Master's Thesis:

Improving Sentiment Analysis Performance Using Speaker Diarization Classification

by

Geert Telkamp

Data Science: Business and Governance

College year 2018-2019

In collaboration with Totta Research

Supervisor: Menno van Zaanen

TABLE OF CONTENTS

1. INTRODUCTION

Sentiment analysis, the computational study of people's opinions, attitudes and emotions (Medhat, Hassan, & Korashy, 2014), received an increasing amount of attention from both scientific and commercial fields in recent years. It has been found to be a valuable application in various domains such as customer relations, political elections and financial markets (Feldman, 2013).

One core application in the domain of customer relations is the analysis of customer service phone calls. Sentiment analysis is used by companies to efficiently distinguish the informative, high sentiment calls from the other, less informative calls. However, performance of these models or sentiment analysis models in general is often found to be suboptimal (Collomb, et al., 2014). The following sections will address various challenges in performing sentiment analysis on customer service calls and propose a new method, speaker diarization classification, as a solution to these challenges.

1.1 PROBLEM FORMULATION

Although customer service call centers' foremost purpose is to facilitate one-on-one direct communication between customer and company, it serves a second purpose: informing the company about general opinion and sentiment of their customer base and target group. However, for companies that receive large amounts of calls on a daily basis, manual analysis of all calls will be too time consuming. An automated process is needed to select calls that are likely to be informative, allowing for more efficient analysis of customer service calls.

This is where sentiment analysis comes into play. It is able (or aims) to automatically detect the sentiment of an audio or text file, so that cases containing positive or negative sentiment (i.e. cases that are assumed to be informative) can be identified easily. However, systems performing sentiment analysis are known to "exhibit limitations in terms of meeting robustness, accuracy, and overall performance requirements, which, in turn, greatly restrict the usefulness of such systems in practical real-world applications" (Poria, Cambria, Bajpai, & Hussain, 2017). It is of practical importance to develop methods that overcome these performance limitations, as this will enable the companies to find the most informative cases more effectively.

Although many challenges exist for sentiment analysis in general (an overview of these challenges will be provided in the chapter "related work"), for bilateral communications such as customer service calls an additional challenge exists. These calls contain two speakers who can each have their own sentiments. Moreover, these sentiments are not expressed in a structured manner: at any time, each speaker can express their own sentiment, as well as respond to the other speaker's sentiment. Possibly, speakers might even

(partially) match their sentiments. If the objective of analysis is to find the sentiment of just one speaker, truly informative data is intertwined with possibly misinformative data (i.e. data related to the sentiment of the other speaker).

Customer service calls are real-world cases in which the objective of analysis is to find the sentiment of just one speaker, namely the customer. Yet common practice is to treat these calls as a whole, including both the consumer's and the agent's contribution for analysis. If the agent's contribution is indeed more informative to their own sentiment rather than the customer's, the agent's contribution could be considered noise in the data and including it might cause performance to be reduced. Canceling out the irrelevant speaker "purifies" the data, making it easier for the sentiment analysis algorithms to detect true sentiment of the individual speaker.

To capture the unique contribution of each speaker in a bilateral conversation, speaker diarization, the process of distinguishing several speakers within a single audio or text file, is needed. In turn, diarization classification is the process of distinguishing speakers and subsequently classifying the role of each speaker in the conversation. The current research explores the effect of diarization classification on sentiment analysis performance of bilateral conversations. Specifically, it will aim to improve sentiment analysis on customer service calls using diarization classification.

1.2 RESEARCH QUESTIONS

The first objective of this project is to improve sentiment analysis performance by isolating the customer's contribution of a customer service call using diarization classification. Formally, the research question of this paper will be:

Research question 1: Can sentiment analysis performance of customer service calls be improved by isolating the customer's contribution, using diarization classification?

A "common practice" baseline will be set up as a starting point, in which sentiment analysis will be performed without the use of diarization classification and thus the conversations are treated as a whole. Against this baseline will be tested:

H1: Isolating the customer's contribution in customer service calls will increase performance of sentiment analysis.

This hypothesis will be tested using audio and text data separately, as well as a combination of both (i.e. multimodal sentiment analysis), to explore the most predictive combination of features for sentiment analysis. Therefore, the hypothesis can be split into three separate sub-hypotheses:

H1a: Isolating the customer's contribution in customer service calls will increase performance of text sentiment analysis.

H1b: Isolating the customer's contribution in customer service calls will increase performance of audio sentiment analysis.

H1c: Isolating the customer's contribution in customer service calls will increase performance of multimodal sentiment analysis.

In these first hypotheses, the agent's contribution in the call is treated as noise in the data in measuring the consumer's sentiment and is therefore excluded from analysis. However, it is likely that the agent will respond to the customer's sentiment (e.g. "*I understand that you are angry…*") and thus their contribution might contain valuable information. Call center agents are trained to acknowledge consumer's emotions, while expressing little emotion themselves. It is theorized that, for this reason, *what* they say could be relevant for detecting the consumer's sentiment, but *how* they say it is not. In other words, their textual data might inform us, but their audio data will likely not. Therefore, the second research question of this project will be:

Research question 2: Can the best performing set of features extracted from the isolated customer's contribution (as found in testing hypothesis 1) be further improved by adding textual features extracted from the isolated agent's contribution?

In answering this second research question, the following hypothesis has been formulated:

H2: Combining the best performing set of features extracted from the isolated customer's contribution with textual features extracted from the isolated agent's contribution, will increase sentiment analysis performance.

This hypothesis will be tested by taking the best performing set of features in testing hypothesis 1 (either text, audio or multimodal, depending on which one has been found to yield the highest performance) and

combining it with textual features extracted from the agent's contribution. Given results from van Geffen (2018) on a similar dataset, it is expected that using only audio features will yield the highest performance. If this is indeed the case, hypothesis 2 will be tested by combining the isolated customer's audio features with isolated agent's text features and comparing it to the isolated customer's features alone. However, if hypothesis 1 results will indicate differently, the approach will be adapted accordingly (i.e. textual features will be used if they were found to yield the highest performance in testing hypothesis 1, or a combination of textual and audio features will be used if this combination was found to yield the highest performance in testing hypothesis 1).

2. RELATED WORK

The following chapter provides a thorough description of concepts related to the current research. First, a general review of previous work on sentiment analysis will be provided. Next, the scope will be narrowed down to sentiment analysis using text data, audio data, and multimodal data respectively. The subsequent section will elaborate on the topic of speaker diarization, to conclude with a section on earlier work on the dataset that was used in the current research.

2.1 SENTIMENT ANALYSIS

Yadollahi, Shahraki, & Zaiane (2017, page 1) define sentiment analysis as all areas in the field of affective computing related to "detecting, analyzing, and evaluating humans' state of mind towards different events, issues, services, or any other interest." Following this definition, sentiment analysis can be regarded as an umbrella term for various forms of analysis. As is shown in *Figure 1*, the field of sentiment analysis can be divided into two subcategories: opinion mining and emotion mining. In turn, these subcategories can contain various specific forms on classification and detection. With this taxonomy of sentiment analysis, Yadolla, et al. attempt to define what it exactly is and how its terms relate to one another. However, as is noted by Munezero, Montero, Sutinen, & Pajunen (2014) scientific literature lacks definitions that provide clear distinction between subjectivity terms such as sentiment, opinion, emotion, affect, and feeling. Although psychologist have made numerous attempts to provide such definitions, they often end up with extremely complicated formulations that in the end still lack distinctive power.

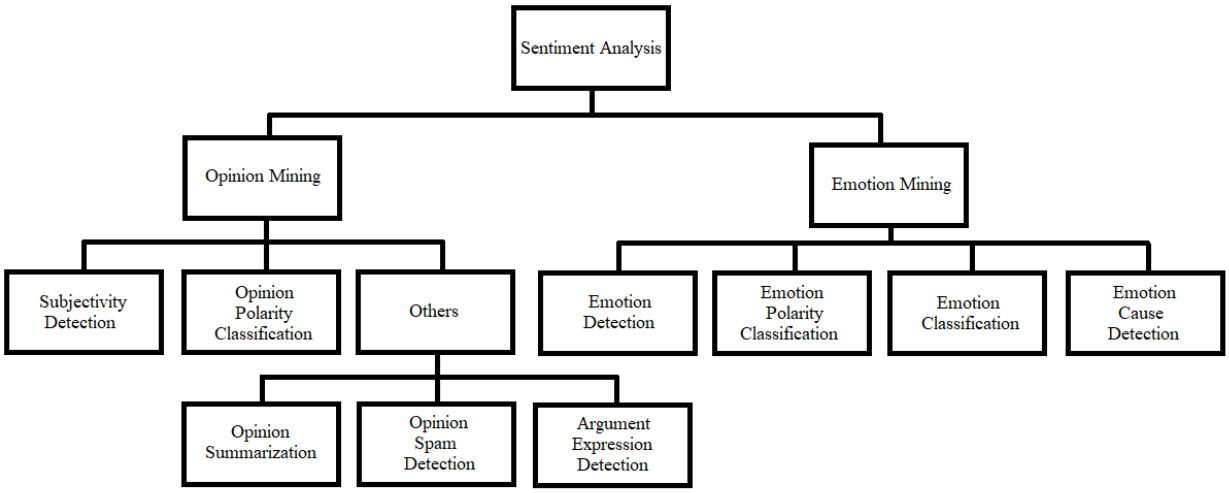

Figure 1: Sentiment analysis as an umbrella term (Yadollahi, Shahraki, & Zaiane, 2017).

Because this lack of clear definitions of sentiment analysis and related terms, both the scientific and the commercial field tend to use these terms interchangeably. For instance, Pang, & Lee (2008) claim the terms "sentiment analysis" and "opinion mining" refer to the same thing and are both a subcategory of "subjectivity analysis". In contrast, Cambria, Schuller, Xia, & Havasi (2013) claim that the two concepts are actually different because "opinion mining" focuses on polarity classification whereas "sentiment analysis" focuses on emotion recognition. However, they do not elaborate on in what aspects these tasks differ, again failing to clearly distinguish them. These different definitions and taxonomies indicate the scientific field lacks consensus.

