Universiteit Utrecht

Bachelor Thesis Artificial Intelligence

Classifying Elderly Emotion

Tobias Cadée
5550904

7.5 ECTS

Supervisor: Dr. Heysem Kaya
Second reader: Dr. Tejaswini Deoskar

April 21, 2020
Words: 6633

Abstract

Classifying human emotion has been a research field of interest for quite some time now. This research will add to existing work on the subject, and look into the classification of emotion of elderly people. As a part of the 2020 ComParE paralinguistics sub-challenge on elderly emotion, I will conduct experiments with different machine learning models, feature representations, principle component analysis and feature level fusion, to find which combination leads to the best results on this classification task. The support vector machine classifier shows promising results on the arousal classification task, and feature level fusion also shows promising results on the valence classification task. The application of principle component analysis shows mixed results, some positive and some negative. Other researchers, as part of the team working on the sub-challenge, might benefit from these results, but should be cautious to apply them in all cases.

Index terms: Computational Paralinguistics, Challenge, Elderly Emotion, Emotion Classification, Support Vector Machine, Gradient Boosting Machine, Principle Component Analysis

Contents

Chapter 1

Introduction

Classifying human emotion[1] has been a field of study for some time now. It has received more attention over the past twenty years because of the increase in plausible application areas, which are numerous by now. Human computer interaction is one field of study where a lot could be gained with better human emotion classification (HEC). One could develop chatbots that are equipped with emotion recognition software to be able to give better replies when customers are expressing certain emotions. For a complete overview of the applications of HEC and relevant research see work by Saif M. Mohammad (Mohammad, 2016). This meta-study gives an overview of the research of emotion classification and sentiment analysis from textual sources. I will now move on to a description of HEC and what choices have to be made for it to perform correctly.

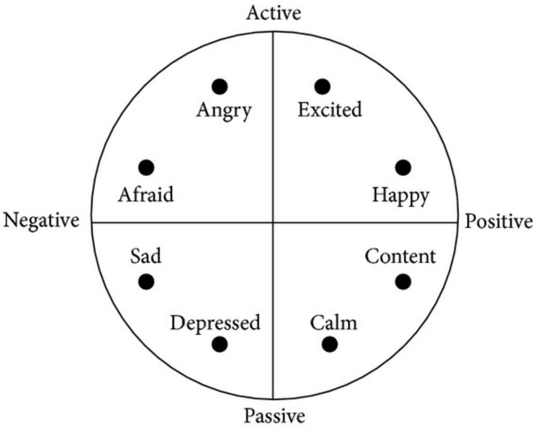


Figure 1.1: *Valence-arousal dimensional model* (Jirayucharoensak, Suwicha & Pan-Ngum, 2014).

The task itself is quite a complex one, as it is not easy to classify human emotion (Ortony, Collins, & Clore, 1988). If you would want to build a system that classifies emotion the first question you would have to ask is: ”Which kinds of emotions do I distinguish?”. In 1957, Osgood, Suci and Tannenbaum showed that the best measures when it comes to distinguishing emotion were potency (positive and negative), activity (active, passive) and evaluation (good, bad)(Osgood & Tannenbaum, 1957). Later, Russell developed a new model that has two axes; valence and arousal (see Figure 1) ((Russell, 1980)). Valence is closely linked to evaluation and refers to the goodness of emotion, in the figure noted as negative and positive. For instance, joy has positive valence. Arousal on the other hand refers to the energy of the emotion, which corresponds roughly to the activity mentioned above. In the figure arousal is referred to as passive and active. By combining these two measures you can explain other emotions like being angry, which is shows in the figure as having high activity (arousal) and is on the negative side of the valence spectrum. The valence-arousal model will be used in this thesis for HEC.

There are multiple ways to accomplish sentiment analysis or HEC. As early as 1968 there were scientists who proposed ways to quantify patterns in texts (Gilman, 1968). Nowadays, state of the art systems are mostly machine learning models as they excel in finding

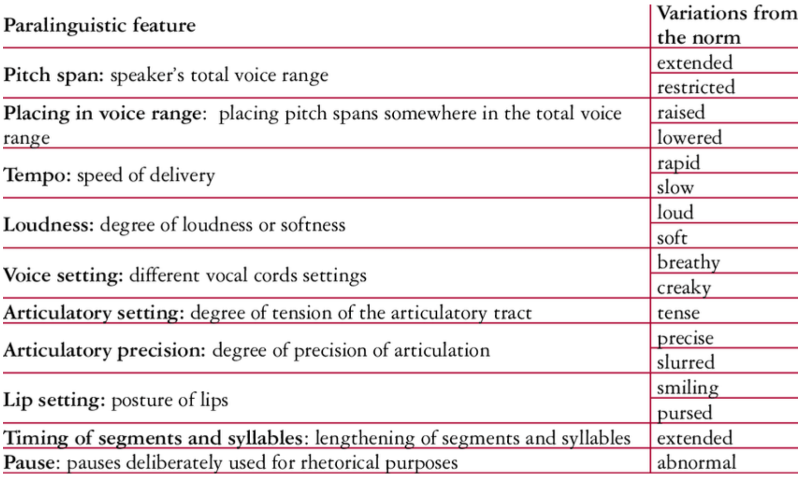
[1] When I say ”Classifying human emotion” I mean automated sentiment analysis, not the psychology field that deals with the classification of human emotions.

patterns if there is a sufficient amount of data available. This will also be my method in doing sentiment analysis. The question is, what type of data do we present to these models?

A much used type, which is also the subject of the study mentioned above by Saif M. Mohammad is 'text'. There is a hurdle when it comes to using text as data, it becomes almost impossible to measure arousal. Since you can not easily infer the activity of the person who has written the text from words alone. Other types of data used for HEC could be video recordings of people where you infer emotion based on their face (Padgett & Cottrell, 1996). Another method would be to use spoken words (audio), on which I will elaborate next.

When one would want to use audio data for an emotion classification problem, two points of attention have to be mentioned first. The first point of attention being the way data is labelled. Are its labels generated using actors which try their best at faking every emotion, or is the data labelled in some other way? Audio files could also be labelled afterwards and experts could also listen at the files as well. The second point of attention when using audio input is your feature representation. There are a number of different feature representations to chose from, on which I will elaborate later in chapter 2. First I will look into whether one look at sentiment analysis from a linguistics or paralinguistics angle.

The linguistic research domain is well known, and is about the meaning and structure of the spoken word. Paralinguistics is less known and therefore deserves its own introduction.



Paralinguistic feature	Variations from the norm
Pitch span: speaker's total voice range	extended restricted
Placing in voice range: placing pitch spans somewhere in the total voice range	raised lowered
Tempo: speed of delivery	rapid slow
Loudness: degree of loudness or softness	loud soft
Voice setting: different vocal cords settings	breathy creaky
Articulatory setting: degree of tension of the articulatory tract	tense
Articulatory precision: degree of precision of articulation	precise slurred
Lip setting: posture of lips	smiling pursed
Timing of segments and syllables: lengthening of segments and syllables	extended
Pause: pauses deliberately used for rhetorical purposes	abnormal

Figure 1.2: *Taxonomy of paralinguistic features.*
(Bombelli, Griselda & Soler, 2013)

I was unfamiliar with paralinguistics when I started this research but soon noticed that with the rise of Artificial Intelligence (AI) and its various applications it is a field we should all be acquainted with. Linguistics focusses on words and their meaning, for instance constructing models based on word frequency or connotation, whereas paralinguistics is all about speech 'beyond' words. Sometimes paraphrased as the aspects of spoken language that do not have to do with the actual words being spoken. For an example of paralinguistic features of speech, see Figure 1.2. Following the introduction of the field of research into emotion recognition, types of data you could use and the paralinguistic domain, I will now introduce my own research.

I was part of a research team working on the annual Interspeech Computational Paralinguistics ChallengE (ComParE). This challenge is part of the Interspeech conference which is the largest conference on speech processing and application [2]. Each year a number of sub-challenges are released ranging from detecting Parkinson's decease to classifying the type of snoring of a person(Kaya & Karpov, 2017; Grósz, Busa-Fekete, Gosztolya, & Tóth, 2015). I will be doing research on one of the sub-challenges, namely the one on classifying elderly emotion[3]. For detailed information about this challenge, read the following paper(B. W. Schuller et al., 2020).

With the average age of our population steadily increasing (Garssen, 2011) a greater need has arisen to be able to monitor the (mental) health of the elderly. This is where the challenge steps in. If it would be possible to classify the emotional state of the elderly, based on audio, you could have systems in place which could respond appropriately to individual persons. This could have positive effects on their mental health (Luxton, 2014; Kenny, Parsons, Gratch, & Rizzo, 2008).

My input for the challenge will be to compare two different machine learning models, namely Support Vector Machines (SVM) and Gradient Boosting Machines (GBM). I will motivate my chose for these models in the next chapter. Furthermore, I will look at which feature set performs best on the arousal classification task, and which on the valence classification task. I will also look at the effect of Principal Component Analysis on the results. Finally I will explore the effects of combining the best performing feature sets to hopefully improve the score[4]. Both tasks are not explicit classification tasks, as the labels could be continuous values. The reason that this is a three class classification problem in this case, is that the data is labelled in this way.

1.1 Overview

In Chapter 2 I will provide background, starting with earlier research into this task done within the paralinguistic field and on the corpus provided with the challenge. After this I will go into more detail about possible feature representations. And finally I will provide information about how the models I will use in my research work. Chapter three is on the methodology, I will go into more detail on how and which experiments will be conducted, including the machine learning pipeline. Chapter four will give the results followed by a conclusion of the results in chapter five. Finally, in chapter six I will reflect on my research and make suggestions for future research. This last chapter is of great importance as my research is part of the challenge on which more researchers are working on at this moment.

