final list of aspects. The second (green) phase entails data collection and cleaning. The third (blue) phase incorporates processing labeling outcomes towards model inputs. The fourth (orange) phase encompasses the model training and evaluation. The fifth (red) phase represents the model prediction and visualization of results.

Appendix 2: *Network architecture LCF-ATEPC-BERT* adapted from Yang et al. (2021)*



**Multi-task model for Local Context Focus Aspect Term Extraction and Aspect Sentiment Classification BERT*

Appendix 3: *Local Context Focus features generator architecture adapted from Zeng et al. (2019)*



***Appendix 4**: All hyperparameters, their function and their explored range*

#	Hyperparameter	Function	Range
1	Pretrained BERT	Which pretrained model is used?	mBERT, BERTje, KPN BERTje, RobBERTv2
2	Model architecture	Which model is used (model architecture)?	fast_lcf(s)_atepc | lcf(s)_atepc
3	SRD (threshold)	How large is the local context focus (SRD) window surrounding the aspect	1 - 8
4	LCF mechanism	Type of local context focus mechanism. (Masking, Weighting of both)	CDM, CDW or Fusion
5	Learning rate	How quickly the model converges (learns)	1e-05 - 1e-04
6	Number of Epochs	Number of epochs model runs on.	3, 4, 8, 10, 30
7	Dropout	Probability of deactivating certain neurons.	0.0 - 0.5
8	Seed	Seed to initialize all 'randomness' in the model.	52, 1, 193 …
9	Batch size	How many input sentences does the model accept per batch?	4, 8, 16, 32
10	Dynamic Truncation	Applying an aspect-centeried truncation instead of head truncation?	True | False
11	Max. Sequence length	Max. input length, longer text is truncated.	100 | 70
12	BERT SPC	Regards to the ABSA input format (False best)	True | False
13	Syntax-based SRD	Use syntax-based SRD (no effect found)	True | False
14	SRD alignment	Tries to align the nodes of syntax tree and tokenization (transformers) (no effect found)	True | False
15	L2 regularization	L2 Regularization parameter (no effect found)	1e-05 - 1e-04
16	Optimizer	'Adamw' algorithm (industry standard)	Not explored

Appendix 5: *Quantiles table for the F1 Scores for different pretrained models*

	All	**KPN BERTje**	**BERTje**	**RobBERTv2**
N	742	316	261	164
Aspect Sentiment Classification F1 Score				
1%	29.9	29.9	261	19.0
5%	29.9	29.9	29.9	29.9
25%	39.7	51.1	48.9	29.9
50%	52.3	55.0	54.2	33.0
75%	57.3	58.4	57.8	40.7
95%	61.8	62.7	60.5	59.4
99%	64.8	65.6	64.2	64.1
Aspect Target Extraction F1 Score				
1%	0	0	0	0
5%	0	62.8	67.1	0
25%	71.37	68.7	85.7	1.4
50%	85.66	75.5	88.7	96.6
75%	90.47	85.1	93.3	96.7
95%	96.7	89.8	95.7	96.7
99%	96.7	90.8	96.3	96.7

Appendix 6

Appendix 6: *20 best performing fine-tuned models by their summed metric*

Session	ASC F1	ATE F1	ASC Acc	ASC ASC ATE	Model	BERT	LCF type	SRD	Epochs	Dropout	Learning rate	L2reg	Seed	Batch size
A600	62,16	89,19	83,37	234,72	fast_lcfs_atepc	BERTje	fusion	5	4	0,2	2E-05	1E-05	52	4
A619	60,82	90,28	83,17	234,27	fast_lcfs_atepc	BERTje	fusion	5	4	0,2	2E-05	1E-05	1	2
A621	58,2	92,52	83,37	234,09	fast_lcfs_atepc	BERTje	fusion	5	4	0,2	2E-05	1E-05	1	4
A613	56,7	94,13	83,17	234	fast_lcfs_atepc	BERTje	fusion	5	4	0,2	2E-05	1E-05	1926	8
A612	56,39	94,1	82,41	232,9	fast_lcfs_atepc	BERTje	fusion	5	4	0,2	2E-05	1E-05	1926	8
H7	57,42	91,67	83,17	232,26	fast_lcfs_atepc	BERTje	fusion	5	4	0,2	2E-05	1E-05	200	4
A598	58,56	90,12	83,17	231,85	fast_lcfs_atepc	BERTje	fusion	5	4	0,2	2E-05	1E-05	52	2
A684	62,33	86,25	82,6	231,18	fast_lcfs_atepc	BERTje	fusion	5	4	0,4	1E-05	1E-05	52	4
F1	62,33	86,25	82,6	231,18	fast_lcfs_atepc	BERTje	fusion	5	4	0,4	1E-05	1E-05	52	4
A618	57,46	89,52	83,56	230,54	fast_lcfs_atepc	BERTje	fusion	5	4	0,2	2E-05	1E-05	1	2
A648	55,33	96,59	78,59	230,51	fast_lcfs_atepc	RobbertV2	fusion	3	10	0,5	3E-05	1E-05	193	4
A644	55,23	96,7	78,39	230,32	fast_lcfs_atepc	RobbertV2	fusion	3	10	0	3E-05	1E-05	193	4
A583	61,73	83,4	85,09	230,22	fast_lcfs_atepc	KPN_BERTje	fusion	1	10	0,25	1E-05	1E-05	52	4
A649	53,69	96,7	79,54	229,93	fast_lcfs_atepc	RobbertV2	fusion	4	10	0,5	3E-05	1E-05	193	4
F4	59,09	87,8	82,98	229,87	fast_lcfs_atepc	BERTje	fusion	5	4	0,4	1E-05	1E-05	1926	4
A610	54,92	91,11	83,75	229,78	fast_lcfs_atepc	BERTje	fusion	5	4	0,2	2E-05	1E-05	1926	4
A674	64,76	80,56	84,32	229,64	fast_lcfs_atepc	BERTje	fusion	3	4	0,4	1E-05	1E-05	1	4
A615	53,54	94,17	81,84	229,55	fast_lcfs_atepc	BERTje	fusion	5	4	0,2	2E-05	1E-05	1926	16
A680	62,33	84,15	82,98	229,46	fast_lcfs_atepc	BERTje	fusion	5	4	0,1	1E-05	1E-05	52	4
A661	52,86	96,7	79,54	229,1	fast_lcfs_atepc	RobbertV2	fusion	3	10	0,5	3E-05	1E-05	52	4