Similarly, in the commercial field, what is called sentiment analysis often refers to the more specific task of emotion polarity classification (EPC). In EPC, classification is not based on clearly defined emotions, but rather on their polarity (positive/neutral/negative). For most practical applications, it is enough to know the polarity of a given document, rather than the specific emotion and because sentiment analysis in general is not an easy task, this simplified approach is often preferred. The current research will treat sentiment analysis as an EPC task.

Scholars have developed various methods of sentiment analysis, ranging from machine-learning approaches (e.g. Naive Bayes, Support Vector machines, Neural Networks) to lexicon-based approaches (e.g. dictionary-based approaches, corpora-based approaches) (Medhat, Hassan, & Korashy, 2014). In the machine-learning approach, there is a set of training records where each record is labeled according to its sentiment. From these training records, features are extracted. What kinds of features are extracted depends on the nature of the data. Sections 2.2 and 2.3 will describe feature extraction methods for both text and audio data respectively.

After extracting these features, each training record is represented as a feature vector which is then used to train a machine learning classifier. The classifier learns the relationship between the labels and the features in order to subsequently classify new cases based on this learned relationship. Lexicon-based approaches on the other hand do not learn from training records, but from a lexicon that provides information. An example of such a lexicon is a dictionary with sentiment scores for each word (e.g. WordStat, Lougghran and McDonald Financial Sentiment Dictionary, SenticNet). The words in the records are thus assigned scores and the scores for each record are then counted. In addition to dictionaries, sentiment tagged corpora (e.g. IMDB movie ratings, Sanders Analytics Sentiment Corpus, Amazon Fine Food Reviews) can provide information, not only regarding the sentiment value of individual words, but also regarding the sentiment value of words within their contexts.

Although these methods have been studied extensively in the last few decades and their applications in the commercial fields are more prevalent every day, sentiment analysis models generally perform suboptimal (Collomb, et al., 2014). One major reason for their inaccurate results is that human language contains various linguistic constructions such as negations, sarcasm, or internal dependencies (e.g. "*that movie was only slightly <u>better</u> than ...*") that make computational analysis hard. The semantics of these constructions cannot be found by just finding the semantics of the individual words, as the true meaning of these constructions is found in the relationship between the individual words. In addition, real-world data containing sentiment is often unstructured and noisy (Douglas-Cowie, et al., 2007), meaning the vast majority of data is neutral in sentiment and when the data does contain sentiment, the expression often lacks proper grammatical structure (Gokulakrishnan, 2012).

Results may vary strongly depending on the data and methods that were used. Collomb, et al. (2014) compare various methods yielding accuracy scores between 70% and 91%. Similarly, a meta-analysis by Govindarajan, & Romina (2013) compared accuracy scores of 10 independent studies, with accuracy scores found to be between 61% and 93%. Likewise, Gonçalves, Araújo, Benevenuto, & Cha (2013) found average accuracy scores between 59% and 89% when comparing 8 different methods. Although these accuracy results bear little meaning without their context (e.g. the nature of the data and the goals of analysis), they all lead the authors to the same conclusion: there are still big steps to be made in gaining more accurate results.

2.2 TEXT-BASED SENTIMENT ANALYSIS

A possible approach to sentiment analysis is to use text data as the object of analysis. Commonly used sources of text data for sentiment analysis are social media, product reviews, and movie ratings. Additionally, audio data such as customer service calls can be transformed into text data by use of speech recognition software such as Nuance Recognizer or Google Cloud Speech API. In text sentiment analysis, elements of the text are represented by numerical statistics. For simpler tasks, the elements to be represented can be all individual words, for instance in a binary or word count representation. A tf*idf representation (Salton, & Buckley, 1988) can be used as a weighting factor to emphasize elements that appear often in some documents, but not often in the entire corpus. If information is to be found in the relationship between words, N-grams (Damashek, 1995) can be used to represent groups of consecutive words as opposed to each feature representing an individual word. If the documents contain sentences with (somewhat) correct grammatical structure, more advanced methods such as part-of-speech tagging (Voutilainen, 2003) and syntactic parsing (Hale, 2001) can be used to capture deeper relationships between words. What elements are represented in what way depends on the nature of the documents and the goals of analysis.

The main reason for the common use of text data for sentiment analysis is that real-world text data is abundantly available. Social media provide scholars with enormous amounts of textual data containing possible expressions of sentiment. New data is created every day without a need for intervention by researchers. It is because of this abundant nature that sentiment analysis on these data provide an alternative that is must less costly, both in time and money, compared to more traditional research methods (e.g. polls, surveys, panels). That is, if sentiment analysis is able to answer the same questions. Traditional research methods often focus on specific topics such as customers' satisfaction with products or brands. Sentiment analysis, however, is not able to actively target calls related to specific topics. It merely finds certain topics, given that conversations about them contain high sentiment expressions. Still, sentiment analysis has been found to be able to yield similar results as traditional research methods, as indicated by O'Connor, Balasubramanyan, Routledge, & Smith, (2010). They compared results from sentiment analysis of Twitter messages to different consumer confidence and political opinion polls. Although some comparisons yielded a correlation as high as 80%, others were as low as 7%. These results indicate that sentiment analysis could function as an alternative to traditional research methods, if implemented correctly.

The use of social media text data has also been found to provide a solution to the labeling problem in supervised machine-learning. Supervised algorithms require class labeled data. Without these labels, the algorithms have no "true" values to learn from. However, annotation of the data by hand can be extremely time consuming. Luckily social media data can provide us indications of the true labels. These indications can then be used to assume the true label of a record, a process called distant supervision. For instance, Go,

Bhayani, & Huang (2009) use emoticons in Twitter data as indicators of labels, also referred to as "noisy labels". Using distant supervision, researchers can obtain large amounts of labeled data with relatively little effort. However, one could question the accuracy of these labels, as they are assumed to be imperfect representations of true labels (hence the term noisy). Luckily, scholars have examined the effect of noise in the labeling of data and found that several methods of classification, such as biased SVM and weighted logistic regression, are provably noise-tolerant (Natarajan, Dhillon, Ravikumar, & Tewari, 2013).

2.3 AUDIO-BASED SENTIMENT ANALYSIS

A much less common approach to sentiment analysis is the use of features extracted from audio. Mairesse, Polifroni, & Di Fabbrizio (2012) attribute this difference to the fact that technologies for text sentiment analysis are much more mature, compared to audio sentiment analysis. In addition, there might be differences in the extent to which data is sensitive to privacy issues. For instance, text data such as reviews can be found fairly easily online, whereas customer service data are often stored on highly secured servers.

The approach to analysis of audio files such as customer service calls is often to use speech-to-text transcription software to subsequently analyze the audio records as if they were text records (e.g. Vaudable, & Devillers, 2012). However, some studies indicate that characteristics of the audio itself might contain information regarding sentiment as well.

Prosody is defined as "suprasegmental information in speech, such as phrasing and stress, which can alter perceived sentence meaning without changing the segmental identity of the components" (Price, Ostendorf, Shattuck-Hufnagel, Fong, 1991, p. 2956). As this definition implies, scholars have studied prosody mainly as a means to spoken language disambiguation. Cutler, Dahan, & Van Donselaar (1997) identified three levels on which prosodic information can be used by humans to get a deeper understanding of spoken language: it informs us in the process of word recognition (e.g. where does a word start or end, in what sense is the word used), it informs us about the syntactic structure of a sentence (e.g. words are pronounced differently depending on the position in the syntactic structure) and the discourse of a spoken language (e.g. highlighting parts that contain salient information throughout the discourse). However, can prosody also be used to inform sentiment analysis? I.e. do statistics on prosody differ for different classes of sentiment?

It seems intuitively understandable that people experiencing high sentiment (e.g. very positive or very negative) use different prosody, like raising or lowering their voice. However, changes in prosody can be similar for different classes of sentiment. For instance, Ververidis, & Kotropoulos (2006) found that pitch and intensity of spoken language increase for both positive and negative sentiment. Mairesse,

Polifroni, & Di Fabbrizio (2012) compared the use of prosody to the use of of text in sentiment analysis performance. They transcribed spoken reviews both by hand and using automatic speech recognition software and compared models using these transcriptions to models using prosody statistics extracted from the audio files. Although they found that the use of prosody outperforms a majority class baseline, it did not outperform a model using only text. These results imply that, although less informative than text, valuable information can be extracted from the analysis of prosody. However, the use of one does not exclude the other.

2.4 MULTIMODAL SENTIMENT ANALYSIS

Mairesse, Polifroni, & Di Fabbrizio (2012) tested whether a combination of text and audio analysis is able to outperform analysis of either text or audio separately. They first created text sentiment model and later added a feature that contained the prediction of this text sentiment model to their prosody models and found that this combination increased performance in three out of four models compared to the models using just prosody. Whether these multimodal models performed better than models using just text depends on the quality of the text. When using text transcribed by hand (which can be seen as a gold standard for transcription), text only models perform best. However, when comparing to models using Automatic Speech Recognition (ASR) to transcribe the speech to text, the multimodal models actually performed better than the text only models. The authors conclude that prosody is able to correct for the noisy transcription by ASR software.

Similarly, various other scholars advocate multimodality in sentiment analysis. Multimodal sentiment analysis is an approach to deeper and more accurate understanding, in which text data is combined with audio or even visual data. Indeed, meta-analysis of multimodal models revealed that 85% of these models perform more accurately than their unimodal counterparts with an average improvement of 9.83% (D'mello, & Kory, 2015). The multimodality of audio and text data accounted for 11.1% of all models used in the meta-analysis.

The authors hypothesize multimodality has three advantages compared to unimodal methods. First, human expression is always a combination of multiple responses at the same time (e.g. the use of different words, different pitch and amplitude of the voice, different facial expressions). Multimodality is able to capture combinations of these responses, rather than individual responses and therefore reflects the nature of human expression better than unimodality. Second, where data is missing in unimodal models, in multimodal models it is less likely that all modalities lack data points at the same time, creating a possibility to more continuous observation given the available channels. Last, unimodal measurement of sentiment is inherently noisy, as a single response could imply multiple sentiments. For instance, an increase in pitch

could imply both positive and negative sentiment (Ververidis, & Kotropoulos, 2006). Multimodality is able to dissect the unimodal effects by looking at combinations of effects, thus reducing noise. For example, this increase in pitch combined with words like “not good” might indicate negative sentiment, whereas a combination with “really good” might indicate positive sentiment.