[2] http://www.interspeech2020.org/

[3] http://www.compare.openaudio.eu/compare2020/

[4] By improving score I mean to improve the emotion classification

Chapter 2

Background

2.1 Classifying Emotion

There has been a great deal of earlier research done in the field of HEC within paralinguistics. I will give an overview of the research done.

One of the earlier studies dealing with this topic was done in 1998 on the transcription of prosodic and paralinguistic features of emotional speech (Roach, Stibbard, Osborne, Arnfield, & Setter, 1998). The researchers do not attempt any kind of classification themselves, but do create on of the first feature representations when it comes to classifying emotion using paralinguistics. Later in 2006, more research was done on a corpus containing spoken dialogues from a medical emergency centre, which show more extreme emotions (Devillers & Vidrascu, 2006). Relief, Anger, Fear and Sadness were used here, opposed to the valence-arousal model I work with. These four types of emotion are part of the six basic emotions defined by Eckman (1999). Although the linguistic cues gave better emotion detection overall, the paralinguistic cues were also quite good. Similar research was done on non-artificial data one year later (Vidrascu & Devillers, 2007). Other interesting research, of the more applicative sort, was done on detecting emotion in older car drivers, and on designing an ”emotionally responsive car” (Jones & Jonsson, 2008). This is in a way similar to my research as it also applies to elderly people. One of the leading researchers within the field of paralinguistics, and responsible for coming up with new challenges every year is Björn Schuller. I will mention some of his work, concerning developments in the field of paralinguistics and especially on emotion classification. In 2008, B. Schuller et al. developed a new method to automatically segment data to overcome statistical noise generated by outliers. This technique showed good results on emotion classification. Participating and eventually winning in the 2009 challenge on emotion was research done by Dumouchel et al. on different feature representations for emotion recognition (Dumouchel, Dehak, Attabi, Dehak, & Boufaden, 2009). More research, co-authored by Schuller, was done in 2012, this time developing a new method to deal with data sparsity. The method makes use of co-training to avoid the need for intensive human labelling. This method showed good results when compared to the baseline system on classifying emotion(Z. Zhang, Deng, & Schuller, 2013). Furthermore, the winners of the 2013 sub-challenge on emotion proposed to use Ada-boost, which is a boosting machine learning algorithm that combines many weak learners into a boosted classifier (Gosztolya, Busa-Fekete, & Tóth, 2013). Gradient boosting machines are also an example of a boosted classifier, but more on that later at the end of this chapter.

Additionally to research specifically on emotion classification, is research giving an overview of the current paralinguistic landscape. Schuller (2012) described ten trends in paralinguistics. Adding to this is research from one year later in 2013, where the field of paralinguistics is defined properly and which describes common practices in paralinguistics (B. Schuller et al., 2013).

2.2 Corpus

The corpus which was supplied to researchers participating in the elderly emotion sub-challenge is the Ulm State-of-Mind in Speech elderly (USoMs-e) corpus. This is the first data set released for scientific purposes containing elderly speech annotated for emotion recognition. The data consists of 352 narratives, from 88 participants reporting two positive and two negative records. Arousal and valence was assessed after each narrative. For both valence and arousal there are three possible labels; Low, Medium and High. These will also be the labels for my classification. The audio files are chunked for feature extraction. These chunks also have labels, corresponding to the label of the original file. The distribution of the classes over the training and development data set are shown in Table 2.1 and 2.2. The development set is used for testing in the experiments. There is a test set available with the corpus but the labels are hidden, this will be explained in the discussion.

	Before Chunking		After Chunking	
	Arousal	Valence	Arousal	Valence
Total	87	87	2496	2496
L	24	14	1021	383
M	28	42	763	1228
H	25	31	712	885

Table 2.1: Distribution of classes for the training data set.

	Before Chunking		After Chunking	
	Arousal	Valence	Arousal	Valence
Total	87	87	2466	2466
L	39	18	1136	428
M	29	49	822	1496
H	19	20	508	542

Table 2.2: Distribution of classes for the development data set.

2.3 Feature Representation

Choosing the right feature representation is key when it comes to paralinguistics. All participants for the 2020 challenges were supplied with a number of baseline feature representations on which I will elaborate in this section. The first are all acoustic representations, which mean that they fall within the scope of the paralinguistic domain. The other features are linguistic features that do rely on transcription of speech. These features might be useful in classifying valence, as acoustic features tend to fall short on this task. The acoustic features are all generated per short chunk of the whole audio input of approximately three seconds length each. The linguistic features are generated per audio file.

2.3.1 ComParE

The ComParE feature set is the standard feature set for the annual challenge. It consists of the functionals of low-level descriptors (LLD's) and their delta's. Low-level descriptors can for instance be the energy, pitch or voicing probability. The functionals are functions applied to these LLD's like minimum, maximum and mean. For a detailed explanation on these descriptors and how they are retrieved see B. Schuller (2011). The ComParE set is generated using the openSMILE toolkit (Eyben, Weninger, Gross, & Schuller, 2013).

2.3.2 Bag of Audio Words

The Bag of Audio Words (BoAW) feature set is similar to the Bag of Words technique used in linguistics (Y. Zhang, Jin, & Zhou, 2010). The key difference is that the word occurrences cannot be counted, as you are dealing with paralinguistics. So, the BoAW representation uses the LLD's to generate term-frequency histograms of short intervals of audio data. In other words. For every short audio fragment the unique audio words are computed, which are audio pieces that are alike in their LLD's. This is done via K-means clustering from the acoustic LLD's of the training data set. Next an array is created for each audio fragment containing the occurrences per audio word. The BoAW representation can vary in length as the quantity of audio words can be chosen. The BoAW features are created with the openXBOW toolkit (Schmitt & Schuller, 2017).

2.3.3 auDeep

AuDeep features are constructed using the auDeep toolkit for Python (Amiriparian, Freitag, Cummins, & Schuller, 2017; Freitag, Amiriparian, Pugachevskiy, Cummins, & Schuller, 2017). The audio spectrograms are fed to a sequence to sequence deep recurrent auto-encoder which can then be used to create features. The auDeep toolkit can generate different kinds of feature sets of a set length. The feature set that is generated depends on the number of frequency bands the auto-encoder uses. This number is notated in the feature set name for transparency.

2.3.4 DeepSpectrum

DeepSpectrum features are constructed using a Python toolkit as well (Amiriparian, Gerczuk, et al., 2017). This toolkit provides a pipeline to first create visual representations of the audio data that are then fed to a pre-trained image Convolutional Neural Network. The final layer is then used to create the feature vectors. Both the auDeep and DeepSpectrum features are hard to interpret by looking at the feature vectors alone because they have already been through a neural network as opposed to the ComParE and BoAW features.

2.3.5 Linguistic Features

New to this years challenge are the linguistic baseline features. They are constructed using a Python pipeline to extract linguistic features from German audio data (Irie, Zeyer, Schlüter, & Ney, 2019). This pipeline uses a language model developed by researchers at Google in 2018 called BERT (Bidirectional Encoder Representations from Transformers). The features are word embeddings, they predict the chance of a word given its context (Devlin, Chang, Lee, & Toutanova, 2018). BERT uses a bidirectional transformer to learn these contextual relations between words. It is bidirectional because it looks at all of the words at the same time, not just from left to right or the other way around. After learning the relations between words, using a neural network, BERT can be used to create linguistic feature vectors. It would be possible to create different linguistic features with another tools, although the BERT approach is currently the state-of-the-art.

2.4 Machine Learning Models

2.4.1 Support Vector Machines

Support Vector Machines (SVM) are part of supervised learning algorithms. The standard SVM is a binary classifier, which creates a hyperplane, which is a line in two dimensions, to separate the input data into two classes, usually -1 and +1, but it can be any two classes. The goal of the algorithm will be to find the optimal hyperplane to separate the data. There could be many hyperplanes that lead to the same classification, but only one hyperplane which is the optimal solution. This is the hyperplane with the greatest distance to the nearest points on both sides of the hyperplane. These points are called the support vectors (hence the name). For an illustration of two hyperplanes of which one is optimal, see Figure 2.1

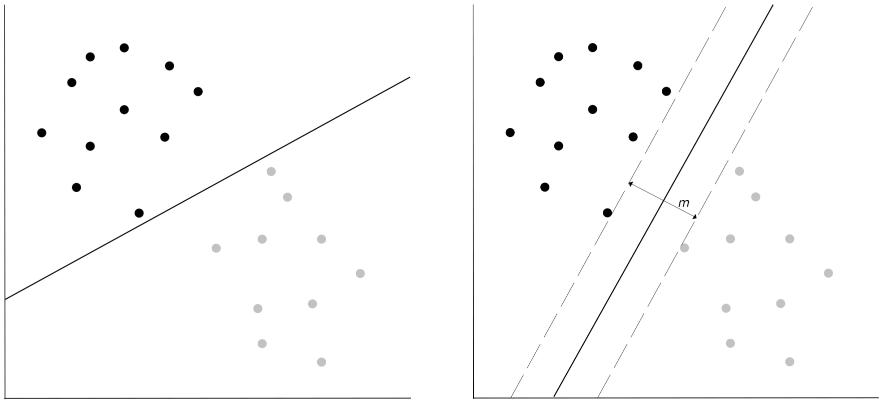


Figure 2.1: *Two hyperplanes in the same linearly separable data.* The two lines separate the data perfectly, but the right one does so while maintaining the maximum distance from both sides.