Appendix 7: *Average Test performance of the final model on multiple (10) seeds*

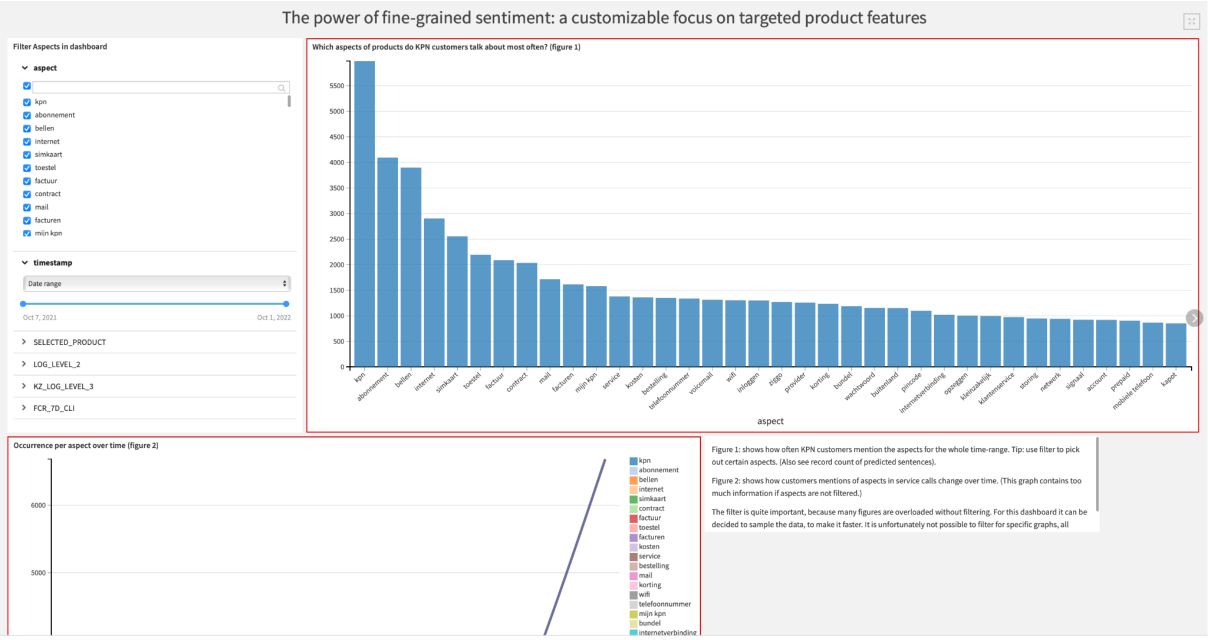
Metric	ASC Accuracy (%)	ASC F1 (%)	ATE F1 (%)
Average	82.24	59.6	79.16*
Standard	1.53	2.05	17.64*
Range	78.78 - 83.56	54.45 - 62.33	67.13 - 88.19
All seeds	[82.6, 82.41, 83.37, 82.98, 78.78, 81.07, 83.56, 83.17]	[62.33, 54.45, 61.84, 59.09, 59.14, 60.5, 59.53, 59.89]	[86.25, 67.13, 81.6, 87.8, 88.19, 69.64, 83.63, 69.13]

**incl. 3 bad seeds*

Appendix 8: *Distribution of labeled Sentiment classes over the datasets*

Set	Positive	Neutral	Negative
Train	134	3403	623
Validation	18	425	79
Test	16	426	77
Relative distribution	3%	81%	15%

Appendix 9: *A sample of a possible dashboard interface for visualizing aspect-sentiment insights for business stakeholders*



The power of fine-grained sentiment: a customizable focus on targeted product features

Filter Aspects in dashboard

Which aspects of products do KPN customers talk about most often? (figure 1)

Occurrence per aspect over time (figure 2)

Figure 1: shows how often KPN customers mention the aspects for the whole time-range. Tip: use filter to pick out certain aspects. (Also see record count of predicted sentences).

Figure 2: shows how customers mentions of aspects in service calls change over time. (This graph contains too much information if aspects are not filtered.)

The filter is quite important, because many figures are overloaded without filtering. For this dashboard it can be decided to sample the data, to make it faster. It is unfortunately not possible to filter for specific graphs, all