2.5 SPEAKER DIARIZATION

Speaker diarization is the process of labeling a speech signal with labels corresponding to the identity of speakers (Moattar, & Homayounpour, 2012). In the process, patterns in the audio files are recognized by various audio characteristics, to subsequently recognize which parts of the audio belong to which speaker. The result is a set of time frames corresponding to the audio file, with a labeled speaker for each time frame.

Although many different approaches to speaker diarization exist, Tranter, & Reynolds (2006, page 1558) identify a “prototypical combination of the key components of a diarization system”. The first of which is detection of the parts of an audio recording that contain speech, as opposed to noise, music, silence, etc. This step is often approached as a multiclass classification task. The next step aims to detect possible change points in time at which the speakers are likely to switch. This step is often approached by use of distance measures and thresholds between the two parts of audio enclosing the change point. Next, each audio part is classified according to gender and bandwidth (a measurement of audio quality). This step is included to improve subsequent steps, for instance by finding different optimal parameter for each gender. Clustering is then used to find which parts of the total audio file belong to which speaker. The last step, resegmentation, can be seen as a reconsideration of the original segment boundaries using the information found in the clustering step.

Little research has been performed on the relationship between speaker diarization and sentiment analysis. Vaudable, & Devillers (2012) compare an approach to sentiment analysis using automatic segment detection to an approach using manual segmentation. Here, automatic segment detection equals speaker diarization. Manual segmentation was performed by not only separating speakers, but also separating segments on emotional changes. They found that manual segmentation yielded a higher performance as compared to automatic segmentations. However, they did not test the results of either segmentations against an unsegmented method. In contrast, Mairesse, Polifroni, & Di Fabbrizio (2012) do compare a speaker dependent and a speaker independent approach. They report that without speaker diarization, none of the models outperform a majority class baseline, likely due to the large inter-speaker variation. After separate normalization for each speaker using speaker diarization, both audio only models and models using audio-text combinations outperformed the baseline. However, after normalization for each speaker, classification was done again for the two speakers combined. This way, they only used speaker diarization to test the

effect of speaker dependency, rather than using speaker diarization to test the effect of isolating the contribution of a single speaker. A method that actually separates segments for each speaker and test this against an unsegmented model seems to be non-existent, to our knowledge.

2.6 EARLIER WORK ON A SIMILAR DATASET

Earlier work at the internship company was performed by van Geffen (2008). In this research, a comparison was made between sentiment analysis models using only audio features, only text features and a combination of both. The main findings indicate that the use of audio features only yielded higher results, than the use of a combination of audio and text features or text features only. The research briefly touched the topic of speaker diarization, but it never formally tested its effect on sentiment analysis performance. It did test a speaker dependent and speaker independent model against each other (i.e. it did test an effect of speaker diarization), however diarization was only used for separate normalization, not complete separate analysis. Moreover, the speaker dependent and independent models differed in more aspects than just normalization, namely the software they used for feature extraction. This makes formal testing of the effect of speaker dependency impossible.

In addition, the research was susceptible to several quite severe limitations. First, the amount of annotated sentiment labels was reasonably low ($N = 3716$, labeling was done on 30-second audio fragments of customer service call recordings). Moreover, the vast majority of this annotated data was labeled as neutral, with highly negative data only accounting for 6% of the total ($N = 223$) and highly positive data only accounting for 5% of the total ($N = 186$). On top of that, the distributions of the features displayed high similarity between classes, making it hard for classifiers to distinguish between classes.

3. METHODS

The next sections will provide a thorough description of the methods used to answer the formulated research questions. Specifically it will describe the dataset that has been used, the methods of feature extraction and feature selection for both audio and text, as well as the method of combining these two and an overview of all classification algorithms will be given. The chapter will conclude with a description of the methods of evaluation.

3.1 DATASET DESCRIPTION

The original dataset consists of *N* = 217,219 audio recordings stored as .wav files. These audio recordings contain 30 second fragments of customer service calls of three big commercial companies and one public organization: Nuon, Tele2, Aegon, and the municipality of Zeist (respectively: *N* = 106,931, *N* = 53,618, *N* = 34,069, & *N* = 23,601). Each fragment has a filename corresponding to the complete call it is part of, as well as a unique fragment number and time frame indication within that call, allowing for tracing back the fragments to the complete conversation. In total *N* = 5,543 of these fragments were annotated by hand, accounting for 13.1%. These annotations were performed according to Thayer's arousal-valence model of emotion (Thayer, 1982), which can be found in *Figure 2*. In this model, emotion can be classified over the two dimensions of arousal and valence, where arousal represents the intensity of the emotion and valence represents whether an emotion is positive or negative. In the current research, the annotation on the arousal scale was not regarded. The annotation of valence was performed on a scale from a to e ("a" = very negative, "b" = negative, "c" = neutral, "d" = positive, "e" = very positive). The ratios of these annotations per company, as well as combined can be found in *Table 1*.

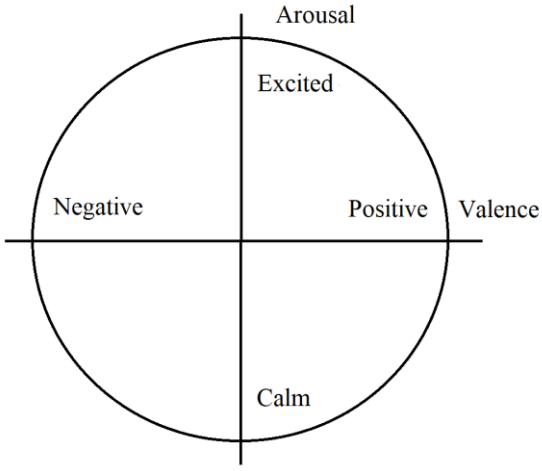

Figure 2: Thayer's arousal-valence model of emotion (Thayer, 1982).

	A	b	c	d	e
Nuon	88	403	554	454	67
Tele2	153	521	871	375	75
Aegon	85	407	415	436	50
Zeist	33	169	243	122	22
Total	359 (6.5%)	1500 (27.1%)	2083 (37.6%)	1387 (25%)	214 (3.9%)

Table 1: Numbers of annotations per sentiment class for each company and their combined amounts.

Speech recognition software by Nuance was used to transcribe the audio files into text, as well as to perform speaker diarization. This transcription was not performed on the separate fragments, but on the entire conversations so that in a later stage speaker numbers will be consistently matched to the same speakers throughout different fragments. Nuance returns transcriptions in JSON format for each recognized word as is depicted in *Figure 3*. Because each word is tagged with a speaker number and a time frame, it was possible to separate the words by speaker for each 30 second fragment. Concatenating these words for each speaker within each time frame results in a transcription of what is being said by each speaker separately for each fragment. These transcriptions were added to the dataset as a column of strings, each transcription represented as a collection of words in a single string. Another column was added indicating whether the transcription contained the words of speaker 1, speaker 2 or both. Based on the assumption that the call center agent always answers the phone and given that speaker numbers are labeled to the same speaker throughout fragments within each call, three diarization subsets can be separated: a subset containing only measurements of the agent's contribution (speaker 1), a subset containing only measurements the customer's contribution (speaker 2) and a subset containing measurements of their combined contributions as a whole. On average, the fragments contained 26.2 words for the agent, 36.1 words for the customer and 62.3 words for both speakers.


```
"13":{
   "start":1.82,
   "end":1.92,
   "weight":0,
   "best_path":true,
   "word":"een",
   "intensity":0
},
"14":{
   "start":1.92,
   "end":2.4,
   "weight":0,
   "best_path":true,
   "word":"tweetal",
   "intensity":0
},
"15":{
   "start":2.4,
   "end":3.24,
   "weight":-0.529047,
   "best_path":true,
   "word":"facturen",
   "intensity":0
},
"16":{
   "start":2.4,
   "end":3.24,
   "weight":-1.214662,
   "best_path":false,
   "word":"sturen",
```


Figure 3: Nuance output

3.2 FEATURE EXTRACTION

The recordings and their transcriptions were used to extract textual features and features based on audio. Both processes will be described in the following sections. Some parts of the preprocessing were identical for each of the three subsets (customer, agent, both) and therefore performed on the entire dataset, others were tailored to each split and therefore performed after the dataset is separated into three subsets.

3.2.1 TEXTUAL DATA

Textual data was obtained by using the speech-to-text transcriptions by the Nuance software. These transcriptions were all in Dutch, including English words that are commonly used in daily Dutch language, such as “no” and “shit”. Nuance was not set to recognizing other languages, causing conversations in other languages to be problematic. However, these conversations seemed to rarely occur. Although performance of this software seemed suboptimal based on face validity, a single transcription was tested for accuracy, which was found to be 80%. More testing needs to be done to formally assess its performance on this specific dataset, however because the short time frame for this project, this was not possible.

First, transcripts that were too short (less than 4 words) were removed, as they were believed to contain too little information to be valuable. A list of stop words set up by the internship company Totta was combined with a more extensive list found at https://eikhart.com/nl/blog/moderne-stopwoorden-lijst. However, initial experiments showed that accuracy scores were higher for no exclusion of stop words, compared to each list of stop words as well as the combined list. For this reason, it was decided to not exclude stop words.

Next, the Nuance software does not identify sentences and thus only capitalizes words that are ought to be capitalized regardless of their syntactic position, such as names of people and streets. Although the effect of exclusion of capitalized words on final accuracy scores can be tested too, any possible increase is likely due to coincidental effects in the test set and is not expected to generalize to test data, given the nature of these capitalized words. These capitalized words are not expected to be informative regarding sentiment. Capitalized words will therefore be excluded regardless (i.e. the entire words were removed from the transcripts).

Next the transcriptions were transformed into tf*idf representations (Salton, & Buckley, 1988). These representations consist of two components: a term frequency indicating how often a term appears in a particular document and an inverse document frequency, indicating the inverse number of documents containing the term. In this case the term frequency will represent how often that term is found in a particular call and the inverse document frequency will represent the inverse number of calls that contain the term,

over all calls. The product of these components can be seen as a weighting for terms, assigning higher weights to terms that appear frequently in specific documents, but infrequently over all documents. This way, the tf*idf representation aims to emphasize informative words. The decision for this method was based on the fact that the transcriptions needed to be analyzed as individual words or N-grams, as transcriptions contained too few correct grammatical structures to include methods that analyze syntactic structure. In contrast, tf*idf allows for analysis of individual words or N-grams, while still weighing their importance in relation to their contexts. Because the tf*idf algorithm was run for each diarization subset (consumer, agent, both) separately, all words were weighted against all other words within the same subset.