To find the optimal hyperplane, the algorithm has to find the weight vector w for the hyperplane. When the data is linearly separable, the following should hold for every correct classification of a data point x: $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) - 1 \geq 0$ (for any $i = 1, .., n$). The margin can be mathematically expressed as $\frac{2}{||\mathbf{w}||}$, this margin is shown in figure 2.1 in the right figure as m. The optimal hyperplane has the largest margin, so m needs to be maximized. This means that $||\mathbf{w}||$ has to be minimized. The final optimization problem can now be defined as follows:

$$\begin{gathered}\text{Minimize } ||\mathbf{w}|| \\ \text{subject to } y_i(\mathbf{w} \cdot \mathbf{x_i} + b) - 1 \geq 0 \\ \text{for any } i = 1, .., n\end{gathered}$$

To deal with non-separable data, a soft margin rather than a hard margin can be used. This means that some percentage of the training data is allowed to be misclassified. The error from the soft margin is penalized by C. If a larger value for C is chosen, the classifier will try to separate the instances as best as possible, but the distance from the support vectors may be smaller. As a consequence the dividing line will be less optimal. Finding the right value for C is important when using SVM's. This becomes even more essential when the number of features is greater than the number of samples.

A quality of SVM's is that it is still effective in high dimensional feature spaces, as well as if the number of dimensions is greater than the number of samples. This can be considered motivation for using this model in my research. There are also different kernels available for using the SVM model, for this thesis I will be using the Linear SVM model that tries to linearly separate the data, after applying the kernel function.

SVM models can be trained on more than two classes. In this case a combining strategy will have to be used. The linear SVM model that I will be using, uses the one-vs-the-rest strategy. In this strategy the model is run for every class. And every data point is classified as being this class or not. This is then repeated for the number classes there are. This strategy is illustrated below in Figure 2.2.

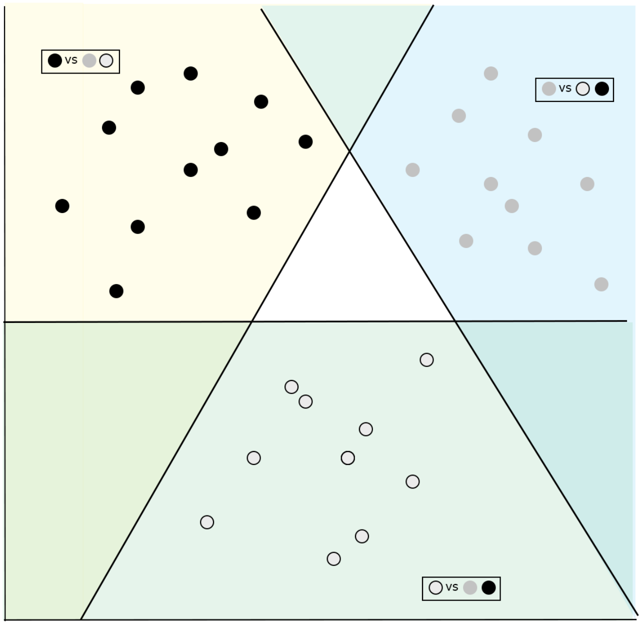


Figure 2.2: *One-versus-rest multi-label support vector machine.*

2.4.2 Gradient Boosting

Gradient boosting is an ensemble learning method that was first formulated in 1997 by Leo Breiman (Breiman, 1997). He based his research on other work by Yoav Freund and Robert Schapire, who invented the AdaBoost algorithm two years earlier (Freund, Schapire, & Abe, 1999). The idea of an ensemble learning model as opposed to a kernel learning method is that it combines weak classifiers, like single decision tree's, into a strong classifier. Within this class of machine learning algorithms bagging and boosting algorithms are distinguished. It should come as no surprise that gradient boosting is of the latter kind. Bagging methods combine the weak learners in parallel, whereas boosting methods combine them sequentially, updating the weights after each iteration. Both algorithms are good at reducing variance and provide stability. Bagging algorithms also take care of over-fitting, where boosting algorithms can still over-fit the training data. On the other hand, boosting algorithms also try to reduce bias where bagging algorithms do not. I will now examine how the gradient boosting machine works.

The goal of the gradient boosting algorithm is to minimize the loss function L. This can be any loss function. The loss function the distance from the current to the final predictions. In the first iteration the model trains a single decision tree on the data, while minimizing this loss function. The result of this can be mathematically expressed as:

$$F_0(x)$$

After this first step, the algorithm repeats the following steps in each iteration: Firstly, the gradient of L is calculated. This gradient is needed because the algorithm makes use of gradient descent to be able to minimize loss functions which can not be minimized directly. This gradient returns the pseudo residuals. This is the distance to the goal function, which has to be minimized as stated before. These residuals are calculated as follows:

$$r_i m = -[\tfrac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)} \text{ for } i = 1, .., n.$$

Where $m \in M$ is the iteration, and x is the input vector.

Then, a new tree is trained on the residuals. The leafs of this tree will produce an average gradient. The algorithm then takes a metaphorical step of the size of the residual, which is multiplied with a learning rate parameter. This is one of the parameters that will be tuned in experimentation, it defines how fast the model converges towards the goal function. This step is mathematically expressed as follows:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Where $h_m(x)$ is the new tree that has been learned on the residuals,
and γ_m is the step size.

After the final iteration the model returns a decision tree which can then be used for classification. The number of iterations is another important parameter to tune. Too many iterations can cause the algorithm to over-fit, and too few iterations will return a model which has not been fitted enough. The number of iterations will also be tuned in the experiments. The final parameter that will be tuned is the maximum depth of each new decision tree.

2.5 Principal Components Analysis

Principal Components Analysis (PCA) is a statistical procedure to reduce the dimensionality of the feature space. This is especially useful in the case of acoustic features, because they tend to be high dimensional while the training data size can be limited. This can lead to over-fitting. PCA was invented in 1901 by Karl Pearlson (1901). The idea behind PCA is to find the components in the data that explain the largest portion of variance. The procedure firstly finds the first component with the highest explained variance, then iterates until all of the variance is explained. These components are uncorrelated, which is a good attribute for training. This is illustrated in Figure 2.3, where you see the first two principal components of a random data set as arrows.

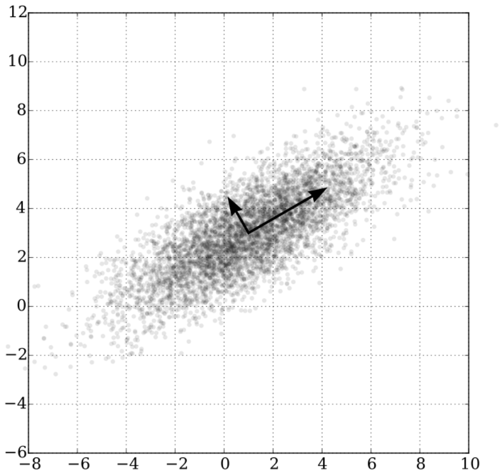


Figure 2.3: *The first two principal components of a data set.*

Chapter 3

Methods

The goal of my thesis will be to find out which machine learning model, together with which kind of feature set, leads to the best results on the data set that has been supplied with the challenge. In addition to this research question I will also experiment with the application of PCA on each feature set to see if this leads to a score improvement. Finally I will conduct experiments on combinations of best performing feature sets to see if this will boost performance.

Although the arousal and valence classifications are two separate tasks, I will conduct the same experiments for both classifications. These experiments will follow a pipeline that I have constructed to visualize my process and to maintain consistency throughout my research. This pipeline is shown in Figure 3.1. In this chapter I will go into detail about the methodology of my experiments. The following sections will correspond to the categories in the pipeline figure.

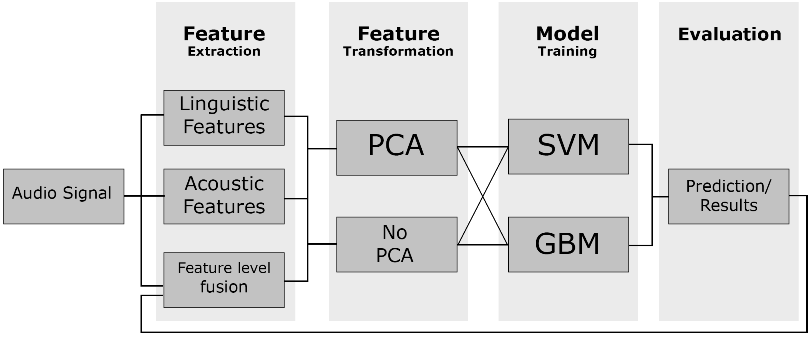


Figure 3.1: *Machine learning pipeline.*

3.1 Feature Extraction

The first task in the machine learning pipeline is to select the features on which the model will be trained. These are the features mentioned in the previous chapter, they will be mentioned again for clarity. The linguistic features are a set of four feature sets supplied with the challenge, called frozen-bert-gmax, frozen-bert-rnnatt, frozen-bert-pos-fuse-rnnatt and frozen-bert-fused. The acoustic features are ComParE, BoAW-125, BoAW-250, BoAW-500, BoAW-1000, BoAW-2000[1], auDeep-30, auDeep-45, auDeep-60, auDeep-75, auDeep-

[1] The numbers correspond to dimension of the feature vector.

fused [2] and DeepSpectrum. These features are all split up into training and test partitions. After evaluation, the four best acoustic features are selected and then combined to pairs and triples by ways of concatenation. This is called feature level fusion. These combined features then follow the same path down the pipeline to see what results they yield. A more detailed description on evaluation will be discussed at the end of this chapter.

3.2 Feature Transformation

After feature extraction there are two options when to comes to feature transformation. The first is to apply PCA to the feature set and the other is to leave the feature set intact. The feature transformation step in the machine learning pipeline is essential to see what the effects of PCA are on the results. A PCA model is fit on the training data and then both the training data and the test data are transformed, while keeping 95% of the variance of the original data set explained, but greatly reducing the feature space. This could have positive effects on the score.

3.3 Model Training

3.3.1 SVM

A linear SVM classifier from the scikit-learn Python package was used for this research (Pedregosa et al., 2011). The C parameter was tuned by repeatedly running the model with a different value for C on the same dataset. The C parameter was chosen from the following values: [0.00001, 0.0001, 0.001, 0.01, 0.1, 1].

3.3.2 GBM

A GBM classifier from the scikit-learn Python package was used for this research. The learning rate was tuned, choosing from the following values: [0.3, 0.2, 0.15, 0.1, 0.05, 0.01]. Subsequently the number of estimators was tuned, choosing from the following values: [25, 100, 200, 500, 750, 1000, 1500, 2000]. Finally, the max depth was tuned, choosing from the following values: [2,3,4,5,6,7].

3.4 Evaluation

After running the machine learning model on the training data, predictions are cast on the test data. The score subsequently used to measure the performance of a feature set/model combination is calculated with the Unweighted Average Recall (UAR). This error measure was set for this challenge, so choosing an alternative error measure while doing experiments would lead to inconsistency. UAR is calculated by taking the recall (the percentage of relevant classifications) of every label and then calculating the mean of these values.

[2] Here the numbers corresponds to the number of frequency bands given to the auto-encoder. This was mentioned in the previous chapter.

Chapter 4

Results

The results are summarized to give an overview of the most important comparisons, because of the large quantity of results gathered. Appendix A shows the full results of the arousal classification task, Appendix B shows the full results for the valence classification task.

The results chapter is divided up into two sections. The first section covers thearousal classification task, and the second section covers valence classification. While the experiments for both classification tasks were done in the same manner, the results are different so they should be treated separately. For both classification tasks I will present my results in the same manner.

Firstly, I will show the results of using the SVM classifier with and without the application of PCA. Secondly, I will show the same results but then using the GBM classifier. The third part will be a comparison of using SVM or GBM, where the score is chosen as the the highest score obtained using PCA and not using PCA. Finally, I will present my results on using the linguistic features. I would like to point out that the y-axis of the figures are not consistent across all figures.

4.1 Arousal Classification

4.1.1 Acoustic Features

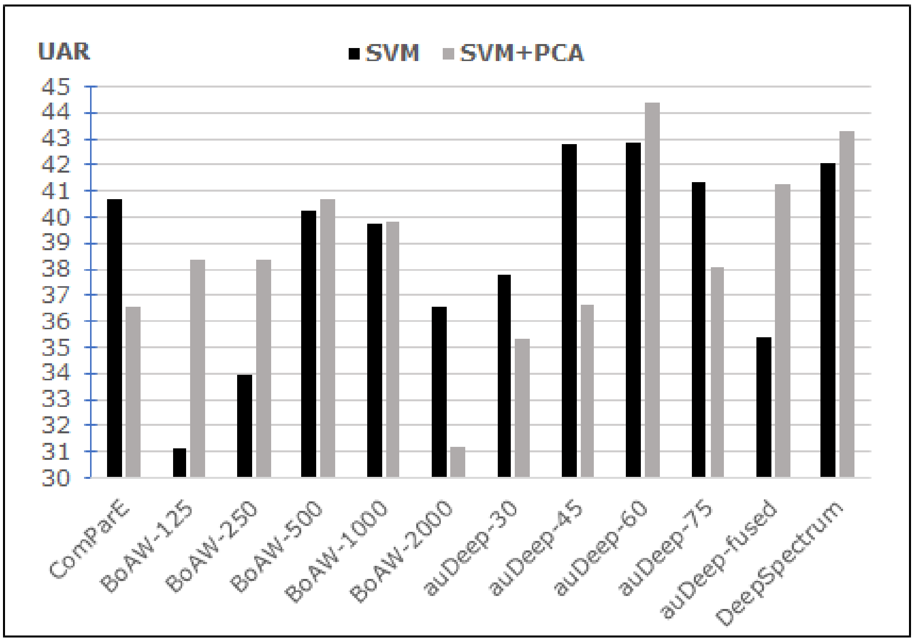


Figure 4.1: *Arousal classification using a SVM classifier with acoustic features.*

Figure 4.1 illustrates the results of using SVM for the arousal classification task, and additionally the effect of applying PCA on each feature set. Scores vary a lot between feature sets and also between using PCA or not. The best performing feature sets when using SVM are auDeep-60 and DeepSpectrum with respectively 44.41% and 43.27% UAR. In both of these cases PCA transformation had a positive effect on the score, with an increase of 1.56 and 1.18 percent.

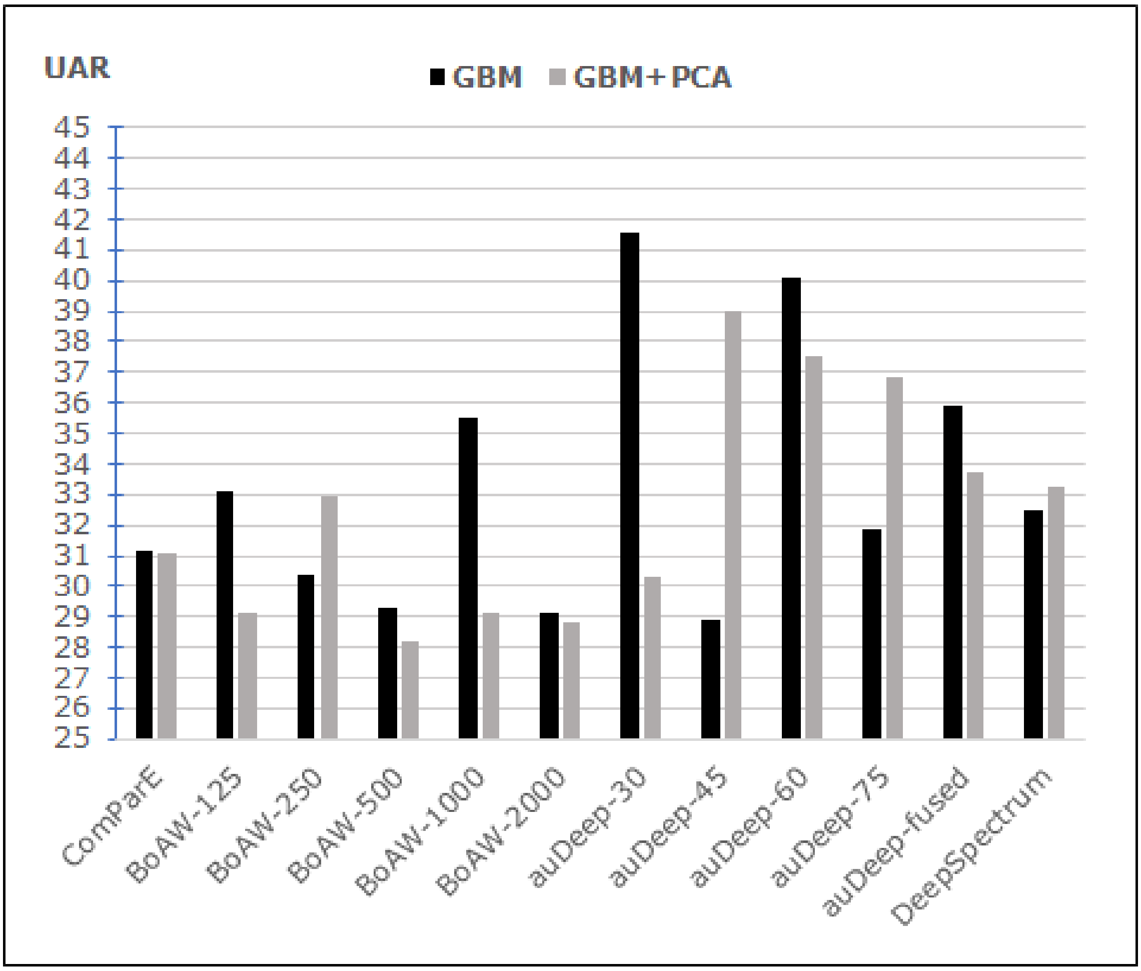


Figure 4.2: *Arousal classification using a GBM classifier with acoustic features.*

Figure 4.2 illustrates the results of using GBM for the arousal classification task, and additionally the effect of applying PCA on each feature set. As was the case with SVM (Figure 8); scores vary a lot between feature sets and also between using PCA or not. In some cases the application of PCA had a significant positive effect, for instance with auDeep-45, where the score increased with 10.10%. On the other hand, PCA also had significant negative effects on the obtained score in some cases. For auDeep-30, the score dropped with 11.24% after applying PCA. The best performing feature sets when using SVM are auDeep-30 and auDeep-60 with respectively 40.08% and 43.27% UAR. PCA had a negative effect on the scores for these feature sets.