In addition, various hyperparameters have been optimized for each diarization subset separately. Every hyperparameter decision was made either because it led to an increase in performance for all models, or because it led to an increase in performance for the best performing model. First, the effect of enabling or disabling add-one smoothing of the tf*idf representations on final accuracy scores has been tested. Smoothing was expected to have a negative effect on accuracy as adding one to absent terms causes representations of calls to be artificially more similar. However, enabling smoothing caused an increase in accuracy for nearly all models in all diarization subsets and was therefore enabled.

Next, the effect of normalization of the tf*idf representations was tested. No normalization was tested against L1 and L2 normalization. Normalization was expected to have a positive effect on accuracy scores, as it counters effects caused by differences in length of the transcriptions. However, disabling normalization caused an increase in nearly all models for each of the diarization subsets and normalization was therefore disabled.

Subsequently, different ranges of N-grams were examined. Including N-gram ranges was expected to have a positive effect on accuracy as bi- or trigrams have the ability to detect simple negations (e.g. “not good”). Indeed, for the customer subset and the agent subset adding a range of N-grams increased accuracy. However, this effect was not found for the both subset and was therefore disabled in this subset. Last, different cutoff values have been tested for terms that appeared in more than a certain percentage of all calls. Setting a cutoff value was expected to have a positive effect on accuracy scores, as there were thus far no steps taken for removing stop words (recall that stop word exclusion by using a stop word lexicon decreased accuracy and was therefore not included). Excluding the most frequently appearing terms is expected to function as a form of stop word exclusion. Indeed, for the customer subset and the agent subset, including a cutoff value increased performance. This was not the case for the both subset. For an overview of the hyperparameters and their settings for each diarization subset, see *Table 2*.

	Add-one smoothing	Normalization	N-gram range	Frequent words
Customer subset	Enabled	Disabled	Uni- and bigrams	20% cutoff
Agent subset	Enabled	Disabled	Uni-, bi- and trigrams	10% cutoff
Both Subset	Enabled	Disabled	Only unigrams	No cutoff

Table 2: hyperparameter settings for the tf*idf algorithm for each of the diarization subsets.

3.2.2 AUDIO DATA

Feature extraction for audio data was done by using pyAudioAnalysis and pySPTK. These software packages were used to calculate different prosodic values for each audio record. The process of feature extraction starts with so-called short-term feature extraction: audio files are cut into small time frames and a number of features are then computed for each frame. Each audio record is then represented as a sequence of feature vectors. In total pyAudioAnalysis is able to extract thirty-four different kinds of prosodic features, out of which eight were selected based on past experiences with their performances by the internship company. In addition, pySPTK was used to add another feature. In total, the following features were extracted for analysis:

- F0: a measurement of pitch, where the F0 of a spectrum is equal to the pitch of the lowest frequency.
- Amplitude: a measurement of energy, calculated as the sum of squares of the signal values, normalized by the respective frame length.
- Zero Crossing Rate: a measurement of the amount of times the signal changes sign within a certain time frame.
- Entropy of Energy: a measurement of abrupt energy changes, calculated as the entropy of sub-frames' normalized energies. This is a measurement of irregularities in energy over time.
- Spectral Centroid: the center of the spectrum’s gravity. This is a measurement of central tendency over all frequencies in a sound. Spectral centroids are often a good predictor of a sound’s “brightness”.
- Spectral Spread: the second central moment of the spectrum. This is a measurement of spread in relation to the spectral centroid.
- Spectral Entropy: entropy of the normalized spectral energies for a set of sub-frames. This is a measurement of irregularities in energies of frequencies in a spectrum. Meaning that if all

frequencies have equal energies, spectral entropy is high, whereas if the energies of all frequencies are irregular, spectral entropy is low.

- Spectral Flux: the squared difference between the normalized magnitudes of the spectra of the two successive frames. This measurement effectively cuts the audio in subframes and calculates the distances between adjacent frames in spectral energy.
- Spectral Rolloff: the frequency below which 90% of the magnitude distribution of the spectrum is concentrated.

Each of these features was extracted for words spoken by the customer, agent, and both. Frames in the audio file that did not include speech (e.g. silence, background noise) were excluded. To create a speaker independent measurement, all values were normalized for each speaker. In addition, all models were run using speaker dependent, non-normalized values as well, just for the sake of completeness. Instead of representing each audio file as a sequence of these features, statistics are calculated as final representations. The statistics calculated for each prosodic feature are average, standard deviation, range, absolute average deviation, and peak rate (i.e. the amount of peaks of a given measurement per second). In total, these five statistics were calculated for the nine prosodic features, resulting in a set of forty-five features. The means and standard deviations of each of these statistics on all nine audio features for each of the sentiment classes separately are depicted in *Appendix A.*

3.3 FEATURE SELECTION

After features were extracted from text and audio, it was key to find the most informative (i.e. most predictive) subset of features. Although there are some similarities in selection methods for text and audio features (e.g. the criteria of selection), the exact implementation of these selection methods was different. This difference was mainly due to the fact that each subset contained only 45 features for audio, but between 3436 and 17,943 features for text. The approach for audio features was therefore more exhaustive, whereas the approach for text relied more on testing different selection methods. Here again, the selection methods were performed separately for each subset (consumer, agent both), allowing for finding the optimal selection of features within that specific subset.

3.3.1 TEXTUAL DATA

As mentioned before, the feature extraction approaches led to immense amounts of features. This is not unusual in text analysis using tf*idf representations. In addition, the use of the N-gram ranges increased the number of features drastically (also explaining the big differences in numbers of features between subsets).

These huge numbers make it impossible to explore all possible combinations of features. Instead a smarter approach to feature selection is needed. Three of these approaches were optimized for each subset separately: feature selection based on a chi-squared test, feature selection based on mutual information, and dimension reduction based on truncated singular value decomposition.

- Chi-squared test: this test is a statistic that measures the fit between two variables. It quantifies uncertainty by measuring dependency between the variables. An insignificant test score means differences between variables might be due to chance. Here, a chi-squared test score was calculated for each feature, thus having a measure of dependency between the sentiment class and each feature. The best *k* features were then selected by weeding out the features with the lowest dependency. The optimal value of *k* was found by testing its effect on accuracy scores manually.
- Mutual information: this measurement is similar to the chi-squared test as it also measures dependency between two variables. It is, however, not a quantifier of uncertainty. Instead, it measures how much information about one variable can be gained by observing another variable. This way, all features can be ranked based on how information they share with the sentiment classes. The best *k* features were selected by weeding out the least informative features. The optimal value of *k* was found by testing its effect on accuracy scores manually.
- Truncated singular value decomposition (SVD): whereas the previous two methods selected a subset of features from the total, truncated SVD decomposes the total amount of features into a smaller number of new features. It can be seen as a form of dimension reduction, similar to principal component analysis (PCA). PCA however does not support the use of sparse matrices in Scikit Learn, while SVD does. Features are reduced to a set of *k* dimensions. The optimal value of *k* was found by testing its effect on accuracy scores manually.

As a result, for each subset there are now three different selections (or rather two selections and one reduction) of features. For every classifier to be tested, each of these three selections will be used, allowing for comparison of performance within each classifier.

3.3.2. AUDIO DATA

In contrast, there were forty-five audio features extracted, five of which (all statistics on spectral entropy) registered nothing but zero values and where therefore excluded. This leaves forty features for analysis, far fewer than there were for text data. This allowed for a more exhaustive approach to feature selection. Still, with forty-five features the total number of possible combinations is 2^40-1 = 1,099,511,627,775. Therefore, a selection method is still needed. Similar to feature selection for text features, the chi-squared

test and mutual information statistics were used to quantify dependencies for each feature. However, this time an algorithm was developed that iterated through all possible values of *k* (= 1:40), allowing for testing of the most informative feature alone all the way through testing all features together. This procedure thus resulted in 2x40 = 80 sets of features that were fed to each of the classifiers.

3.3.3. COMBINING AUDIO AND TEXTUAL DATA

Features extracted from text data and audio data were combined to into a single multimodal set of features. The approach of combination is inspired by the approach of Mairesse, Polifroni, & Di Fabbrizio (2012), who took the predictions of their text-based model and added these predictions as a single feature to their best performing set of audio features. This approach seems to have several advantages to simply merging all features. First, prediction by text-based models can be represented as a single feature, resolving issues caused by high dimensionality. Second and related, this method prevents having to work with a huge and mixed pile of features that once again might require selection methods, in turn preventing the possibility of the small amount of audio features being completely weeded out. Moreover, this method has been shown to be able to improve sentiment analysis performance as compared to unimodal approaches and it seemed to be able to correct for noisy automatic speech transcription.

3.4. ALGORITHMS

After performing the previously described extraction and selection methods for both audio and text features separately for consumer, agent and both, it is key to test these subsets of features against each other using different classifiers. Because the primary scientific objective of this research is not to find the best performing sentiment classifier but to test the effect of diarization classification on these classifiers, a selection of commonly used, straight-forward classifiers is made. By testing several of these classifiers rather than a single methodologically advanced classifier, an attempt is made to obtain a generalizable effect of diarization classification. The following classifiers have been tested:

- Multinomial Naive Bayes: Naive Bayes can be considered a family of probabilistic classifiers. Class membership is predicted based on a calculation of conditional probability: the probability that an instance belongs to a certain class, given its features. When using regular word counts, the features follow a multinomial distribution. Although theoretically this is not the case for tf*idf representations, in practice the algorithm often still performs well on these representations. For each diarization class subset the classifier was trained using 10-fold cross-validated grid search (alpha = [0.001, 0.01, 0.1, 1]).

- Support Vector Machine: this is a linear binary classifier in which a decision boundary is found by finding the best line based on each class' edge cases, called support vectors. Different kernel methods can be used to transform the feature space in order to separate cases that are otherwise not linearly separable. For each diarization class subset the classifier was trained using 10-fold cross-validated grid search (kernel = [linear, poly, rbf, sigmoid]).
- K-Nearest Neighbors: this is a very straightforward classifier that calculates distances between training instances and new instances and classifies these new cases based on similarity to the *k* nearest neighbors. The optimal value of *k* has been found for each diarization class subset using 10-fold cross-validation with *k* = [1, 2, 3, 4, 5, 6, 7, 8].
- Logistic Regression: this is a statistical model that uses log transformed odds of binary dependent variables for classification. These odds in turn are linear combinations of one or more independent variables. For each diarization class subset the classifier was trained using 10-fold cross-validated grid search (C = [0.001, 0.01, 0.1, 1, 10, 100]).

In addition, in regular classification, class membership is predicted by calculating a probability of class membership for all classes and selecting the class with the highest probability. For binary classification, this means predicting class membership based on argmax (or a 0.5-threshold) of the class probabilities, selecting the class with a probability higher than 0.5. However, for analysis of the audio features, using this cut-off value of 0.5 yielded poor results. Therefore, all classifications using audio features were performed by predicting 10-fold cross validated class probability rather than class membership. Subsequently, instead of using argmax for these predictions, an algorithm iterated through all possible thresholds for cut-off (*k* = [0.01, 0.02, … 0.99, 1]), selecting the cut-off value that lead to the highest accuracy. Iterating through all possible thresholds was able to slightly improve accuracies.

3.6 EVALUATION

For each subset of the diarization classifications (customer, agent, both), the data was split into a train and a test set (test set proportion = 0.33). Within each train set, performance was measured in accuracy using 10-fold cross-validation. After tuning of the parameters for each model within the training set, final performance was measured using the test set. Given the problem of data scarcity, one could argue it might be better to use all data for training and base evaluation solely on the 10-fold cross-validation scores, rather than taking out a part of the data for testing too. However, taking out this test set allows for testing on previously unseen data, to provide information regarding generalizability of the 10-fold cross-validated results. General performance was evaluated by comparing accuracies to a majority class

baseline. The decision for accuracy was made based on the fact that classes were reasonably equally distributed when comparing the outer two classes to each other (see section 4.1 for an explanation why outer classes were compared), which allowed accuracy scores to serve as an intuitively understandable measurement of performance. Next, to test hypotheses of this research, a comparison of accuracy scores between the customer subset and the both subset was made for each classifier separately. These comparisons will be statistically tested for significance by conducting pairwise independent samples t-tests. In addition, differences between classifiers within each diarization subset will be tested for significance by use of analysis of variance (ANOVA) and post hoc Tukey tests. The performance of the different feature selection methods and all hyperparameter tunings was assessed by regarding their overall effects on the accuracy of all classifiers.

4. RESULTS

The current section will present various results related to the research questions of this paper. First, results for different comparisons of sentiment classes will be discussed. Next, for each diarization subset, a majority class baseline score is reported (there are minor differences between each diarization subset, as will be explained in section 4.2.). Subsequently, results within each of the three diarization subsets will be reported, followed by comparisons of these results between diarization subsets. This will be done for text data, audio data and combined data separately. These comparisons are relevant for answering the first research question of this paper. Next, results of the models that combines features from the customer subset and the agent subset will be reported as they are relevant for answering the second research question of this paper. The chapter concludes with a section on results regarding generalizability.

4.1. CLASS COMPARISONS

As discussed in the method section, the annotation of the sentiment labels was performed on a 5-point scale (“a” = very negative, “b” = negative, “c” = neutral, “d” = positive, “e” = very positive). This allows for various boundaries to be tested in binary classifications. Initially, an attempt was made to test the most negative class against all other classes (i.e. “a” versus “b”, “c”, “d”, “e”). For this comparison, a majority class baseline resulted in the very high accuracy score of 93.5%, as a result of classes being unbalanced. Because of this reason, in the initial approach an attempt was made to compare models based on precision scores, rather than accuracies. However, various preliminary results indicated that precision scores could not be calculated. This was due to the fact that these models just assigned the majority class to all cases (precision scores only regard true and false positives, and the majority class baseline only assigns

negatives). Indeed, for these models accuracy scores were equal to the majority class baseline. A second approach was therefore to artificially create more polarized data, by comparing only highly negative data to highly positive data (i.e. “a” vs. “e”). Preliminary testing on text data showed promising results and therefore this approach was continued to be used. There was one theoretical argument against this approach for audio analysis: some measurements of prosody have been found to show similar patterns for highly positive and highly negative data (Ververidis, & Kotropoulos, 2006). However, testing all models by comparing highly negative data to neutral data (i.e. “a” vs. “c”) showed little improvement for the audio data but a large negative effect for text data. For the remaining parts of this paper, therefore, this second approach to class comparisons (“a” vs. “e”) will be used and results on this comparison will be reported.

4.2. MAJORITY CLASS BASELINES

Majority class baselines have been set up for each of the diarization subsets. As classes are reasonably equally distributed (N = 359 for data labeled as highly negative and N = 214 for data labeled as highly positive), accuracy scores were used to assess performance. However, minor differences in accuracy scores of these baselines exist between the diarization subsets. These differences are a result of randomly separating a proportion of the data to later be used as previously unseen test data. For the customer subset, accuracy of the majority class baseline is .644. For the agent subset, accuracy of the majority class baseline is .635. Last, for the both subset accuracy of the majority class baseline is .643.

4.3. TEXTUAL DATA

In the following section, results will be discussed regarding all models that used features extracted from the textual data only. Results will be discussed for each of the diarization subsets separately first, comparing results within each diarization subset. After, comparisons of results between the diarization subsets will be made, in order to examine the hypothesized effect of diarization classification using textual data (*Hypothesis 1a*).

4.3.1. TEXTUAL DATA FOR CUSTOMER ONLY

In total, four different selections of features were used (all features, chi-squared selected features, mutual information selected features and reduced features after SVD). Each of these selections were combined with each of the four classifiers (Naive Bayes, SVM, Logistic Regression, KNN), resulting in a total of sixteen possible models to test. However, the negative values in features created by SVD were not

compatible with the Naive Bayes algorithm. Scaling could have solved this problem, however given the generally low performance of this method in other classifiers and given that scaling might cause effects on accuracy that are not found in the other models, it was decided that this model lacks relevance. Therefore, it was excluded from analysis, leaving fifteen models to be tested. The 10-fold cross validated accuracy results of these models can be found in *Table 3*.

	All features	Chi-squared	Mutual Information	SVD
Baseline	.644	.644	.644	.644
Naive Bayes	.577	.885	**.913**	-
SVM	.664	.866	.826	.653
Logistic Regression	.689	.821	.846	.652
KNN	.423	.700	.667	.646

Table 3: Average 10-fold cross validated accuracy scores for all combinations of feature selection methods and classifiers for the customer subset.

Although no feature selection and SVD yielded similar or worse results than the majority class baseline of .644, the chi-squared and mutual information methods did outperform this baseline. Moreover, for every classifier the best performing model outperformed the majority class baseline considerably. The best results were gained by using a Naive Bayes classifier in combination with a mutual information selection method (.913 accuracy). Analysis of variance (ANOVA) between each classifier's best performing model indicated that differences between classifiers in accuracy scores are significant ($F(3, 36) = 26.59$, $p < .001$). A post hoc Tukey test indicated that for the best performing Naive Bayes model ($M = .913$, SD = .038), differences were insignificant ($p = .261$) compared to the best performing SVM model ($M = .866$, SD = .064), moderately significant ($p = .054$) compared to the best performing Logistic Regression model ($M = .846$, SD = .066) and significant ($p < .001$) compared to the best performing KNN model ($M = .7$, SD = .053). Together these results suggest the Naive Bayes classifier is able to outperform the other classifiers for text data within the customer subset, although not all differences are significant.

4.3.2. TEXTUAL DATA FOR AGENT ONLY

Similar to the analysis of textual data for the customer subset, four different selections of features were used (all features, chi-squared selected features, mutual information selected features and reduced features after SVD) and again each of these selections were fed to each of the four classifiers (Naive Bayes, SVM, Logistic Regression, KNN). Here too the combination of SVD features and Naive Bayes was excluded, leaving again 15 models to be tested. The 10-fold cross validated accuracy results of these models can be found in *Table 4*.

	All features	Chi-squared	Mutual Information	SVD
Baseline	.635	.635	.635	.635
Naive Bayes	.513	**.896**	.872	-
SVM	.659	.878	.819	.644
Logistic Regression	.632	.828	.819	.638
KNN	.641	.712	.682	.540

Table 4: Average 10-fold cross validated accuracy scores for all combinations of feature selection methods and classifiers for the agent subset.

Although no feature selection and SVD yielded similar or worse results than the majority class baseline of .635, the chi-squared and mutual information methods did outperform this baseline. Similar to analysis of the customer subset, the best results were gained by using a Naive Bayes classifier, however this time in combination with the chi-squared selection method (.896 accuracy). Analysis of variance (ANOVA) indicated that differences between each classifier's best performing model in accuracy scores are significant ($F(3, 36) = 23.56$, $p < .001$). A post hoc Tukey test indicated that compared to the best performing Naive Bayes model ($M = .896$, SD = .061), differences were insignificant ($p = .878$) for the best performing SVM model ($M = .878$, SD = .056), significant ($p = .037$) for the best performing Logistic Regression model ($M = .828$, SD = .055) and significant ($p < .001$) for the best performing KNN model ($M = .712$, SD = .042). Together these results suggest the Naive Bayes classifier is able to outperform the other classifiers for text data within the agent subset, although not all differences are significant.

4.3.3. TEXTUAL DATA FOR BOTH CONTRIBUTIONS

For the subset that contained the combined contributions of the customer and the agent, the same combinations of feature selection methods and classifiers were used. The 10-fold cross validated accuracy results of these models can be found in *Table 5*.

	All features	Chi-squared	Mutual Information	SVD
Baseline	.643	.643	.643	.643
Naive Bayes	.625	**.887**	.853	-
SVM	.682	.832	.798	.648
Logistic Regression	.687	.835	.782	.661
KNN	.367	.690	.659	.601

Table 5: Average 10-fold cross validated accuracy scores for all combinations of feature selection methods and classifiers for the both subset.

There were some models that performed worse that the majority class baseline of 64.3%, however here too for every classifier the best performing model outperformed the majority class baseline considerably. Again, the best results were gained by using a Naive Bayes classifier in combination with a chi-squared selection method (88.7% accuracy). Analysis of variance (ANOVA) indicated that differences between each classifier's best performing model in accuracy scores are significant ($F(3, 36) = 41.12$, $p < .001$). A post hoc Tukey test indicated that compared to the best performing Naive Bayes model ($M = .887$, SD = .035), differences were significant ($p = .027$) for the best performing SVM model ($M = .832$, SD = .042), significant ($p = .040$) for the best performing Logistic Regression model ($M = .835$, SD = .045) and significant ($p < .001$) for the best performing KNN model ($M = .69$, SD = .044). Together these results suggest the Naive Bayes classifier is able to significantly outperform all other classifiers for text data within the agent subset.