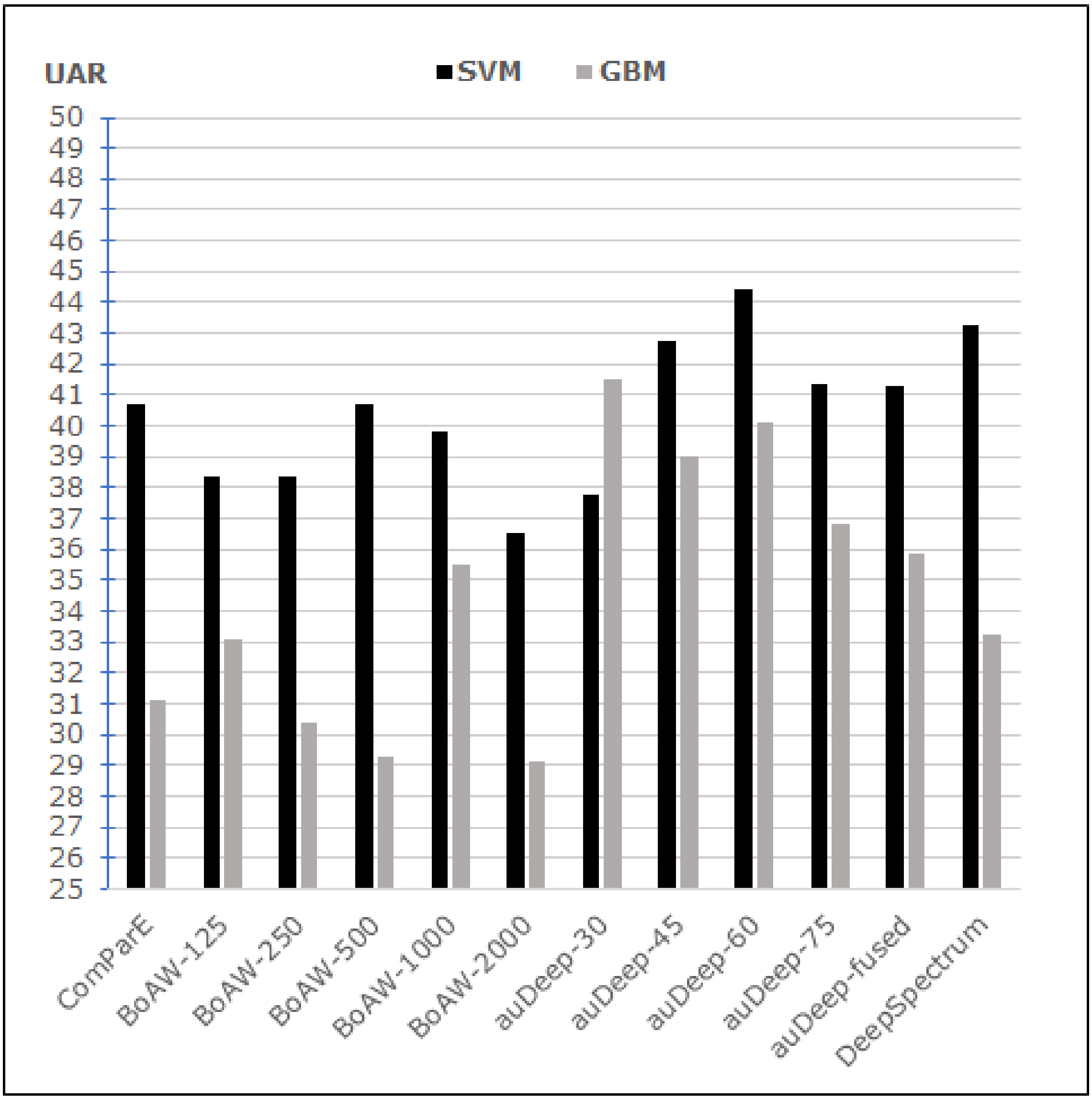


Figure 4.3: *Comparing SVM with GBM for the arousal classification task, using acoustic features.*
Note that the score for each feature set is the maximum of applying PCA or leaving the feature set intact.

Figure 4.3 illustrates the difference of scores between using SVM and GBM, using acoustic features. The SVM classifier produced better results with almost every feature set, except for auDeep-30, where the GBM classifier yielded better results. The difference between using SVM and GBM is significant, as the mean score for SVM is 40.44% UAR and the mean score for GBM is 34.81% UAR. The best scoring feature sets are the same as mentioned above; auDeep-60 and DeepSpectrum. These are also the best scoring acoustic feature sets for the arousal classification task.

4.1.2 Linguistic Features

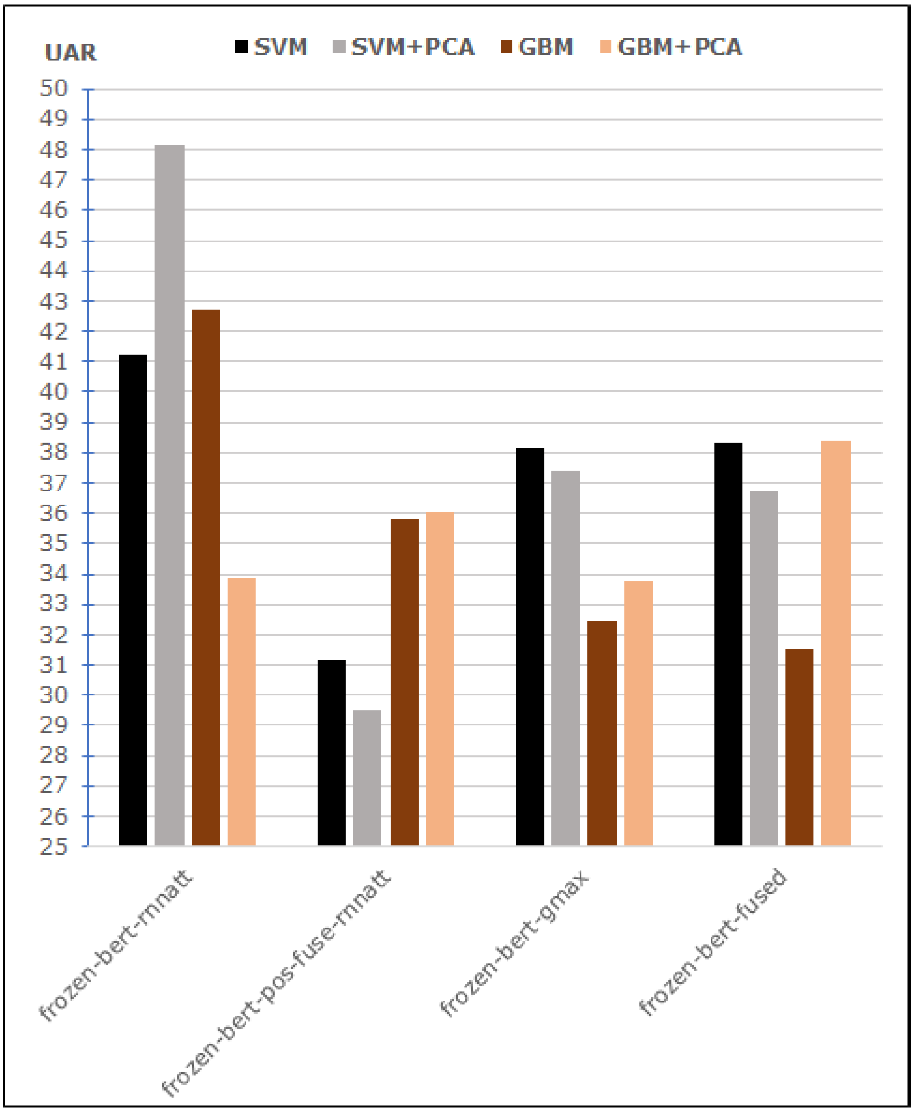


Figure 4.4: *Comparing SVM with GBM for the arousal classification task, using linguistic features.*

Figure 4.4 shows the results of using the linguistic feature sets for the arousal classification task. The frozen-bert-rnnat feature set scores are the highest, both in combination with SVM and GBM with respectively 48.16% and 42.73% UAR. The application of PCA had a positive effect on the score of the SVM classifier, as it increased with 6.92%. The opposite effect was found for GBM. The application of PCA decreased the score with 8.87%.

4.1.3 Feature Level Fusion Features

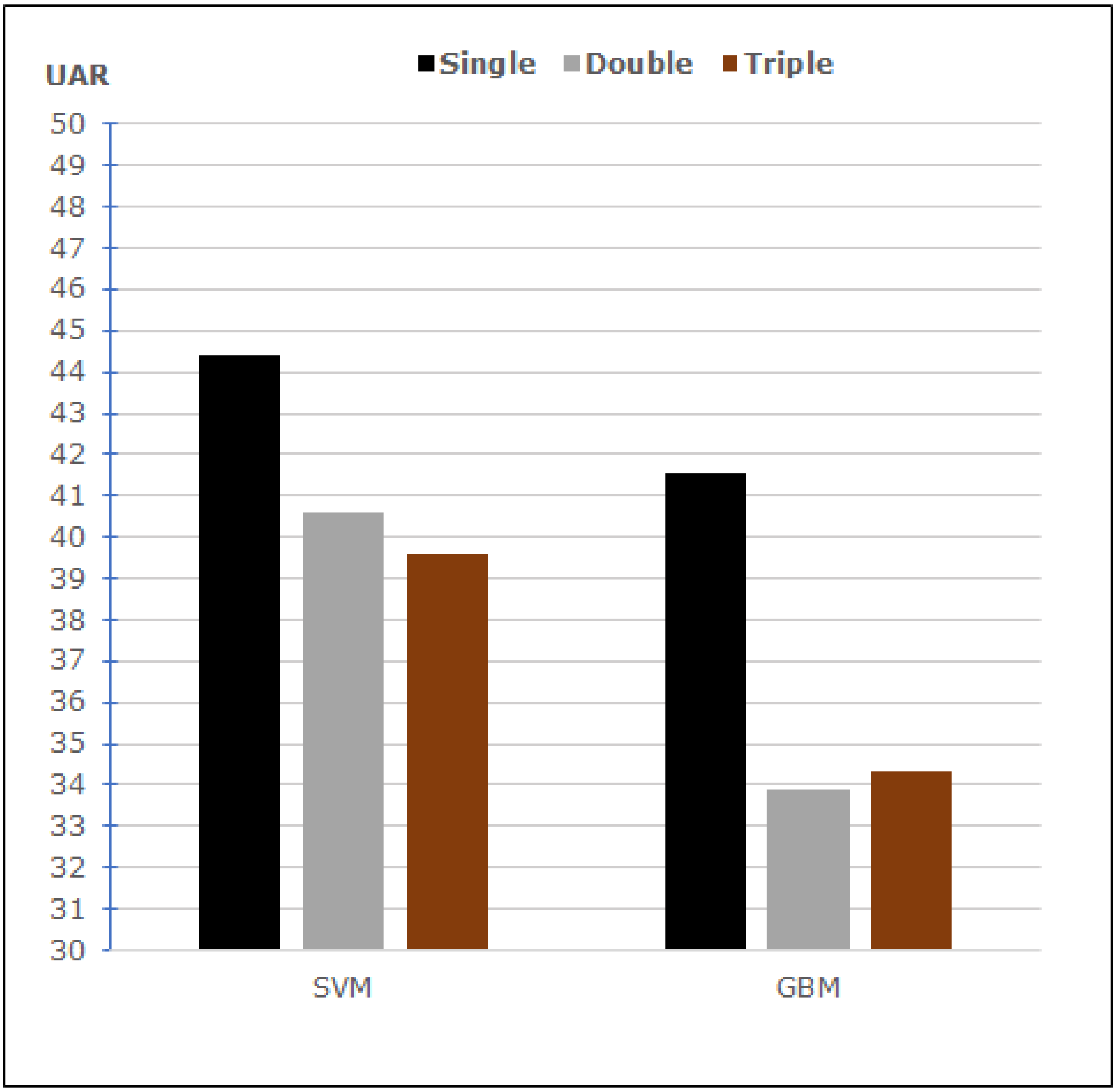


Figure 4.5: *Feature level fusion results for the arousal classification task.*
The best performing feature set was used in this graph, for single, double fused and triple fused features. The best performing feature sets, listed from left to right are: AuDeep-60, AuDeep-45/DeepSpectrum, auDeep-45/ComParE/BoAW-500, AuDeep-30, auDeep-30/DeepSpectrum and auDeep-30/DeepSpectrum/BoAW-125.

Lastly, Figure 4.5 shows the results of experiments with feature level fusion. Fusing the best performing feature sets leads to lower scores for the arousal classification task. In the double fused feature case, the score decreased with 3.81% with the SVM classifier and 7.68% with the GBM classifier. In the case of the GBM classifier, the score increased a bit with the triple fused feature fusion, but not enough to get close to the single feature score.

4.2 Valence Classification

4.2.1 Acoustic Features

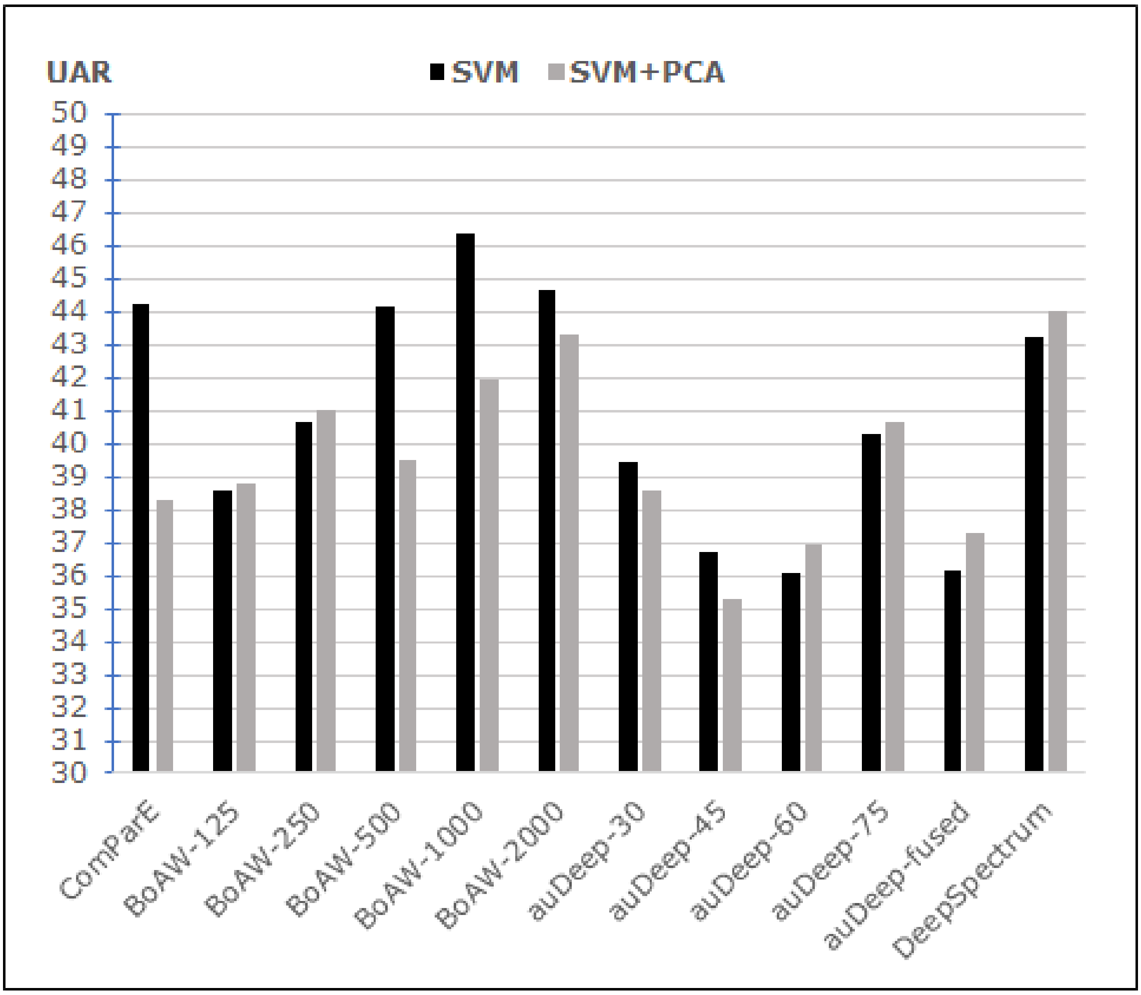


Figure 4.6: *Valence classification using a SVM classifier with acoustic features.*

Figure 4.6 illustrates the results of using SVM for the valence classification task, and additionally the effect of applying PCA on each feature set. Scores vary a lot between feature sets and also between using PCA or not. The best performing feature sets when using SVM are BoAW-1000 and BoAW-2000 with respectively 46.35% and 44.69% UAR. In both of these cases PCA transformation had a negative effect on the score. This impact of PCA transformation on the score when using SVM and for the valence classification is minimal in most cases, but had a significant negative effect when it comes to the ComParE, BoAW-500 and BoAW-1000 feature sets.