4.3.4. COMPARING DIARIZATION SUBSETS

In order to test whether isolating the customer's contribution will increase sentiment analysis performance for text data (*hypothesis 1a*), comparisons must be made between diarization subsets. Specifically, for each classifier the best performing model of the customer subset is statistically tested against the best performing model of the subset containing both consumer and agent, by using independent sample t-tests. Means of

each best performing model and their standard deviations in 10-fold cross validated accuracy for each two subsets can be found in *Table 6*, along with the pairwise t-test results.

	Customer subset		Both subset		t-test results		
	M	SD	*M*	SD	*t*	*df*	*p*
Naive Bayes	.913	.038	.887	.035	1.59	18	.129
SVM	.866	.064	.832	.042	1.40	18	.177
Logistic Regression	.846	.066	.835	.046	0.43	18	.671
KNN	.700	.053	.690	.044	0.46	18	.652

Table 6: 10-fold cross-validated accuracy results for each classifier for the customer and the both subset and t-test results for their comparison.

For each of the classifiers, the accuracy scores are slightly higher when the customer's isolated contribution has been analyzed, compared to when the conversation is analyzed as a whole. These results imply that analyzing the isolated customer's contribution improves sentiment analysis performance, compared to treating the conversations as a whole. However, these differences were all not found to be significant.

4.4. AUDIO DATA

In the following section, results will be discussed regarding all models that used features extracted from the audio data only. Results will be discussed for each diarization separately first, after which comparisons between results of the diarization subsets will be made, in order to examine the hypothesized effect of diarization classification using audio data (*Hypothesis 1b*).

4.4.1. AUDIO DATA FOR CUSTOMER ONLY

For the analysis of audio data, only two methods of feature selection were used (chi-squared and mutual information). However, because the number of features to be selected *k* was set on a range from one to forty (the total amount of features) for each method, in total eighty sets of features were tested for each of the four classifiers (Naive Bayes, SVM, Logistic Regression, KNN), resulting in a total set of 320 models. Because of this large amount, only the best performing set of features for each selection method will be reported for each classifier. In addition, instead of predicting classes themselves, probabilities of classes were predicted and subsequently an algorithm iterated through all possible thresholds as cut-offs for class

predictions, selecting the best performing threshold. The accuracy results of these models as well as their thresholds can be found in *Table 7*.

	Chi-squared	Threshold	Mutual Information	Threshold
Baseline	.644	-	.644	-
Naive Bayes	.647	.61	.644	.00
SVM	.653	.64	.658	.64
Logistic Regression	.653	.56	.653	.40
KNN	.661	.34	**.669**	.25

Table 7: Accuracy scores of the best performing set of features and their threshold values for cut-off, for each selection method, for each classifier.

Although some of these accuracies are higher than the majority class baseline of .644, these differences are very minimal and not expected to be significant. For both feature selection methods, KNN is found to yield the highest performance. However, differences between the four classifiers are also minimal and not expected to be significant. Unfortunately, this significance cannot be tested, as just a single prediction for classes exists for each classifier, as opposed to multiple predictions done by cross-validation. Cross-validation was used to predict probabilities of class labels, rather than class labels themselves. After these predictions, different thresholds for cut-off values have been used to transform these probabilities into class label predictions. This can be seen as a form of hyperparameter tuning after cross validation, which causes testing for significance using the means and standard-deviations of these probabilities to not represent significance in differences between final class label predictions.

4.4.2. AUDIO DATA FOR AGENT ONLY

For the analysis of audio data using the agent subset, the same methods of feature selection and classifiers were used as in the customer subset. The accuracy results of these models as well as their thresholds can be found in *Table 8*.

	Chi-squared	Threshold	Mutual Information	Threshold
Baseline	.654	-	.654	-
Naive Bayes	.653	.62	.644	.62
SVM	**.671**	.59	.656	.57
Logistic Regression	.659	.50	.653	.55
KNN	.659	.25	.650	.00

Table 8: Accuracy scores of the best performing set of features and their threshold values for cut-off, for each selection method, for each classifier.

Here, nearly all models perform slightly better than the majority class baseline of .654, however differences are very small and not expected to be significant. Best performances are found for the SVM classifier, for both selection methods. However, differences between the four classifiers are also small and not expected to be significant.

4.4.3. AUDIO DATA FOR BOTH CONTRIBUTIONS

For the analysis of audio data using the subset containing contributions from both the customer and the agent, the same methods for feature selection and classifiers were used as in the isolated subsets. The accuracy results of these models as well as their thresholds can be found in *Table 9*.

	Chi-squared	Threshold	Mutual Information	Threshold
Baseline	.643	-	.643	-
Naive Bayes	.651	.61	.646	.63
SVM	.661	.57	.667	.51
Logistic Regression	.667	.47	.659	.46
KNN	.658	.00	**.669**	.29

Table 9: accuracy scores of the best performing set of features and their threshold values for cut-off, for each selection method, for each classifier, for audio data.

Here, nearly all models perform slightly better than the majority class baseline of .643, however differences are very small and not expected to be significant. Best performances are found for Logistic Regression for the chi-squared method and for KNN for the mutual information method. However, again differences between the four classifiers are small and not expected to be significant.

4.4.4. COMPARING DIARIZATION SUBSETS

In order to answer the research question whether isolating the customer's contribution will increase sentiment analysis performance for audio data (*hypothesis 1b*), comparisons must be made between diarization subsets. Specifically, for each classifier the best performing model using the customer's isolated data must be compared to the best performing model using the data containing both contributions. *Table 10* summarizes these performances as well as the difference in performance of these diarization subsets.

	Customer subset	Both subset	Difference
Naive Bayes	.647	.651	.004
SVM	.658	.667	.009
Logistic Regression	.653	.667	.014
KNN	**.669**	**.669**	0
Average of all classifiers	.657	.664	.007

Table 10: Accuracy scores for all classifiers for the subset containing the isolated customer's contribution and the subset containing the contributions of both the customer and the agent, as well as the difference between these subsets, for audio data.

For all classifiers, the subset containing both contributions performed better, implying isolating the customer's contributions actually reduces sentiment analysis performance. However, differences between subsets for each classifier are small and expected to be insignificant. In addition, averages over all four classifiers were calculated. Their standard deviations were calculated by taking the standard deviation of the distribution of all four classifiers' means, because no standard deviations are available for the individual classifiers. An independent samples t-test was used to statistically test whether the average performance of the customer subset (M = .657, SD = .009) differed from the average performance of the both subset (M = .664, SD = .008). This test was found to be insignificant (t (6) = 1.12, p = .307), indicating that performances are statistically not different for the two subsets.

4.5. COMBINING AUDIO AND TEXTUAL DATA

In the following section, results will be discussed regarding all models that used a combination of audio and text features. Results will be discussed for the customer subset first, followed by results of the subset containing both contributions. Subsequently, comparisons between results of the diarization subsets will be made, in order to examine the hypothesized effect of diarization classification using combined text and audio data (*Hypothesis 1c*).

For combining audio and text data, the best performing audio models were combined with a single feature containing predictions by the best performing text model. Specifically, for the customer subset, the Naive Bayes text model that yielded an accuracy of .913 was used to predict classes and these classes were added as a feature to all sets of audio features. This combination was then tested for each of the four classifiers. Accuracies of the best performing models for each classifier can be found in the column "Customer subset" in *Table 11*.

A similar approach was used for the subset containing both contributions. The Naive Bayes classifier that yielded an accuracy of .887 was used to predict classes and these classes were added as a feature to all sets of audio features. This combination was then tested for each of the four classifiers. Accuracies of the best performing models for each classifier can be found in the column "Both subset" in *Table 11*.

	Customer subset	Both subset	Difference
Naive Bayes	.933	.940	.007
SVM	.941	**.950**	.009
Logistic Regression	.939	.947	.008
KNN	.941	.942	.001
Average over all classifiers	.939	.945	.006

Table 11: Accuracy scores of the best performing set of combined text and audio features for each classifier, for the customer subset and the subset containing both contributions separately.

For all classifiers, the subset containing both contributions performed better. However, differences between subsets for each classifier are small and expected to be insignificant. In addition, an independent samples t-test was used to statistically test whether the average performance of the customer subset (M = .939, SD = .003) differed from the average performance of the both subset (M = .945, SD = .004). This test was found

to be moderately significant (t (6) = 2.33, p = .058), indicating that performances are somewhat better for the subset containing both contributions when a text and audio are combined.

For both subsets using both text and audio, increased accuracy scores were found for all classifiers compared to all unimodal models. This also means these combined models performed better than the best performing Naive Bayes models that were used to create the prediction features that were added in these combined models. For the customer subset, combining text and audio lead to an increase in accuracy from .913 to .941. For the subset containing both contributions, this led to an increase from .887 to .947. These results indicate that the method used to combine text and audio was successful.

4.6. ADDING THE AGENT'S TEXT FEATURE

Previous sections reported accuracy scores for various models for each diarization subset. In order to answer the second research question of this paper, the best performing model within the customer subset was taken and to it a feature was added that contained class predictions of the best performing text model in the agent subset. This combined model was compared to the original model without the added feature in order to assess whether adding the agent's text feature is able to improve the best performing model within the customer subset. For every classifier, the best results within the customer subset were obtained for models that combined text and audio. Therefore, the agent's text feature will be added to these models. *Table 12* reports accuracy scores for the models that combined text and audio in the customer subset, these models with the agent's text feature added and the difference between these models.

	Customer only	Agent added	Difference
Naive Bayes	.933	.931	-.002
SVM	.941	**.945**	.004
Logistic Regression	.939	.940	.001
KNN	.941	.935	-.006
Average over all classifiers	.939	.937	-.002

Table 12: Accuracy scores for the models that combined text and audio in the customer subset, these models with the agent's text feature added and the difference between these models.

Adding the agent's text feature results in effects in different directions for different classifiers. However, these differences are minimal and not expected to be significant. Indeed, an independent samples t-test indicated that averages over all classifiers for customer only (M = .939, SD = .005) and the agents text feature added (M = .937, SD = .003) are not significantly different (t (6) = .68, p = .518).

4.7. GENERALIZABILITY

All results that have been reported in previous sections were based on 10-fold cross validation within the train set. All models were optimized to yield the highest possible results within these train sets. After the best results were believed to be obtained, performances were measured on a previously unseen test set. Analysis on this test set indicated poor generalizability. Out of all models that were tested during this project, the highest accuracy on the test set was .667. This means that for all diarization subsets, for all modalities, for all feature selection methods, and for all classifiers none of the models were able to achieve an accuracy higher than .677. This best performing model was barely outperforming the majority class baselines. The sharp contrast of these test set accuracies and the high accuracies yielded within the train set imply the models were prone to overfitting.

5. DISCUSSION

The current research aimed to improve sentiment analysis on customer service calls using diarization classification. Specifically, it investigated whether isolating the customer's contribution would lead to more accurate results for text, audio and a combination of both. The following section will elaborate on the results that were found, on the limitations that the research faced and it will put the current findings in context of related work. This will be done for textual data, audio data, and combined data separately.

5.1 TEXTUAL DATA

First of all, for sentiment analysis on textual data, the Naive Bayes classifier outperformed all other classifiers consistently throughout the different diarization subsets, although not always significantly. These results are not surprising as Naive Bayes models have been found to be relatively well performing in text classification tasks (McCallum, & Nigam, 1998). What is surprising is the consistently high performances throughout all diarization subsets, with accuracy scores up to .913. For a difficult task as sentiment analysis, this can be considered a strong performance; the various comparative studies and meta-analyses in the "related work" section reported accuracies ranging from .59 to .93. However, when regarding performances

in predicting previously unseen test data, neither this best performing model, nor all other text models were able to yield an accuracy that was considerably higher than the majority class baselines.

These results suggest the models were strongly overfitted to the training data, causing these high performances to not be found in previously unseen data. Although unfortunate, this was not a big surprise. Previous work by van Geffen (2018) on a similar data set showed that insufficient distinctiveness could be found in the data, due to data scarcity and data neutrality. Since van Geffen (2018) more data and sentiment label annotations have been added, which possibly could have reduced these issues of data scarcity and neutrality. However, the current research suggest that the data was still insufficiently useful for sentiment analysis models. Indeed, the amount of cases that were labeled as highly negative was only 359 for each of the diarization subsets. These cases were compared to 214 highly positive cases, which lead to a total of 573 cases available for analysis. After splitting off 33% of this data for testing, only about 382 cases were left for training. It is expected that this small training set led to the models to be prone to overfitting, as overfitting is a common problem in small datasets (Hawkins, 2004).

Whereas the small dataset made the models prone to overfitting, it seems like the selection methods were actually causing this overfitting. Results for each of the diarization subsets show that if no selection method is used, and thus all representations of all words are used, results are hardly outperforming the majority class baselines and often even underperforming them. In contrast, after feature selection methods were applied and tuned, performances rose considerably. Closer examination of this selection process and the features they selected, revealed a possible reason for why performances in cross validation and the test set are so different: these selection methods assign some score of informativeness to each feature, based on their informativeness within the training set. However, when looking at the words these features actually represent, most of these words seem to not be related to sentiment. The list of selected words by mutual information selection can be found in *Appendix B*.

Although some words seem likely to be truly informative about sentiment (e.g. ‘weer goed, ‘dankjewel’, ‘shit’, respectively translated as ‘good again’, ‘thanks’, ‘shit’), most of them seem semantically unrelated to sentiment. The relationship between these words and sentiment classes, as detected by the selection methods, is only found within the small training set, thus causing the cross validated accuracy scores to increase. In previously unseen data no such relationship could be found. Possible solutions for this problem could have been to remove stop words from the transcriptions or to select features based on a lexicon containing words that are known to be related to sentiment. Both of these methods have actually been performed, however, both methods led to accuracy scores to be reduced to baseline levels again. Altogether, these results seem to show that either a non-generalizable relationship that only exists in the training set can be found, or no relationship at all.

An important limitation to the methods used in this research is that cross validated accuracy results were over-appreciated. Various choices that seemed theoretically sound, were not made because they led to large negative effects on the cross validated accuracy. An example of this is the choice to not include the lexicons of stop words. Stop words are theorized to be uninformative regarding sentiment and therefore often excluded. However, because the effect of inclusion on the held out test data was not known and the cross validated accuracy decreased considerably, the choice was made to not exclude stop words. Perhaps accuracies on the test set would have been slightly higher if these words were excluded. However, removing words would lead to even less data, while data scarcity is already an issue in the current research.

The effects of diarization classification on text sentiment analysis cannot be truly assessed, as the models cannot be considered successful in measuring sentiment. For all classifiers, the cross validated accuracy was slightly higher when the customer's contribution was isolated. This effect was statistically insignificant for all classifiers. Possibly this effect might be larger and even become significant when the models are more accurate in measuring actual sentiment. However, it seems more likely that accuracy scores in the customer subset are higher just because the transcriptions are shorter and thus even less data is available (in turn increasing susceptibility to overfitting even more).

5.2. AUDIO DATA

Whereas for textual data only non-generalizable effects could be found, for audio data it seems like no effects could be found whatsoever. No matter what methods were used, differences with the majority class baselines were minimal. These results were not a surprise either, as examination of means and standard deviations of the audio features (see section 3.2.2) suggests that differences in the distributions of these features between sentiment classes are very small. This was true for both comparison between highly positive and highly negative data, as well as comparison between neutral data and highly negative data. In addition, both data that was normalized and non-normalized data was tested, but here too no improvements could be found.

One possible explanation for this lack of results might be that the wrong types of features were extracted (i.e. these statistics are not informative regarding sentiment). However, even the features that have been shown to be informative regarding sentiment in previous research (e.g. Mairesse, Polifroni, & Di Fabbrizio, 2012), such as F0 statistics, were found to have very little differences between classes. Several possible explanations have been considered, such as the effect of normalization or the possibility of values being measured for audio input other than spoken language such as silence or background noise. However, enabling or disabling normalization did not make a difference and speech recognition software allowed for values only being measured on spoken language.

Two other possible explanations were found that have not been tested. The first being that telephone recordings are just too noisy and of too poor quality. This could have impaired the audio analysis software's performance. Another possible explanation is that the 30 second fragments are too short. For instance, negative sentiment can be measured by a change in pitch and intensity (Ververidis, & Kotropoulos, 2006). However, what if this change lasts several minutes? Then the actual moment of change is captured in only one fragment of 30 second, but not in all the others, while the others are still labeled as being negative. Disabling normalization could counter this effect as then more absolute values are measured, but this also causes the features to represent other absolute differences, such as differences between men and women.

Because no meaningful effects could be found for audio whatsoever, comparisons between diarization subsets does not inform us about the effect of diarization classification on sentiment analysis.

5.3. COMBINING AUDIO AND TEXTUAL DATA

The models that used a combination of audio and textual features were found to yield the highest performances, with the best performing model's accuracy at .947. These combined models were expected to score somewhere near, or higher than .913, as the predictions from the textual model that were added as a feature to these combined models alone already would yield a .913 accuracy. Although these increases could not be tested for significance (given the models did just produce a single prediction, rather than 10-fold cross validated predictions), they are considered substantial and therefore the method of combining textual and audio features is considered successful. This method was inspired by the method of Mairesse, Polifroni, & Di Fabbrizio (2012) and used because in their research it was also found to be successful.

Interestingly enough, the results on the text and audio combined models displayed a trend that countered the trend found in the models that only used textual data: for textual data, the customer subset performed slightly better for all classifiers, whereas for models that combined text and audio slightly higher results were obtained in the subset containing both contribution. These results were consistent throughout classifiers. However, differences are small and expected to be marginally significant at best. Because these small differences and because the lack of generalizability, comparisons between diarization subsets does not inform us about the effect of diarization classification on sentiment analysis.

5.4 ADDING THE AGENT'S TEXT FEATURE

Last, the models that added the agent's text feature yielded a nearly similar result to the equivalent model without this feature added. Adding predictions made by the best performing text-based model within the agent subset as a feature to the best performing model in the customer subset lead to a minor increase from

.941 to .945. Although this result implies that adding this text feature from the agent subset to the customer subset does not improve sentiment analysis, little can be said about its effect given the lack of generalizability, because none of the models were able to yield performances that were significantly higher than the majority class baselines in predicting previously unseen data.

5.5 FUTURE RESEARCH

Based on the results of the current research, one clear suggestion can be made for future research. Because the data was not able to provide features with strong and generalizable distinctive power regarding sentiment, a similar experiment could be performed on different data. A dataset is needed that is larger, and more polarized. Perhaps even acted data can be used in order to get an indication of if this effect of diarization classification even exists, to subsequently test it on real-world data.

Alternatively, given the fact that the dataset actually contained a lot of data (N = 217,219), just a small amount of annotations of class labels (N = 5,543), an attempt could be made to predict class labels for these unannotated data as pseudo-labels. When selecting only the predictions with a very high possibility and adding these predicted labels to the model as if they were true labels, the models might become more accurate. Methods for using unlabeled data to improve supervised learning are called semi-supervised learning and have been found to increase accuracy scores (Li, Ko, & Choi, 2018). Initial attempts to such methods have actually been made in the current research, but given the narrow timeframe of the project, no mentionable results have been reached.

6. CONCLUSION

In the current research, an attempt was made to test the effect of diarization classification on sentiment analysis performance of bilateral conversations. Specifically, customer service calls were used to examine whether isolating customer's contribution and analyzing it separately would lead to higher performances in text-based, audio-based, and multimodal sentiment analysis. A minor increase in performance was found for text-based sentiment analysis, when isolating the customer's, however differences were not statistically significant. In addition, none of the found effects were able to hold when predicting previously unseen data. In this concluding section, the research questions of this paper will be answered by either confirming or rejecting the accompanying hypotheses.

Research question 1: Can sentiment analysis performance of customer service calls be improved by isolating the customer's contribution, using diarization classification?

H1a: Isolating the customer's contribution in customer service calls, will increase performance of text sentiment analysis.