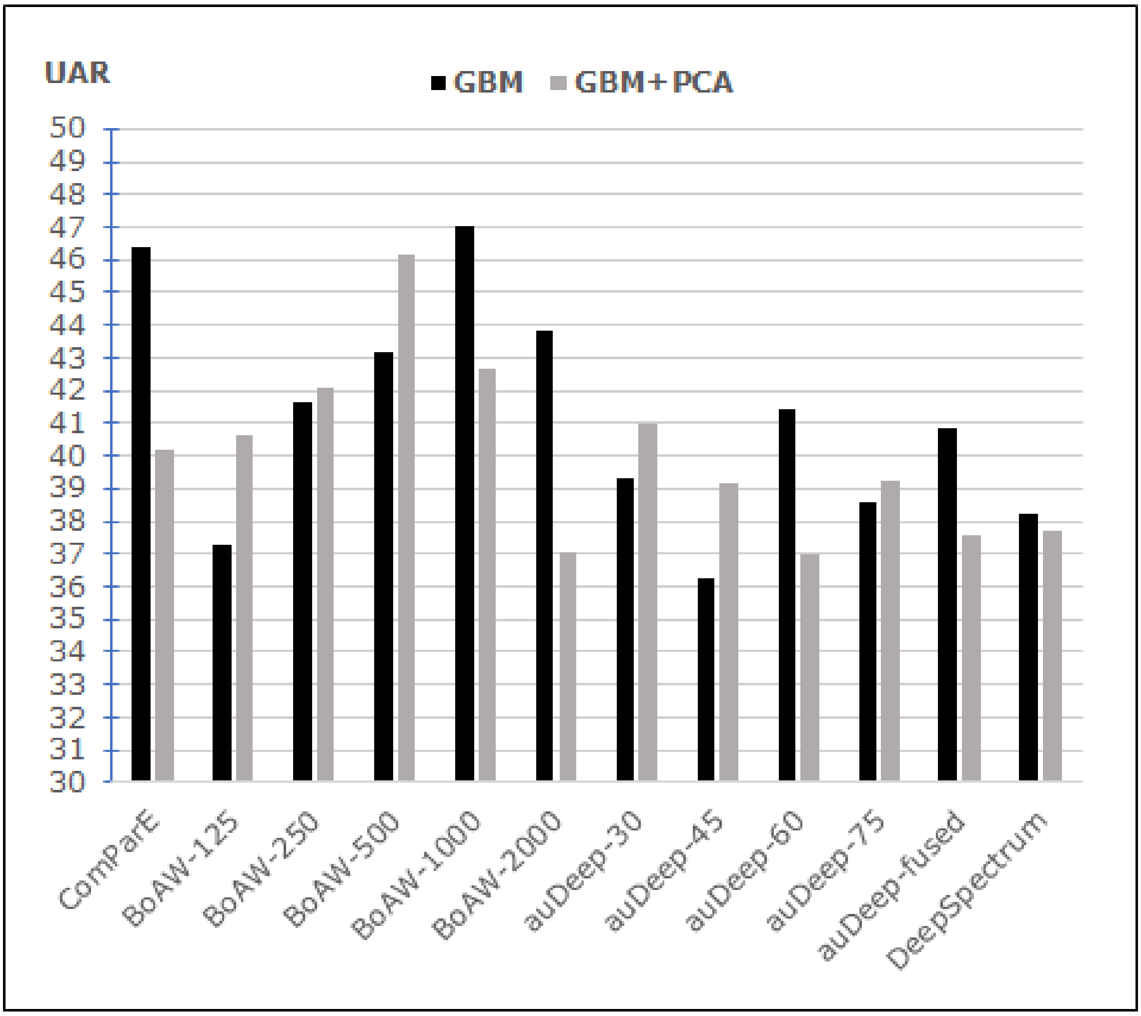


Figure 4.7: *Valence classification using a GBM classifier with acoustic features.*

Figure 4.7 illustrates the results of using GBM for the valence classification task, and additionally the effect of applying PCA on each feature set. As was the case with SVM; scores vary a lot between feature sets. The application of PCA shows some improvement with BoAW-500 with an improvement of 2.98%, although it mostly decreases the scores. This is especially the case with ComParE, BoAW-2000 and auDeep-60 with decreases in scores of respectively 6.17%, 6.74% and 4.50%. The best performing feature sets when using GBM are BoAW-1000 and ComParE with respectively 47.04% and 46.35% UAR. PCA had a negative effect on the scores for these feature sets.

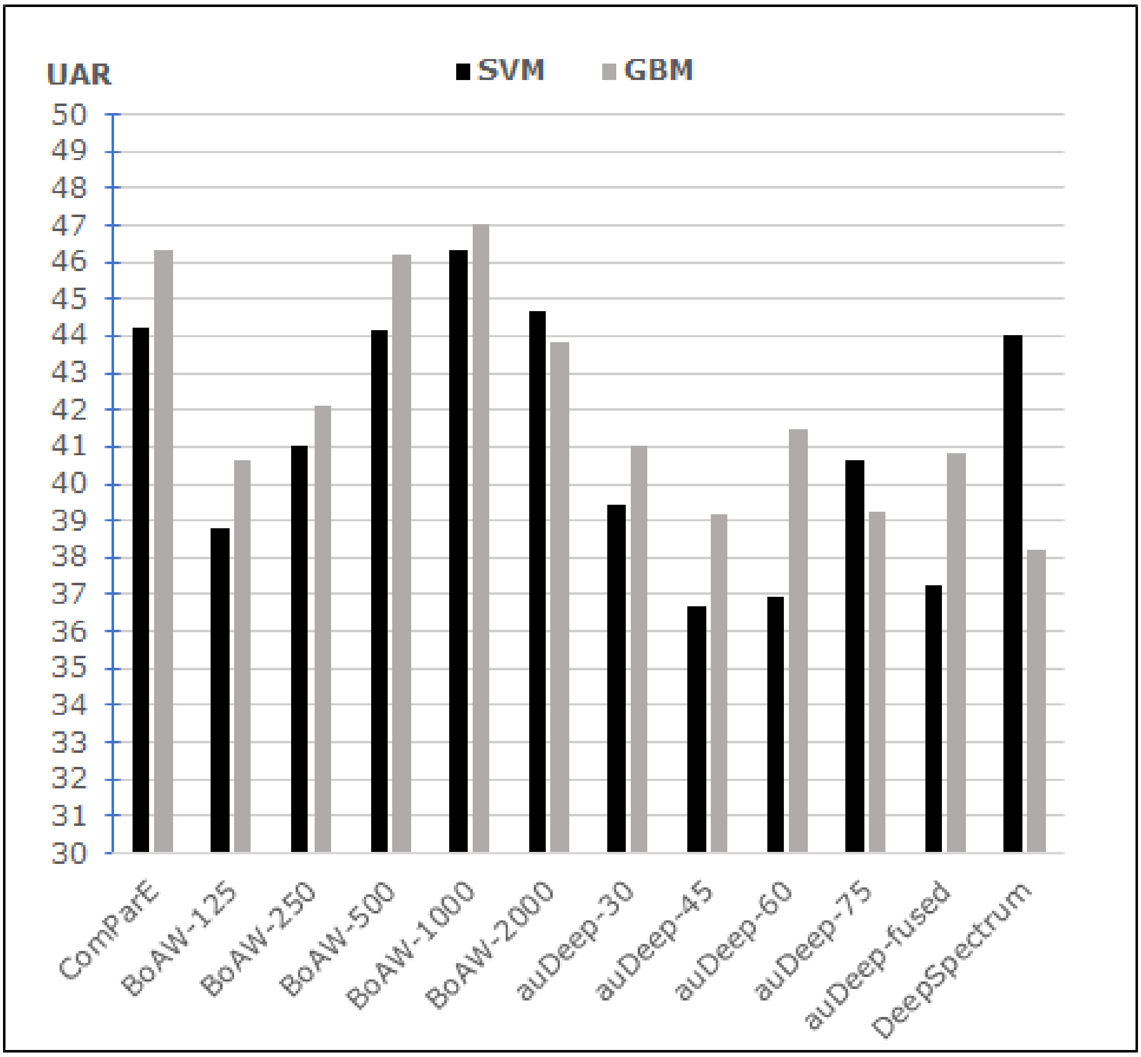


Figure 4.8: *Comparing SVM with GBM for the valence classification task, using acoustic features.*
Note that the score for each feature set is the maximum of applying PCA or leaving the feature set as it is.

Figure 4.8 illustrates the difference of scores between using SVM and GBM. The GBM classifier produced better results with most of the feature sets, except for BoAW-2000, auDeep-75 and DeepSpectrum, where the SVM classifier yielded better results. For BoAW-2000 and auDeep-75 the difference between SVM and GBM is not significant but for DeepSpectrum it is, with this feature set SVM scores 5.77% higher than GBM. For the valence classification task, the GBM classifier performs better, with the highest scoring feature sets being BoAW-1000 and ComParE. These two are also the highest scoring acoustic feature sets for the valence classification task. The mean over all feature sets with the GBM classifier is 42.17% UAR against the 41.19% UAR that the SVM classifier yielded.

4.2.2 Linguistic Features

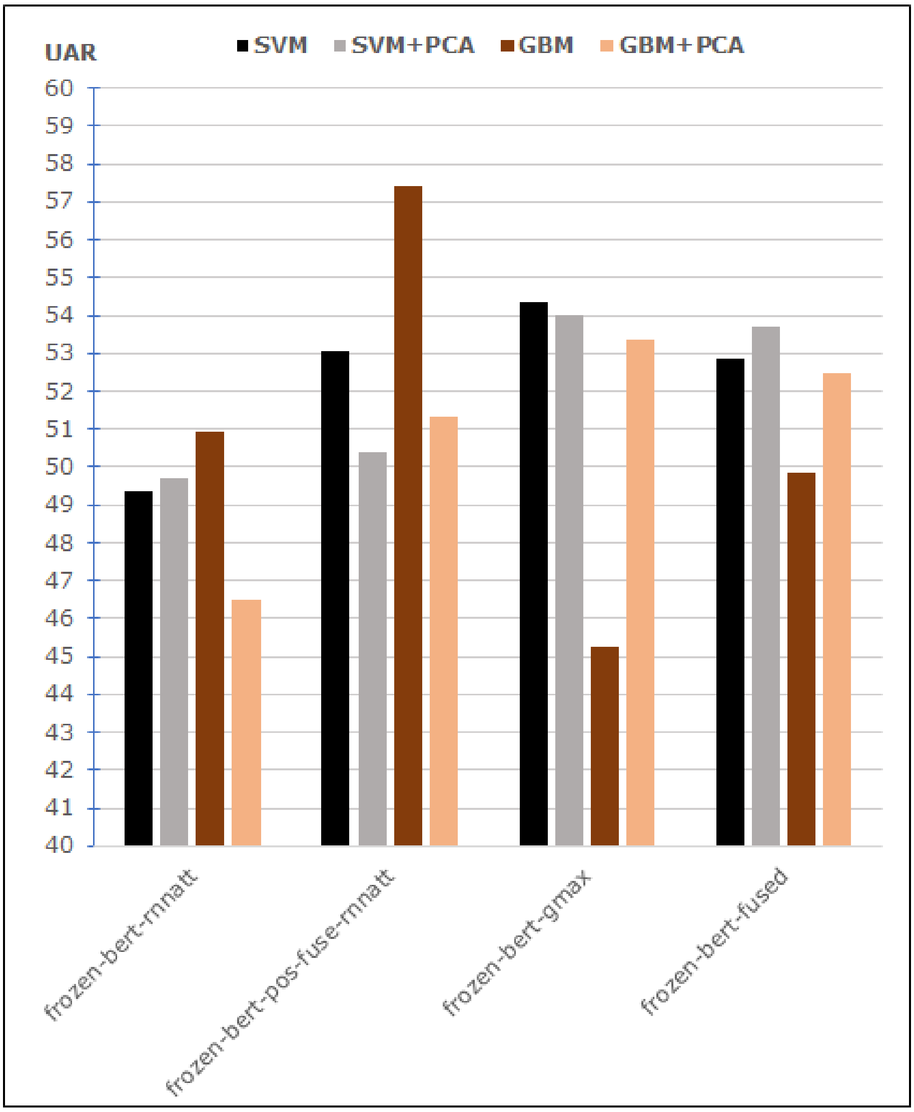


Figure 4.9: *Comparing SVM with GBM for the valence classification task, using linguistic features.*

Figure 4.9 shows the results of using the linguistic features for the valence classification task. The frozen-bert-pos-fuse-rnnatt feature set scores are the highest when used with the GBM classifier, being 57.42% UAR. This is not only the highest score for all linguistic features, it is also the highest score for the whole valence classification task. The application of PCA had a negative effect on the score, when using this combination of feature set and model. The highest score when using the SVM classifier goes to the frozen-bert-gmax feature set, with 54.35% UAR.