For text sentiment analysis, performances were higher when the customer's contribution was isolated, consistently throughout all classifiers. However, differences were not found to be significant, which implies hypothesis 1a has to be rejected. However, given the lack of generalizability, no real conclusions can be made regarding hypothesis 1a.

H1b: Isolating the customer's contribution in customer service calls, will increase performance of audio sentiment analysis.

For audio sentiment analysis, performances were not outperforming the majority class baselines whatsoever. The effect of diarization classification on sentiment analysis can only be tested if there is an effect to begin with. No conclusions can be made regarding hypothesis 1b.

H1c: Isolating the customer's contribution in customer service calls, will increase performance of multimodal sentiment analysis.

For multimodal sentiment analysis, performances were lower when the customer's contribution was isolated, consistently throughout all classifiers. This effect countered the hypothesized effect. However, this difference was not found to be significant. These results imply hypothesis 1c has to be rejected. However, given the lack of generalizability, no real conclusions can be made regarding hypothesis 1c.

Research question 2: Can the best performing set of features extracted from the isolated customer's contribution (as found in testing hypothesis 1) be further improved by adding textual features extracted from the isolated agent's contribution?

H2: Combining the best performing set of features extracted from the isolated customer's contribution with textual features extracted from the isolated agent's contribution, will increase sentiment analysis performance.

Combining the best performing set of features extracted from the isolated customer's contribution with textual features extracted from the isolated agent's contribution yielded a result that was not significantly different from the best performing set of features extracted from the isolated customer's contribution alone.

These results imply hypothesis 2 has to be rejected. However, given the lack of generalizability, no real conclusions can be made regarding hypothesis 2.

REFERENCES

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15-21.

Collomb, A., Costea, C., Joyeux, D., Hasan, O., & Brunie, L. (2014). A study and comparison of sentiment analysis methods for reputation evaluation. *Rapport de recherche RR-LIRIS-2014-002*.

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science*, *267*(5199), 843-848.

D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, *47*(3), 43.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., & Amir, N. (2007). The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In *International conference on affective computing and intelligent interaction* (pp. 488-500). Springer, Berlin, Heidelberg.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12).

Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012, December). Opinion mining and sentiment analysis on a twitter data stream. In *Advances in ICT for emerging regions (ICTer), 2012 International Conference on* (pp. 182-188). IEEE.

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013, October). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks* (pp. 27-38). ACM.

Govindarajan, M., & Romina, M. (2013). A Survey of Classification Methods and Applications for Sentiment Analysis. *The International Journal Of Engineering And Science (IJES)*, *2*(12), 11-15.

Hale, J. (2001, June). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-8). Association for Computational Linguistics.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, *44*(1), 1-12.

Li, Z., Ko, B., & Choi, H. (2018, January). Pseudo-Labeling Using Gaussian Process for Semi-Supervised Deep Learning. In *Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on* (pp. 263-269). IEEE.

Mairesse, F., Polifroni, J., & Di Fabbrizio, G. (2012, March). Can prosody inform sentiment analysis? experiments on short spoken reviews. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on* (pp. 5093-5096). IEEE.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, No. 1, pp. 41-48).

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

Moattar, M. H., & Homayounpour, M. M. (2012). A review on speaker diarization systems and approaches. *Speech Communication*, *54*(10), 1065-1103.

Munezero, M. D., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, *5*(2), 101-111.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., & Tewari, A. (2013). Learning with noisy labels. In *Advances in neural information processing systems* (pp. 1196-1204).

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Icwsm*, *11*(122-129), 1-2.

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

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *the Journal of the Acoustical Society of America*, *90*(6), 2956-2970.

Thayer, R.E., (1982). The Biopsychology of Mood and Arousal.Oxford University Press, New York: NY, USA.

Tranter, S. E., & Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*, *14*(5), 1557-1565.

Vaudable, C., & Devillers, L. (2012, March). Negative emotions detection as an indicator of dialogs quality in call centers. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on* (pp. 5109-5112). IEEE.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication*, *48*(9), 1162-1181.

Voutilainen, A. (2003). Part-of-speech tagging. *The Oxford handbook of computational linguistics*, 219-232.

Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *In ACM Computing Surveys(CSUR), 50*(2), 25.

APPENDIX A

Average values and standard deviations for all audio features, for sentiment label classes “a” and “e”:

		Avg(F0)	Std(F0)	Rng(F0)	Aad(F0)	Pkr(F0)
Class “a”	Avg	0.006894	0.905163	4.696261	0.666072	30.601369
	Std	0.417238	0.257916	1.525843	0.218498	18.540171
Class “e”	Avg	0.025898	0.914930	4.809985	0.670505	26.935897
	Std	0.354231	0.266935	1.643367	0.207585	16.251519
		Avg(A)	Std(A)	Rng(A)	Aad(A)	Pkr(A)
Class “a”	Avg	-0.062537	0.877544	3.425769	0.724972	34.002938
	Std	0.414257	0.249368	0.950681	0.225600	20.370573
Class “e”	Avg	-0.001430	0.864259	3.366258	0.714922	30.839225
	Std	0.364275	0.342811	1.347589	0.288883	17.773490
		Avg(ZCR)	Std(ZCR)	Rng(ZCR)	Aad(ZCR)	Pkr(ZCR)
Class “a”	Avg	0.001790	0.948191	4.692273	0.717245	36.770692
	Std	0.232538	0.194908	1.195690	0.158806	21.251638
Class “e”	Avg	-0.002198	0.948876	4.593753	0.728608	33.768603
	Std	0.227701	0.192407	1.119707	0.155932	19.112467
		Avg(E_ent)	Std(E_ent)	Rng(E_ent)	Aad(E_ent)	Pkr(E_ent)
Class “a”	Avg	-0.010735	0.953795	4.504060	0.729303	43.944398
	Std	0.257005	0.157109	0.985441	0.138637	25.606358
Class “e”	Avg	0.001556	0.956013	4.441188	0.731781	40.503606
	Std	0.201469	0.199802	1.002349	0.173284	23.129071
		Avg(S_cent)	Std(S_cent)	Rng(S_cent)	Aad(S_cent)	Pkr(S_cent)
Class “a”	Avg	-0.027222	0.938086	5.467679	0.657307	43.354195
	Std	0.239564	0.302475	2.395413	0.228632	25.567359
Class “e”	Avg	-0.002925	0.904597	5.210150	0.630897	39.693987
	Std	0.268082	0.360788	2.537104	0.263781	22.656107
		Avg(S_spr)	Std(S_spr)	Rng(S_spr)	Aad(S_spr)	Pkr(S_spr)
Class “a”	Avg	-0.012216	0.952702	4.775254	0.732844	37.779010
	Std	0.232675	0.172615	1.144956	0.146002	22.130070
Class “e”	Avg	-0.000084	0.949569	4.650661	0.740061	34.512094
	Std	0.248771	0.176279	1.087332	0.147492	19.692526

		Avg(S_ent)	Std(S_ent)	Rng(S_ent)	Aad(S_ent)	Pkr(S_ent)
Class “a”	Avg	NaN	NaN	NaN	NaN	NaN
	Std	NaN	NaN	NaN	NaN	NaN
Class “e”	Avg	NaN	NaN	NaN	NaN	NaN
	Std	NaN	NaN	NaN	NaN	NaN
		Avg(S_flx)	Std(S_flx)	Rng(S_flx)	Aad(S_flx)	Pkr(S_flx)
Class “a”	Avg	-0.034914	0.933100	5.265829	0.692887	43.126393
	Std	0.218264	0.213289	2.040694	0.139129	25.275624
Class “e”	Avg	-0.001382	0.946155	5.179558	0.704530	39.576340
	Std	0.230254	0.240126	2.039411	0.151865	22.564526
		Avg(S_rlf)	Std(S_rlf)	Rng(S_rlf)	Aad(S_rlf)	Pkr(S_rlf)
Class “a”	Avg	-0.004091	0.852300	5.173461	0.519042	36.801849
	Std	0.299507	0.409554	2.654962	0.257868	21.579162
Class “e”	Avg	0.001015	0.859036	5.082526	0.538281	33.878075
	Std	0.246720	0.440159	2.541470	0.316799	19.235431

APPENDIX B

List of words that were selected by the chi-squared selection method:

aanwezig	bedank	bel	bel ik	blokkeren	camera
cliënt	dankjewel	dat niet	de camera	de eerste	drucker
duizend negen	een telefoon	er op	even kijken	fijn	fijn weekend
geboren	gelukkig maar	gewijzigd	goed	goed ja	goed te
graag weten	groot zakelijke	haar zus	hadden	hahaha	half tien
hardstikke fijn	heb voor	heeft hij	heel fijn	heel graag	heen
hersteld	het adres	het echt	het er	het goed	het lopen
het moet	het nieuwe	het opgelost	het van	hi	hij de
hij is	hoe dat	hoe ik	hoef	honderd	hoop
houden	id	id kaart	iets anders	ik ben	ik heb
ik hoef	ik hoop	ik je	ik net	in me	in mijn
in voor	ingevoerd	inloggen	is inderdaad	is nul	ja heel
ja hoe	ja klopt	ja wat	ja zeker	ja die	je moet
je weet	je wel	juni	jullie	jus	kaart
kan dan	kat	ken	ken je	kennen	kiezen
kijken ga	klantnummer	korte	laptop	las	lijn
lisanne	lopen en	maar door	maar wel	macht	mee te
met de	moet hebben	money	nat	negen	negen dat
negenentwintig	no	no no	nog in	nou hardstikke	november het
nu nu	nul	nul negen	nul een	nummer	nummer negenentwintig
of	of die	of een	of het	of wij	ok
ok ok	ongeluk	ook in	op mijn	op nul	opgelost
opstellen	over negen	overgemaakt	overgemaakt maar		paul
paul pierce	pierce	pijn	plan	prima	product wat
redelijk	redelijk in	rekening en	saldo	schade	shit
snap	speciaal	super	telefonie	tien	tien november
toen	toestel	toestel is	toesturen	trouwens	twee nul
uit als	uit er	van je	vandaag nog	verkeerd	verlenging
vier nul	vijf over	vijfentwintigste	vijftien	voelt	voor mijn
voor toestel	vooraf	voortaan	wat hebben	wat wij	weer goed
wel goed	weten wie	wij in	wilt	zakelijke	zes twee
zetten	zie het	zit al	zo dat	zo nat	zodat
zullen	zus	een	een drie	een nul	