4.2.3 Feature Level Fusion Features

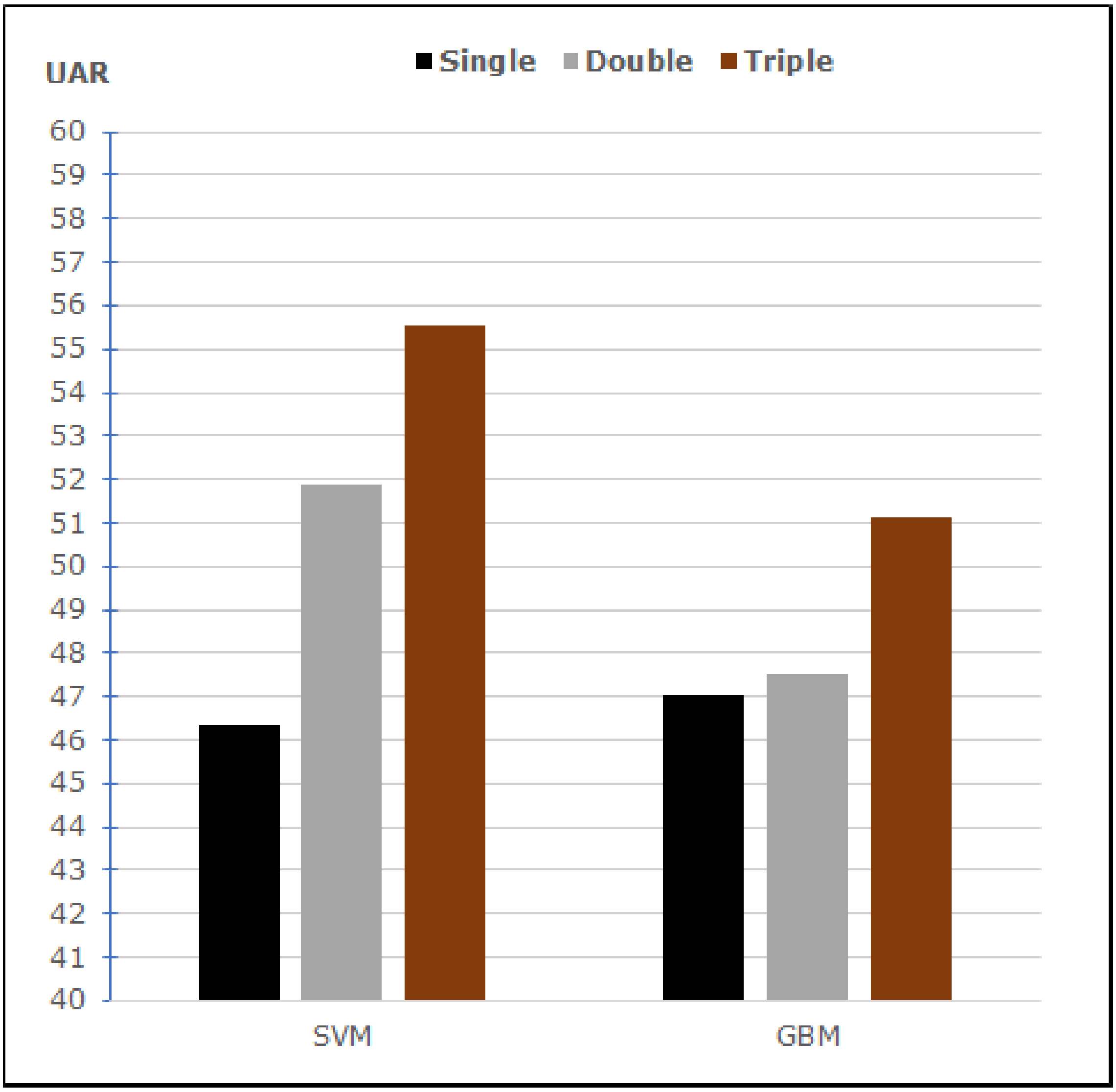


Figure 4.10: *Feature level fusion results for the valence classification task.*
The best performing feature set was used in this graph, for single, double fused and triple fused feature sets. The best performing feature sets, listed from left to right are: BoAW-1000, BoAW-1000/ComParE, BoAW-1000/ComParE/auDeep-75, BoAW-1000, BoAW-1000/auDeep-60 and ComParE/auDeep-60/BoAW-1000.

Lastly, Figure 4.10 shows the results of experiments with feature level fusion. Feature level fusion led to better results for the valence classification task, both with the SVM and GBM classifier. The double fused feature set led to an increase of 5.54% and 0.50%. The triple fused feature set led to an increase of 9.20% and 4.11%.

Chapter 5

Conclusion

In this section I will conclude my research by providing answers to the original research questions. Namely, which machine learning model (SVM or GBM) performs best on the arousal classification task, and which model performs best on the valence classification task? Which feature set leads to the best results on both classification tasks? What are the effects of PCA transformation on the scores? And finally, does feature level fusion lead to better results?

Firstly, conclusions will be drawn for the arousal classification task, whilst addressing each of these questions separately. And secondly, the same will be done for the valence classification task.

5.1 Arousal Classification

5.1.1 Machine Learning model/Feature set combination

The best performing machine learning model for the arousal classification task was the SVM classifier. The mean score of the SVM classifier, when using acoustic features, was 5.63% higher than the mean score of the GBM classifier. The SVM classifier is also computationally cheaper to run. The best scores, when using acoustic features, were obtained using single features with an SVM classifier. These were the auDeep-60 (44.41% UAR) and DeepSpectrum (43.27% UAR) feature sets.

SVM was also the preferred model when using linguistic feature sets. The highest score was obtained using the frozen-bert-rnnatt feature set (48.16% UAR).

5.1.2 PCA

The application of PCA on feature sets for the arousal classification task led to varying results. This makes it difficult to draw conclusions on whether PCA application has a positive effect on the score. There were some significant increases in score when PCA was applied, so it would seem advisable to at least experiment with the application of PCA when attempting to classify arousal.

5.1.3 Feature Level Fusion

Feature level fusion led to clear decreases in score for both the SVM and GBM classifier. With this we can conclude that applying feature level fusion for the acoustic classification task leads to worse performance.

5.2 Valence Classification

5.2.1 Machine Learning model/Feature set combination

When using acoustic features and single features, the GBM classifier performed better on average than the SVM classifier. Using the GBM classifier also led to the highest performing single feature sets, which were ComParE (46.35% UAR) and BoAW-1000 (47.04% UAR), although the SVM classifier performance came close to these scores with the BoAW-1000 feature set (46.35% UAR).

The GBM classifier also outperforms the SVM classifier when using linguistic features. The best performing feature set (frozen-bert-pos-fuse-rnnatt) with an GBM classifier leads to a 3.07% higher score than the best performing feature set (frozen-bert-gmax) with an SVM classifier. When it comes to double and triple fused features the SVM classifier outperforms the GBM classifier. More on feature level fusion for the valence classification task in section 5.2.3..

5.2.2 PCA

PCA transformation had a mostly negative or similar effect when using the SVM classifier. For GBM, PCA transformation led to some significant increases in score, so experimenting with PCA when using a GBM classifier could lead to better performances.

5.2.3 Feature Level Fusion

Feature level fusion led to a significant increase in score, both with the SVM and GBM classifier. The highest increase in score was obtained using the SVM classifier, with the highest score for a single feature set being 46.35% UAR (BoAW-1000), and the highest score for a triple fused feature set being 55.55% UAR (fusion of BoAW-1000, ComParE, auDeep-75). This leads to the conclusion that for the valence classification task, combining the best performing acoustic feature sets could lead to promising results.

Chapter 6

Discussion

The introduction mentioned that I am part of a larger group of researchers, working on the 2020 sub-challenge on Elderly Emotion. The implications of my experiments on the project will be mentioned in this chapter. Furthermore, I will reflect on the choices that were made in my research and what further research could be done on the subject. Part of this further research is already being done at the moment by others working on the challenge.

The following is a warning for researchers who want to use results from this thesis in their research considering the 2020 sub-challenge on Elderly Emotion.

One of the most important aspects of machine learning is generalizability. This means that a machine learning model, after training on the training data, is also good at classifying unseen data. For my own experiments, I trained on the supplied training data set and subsequently casted predictions on the test data set, these predictions were then used to calculate the score.

There is one additional data set available with the challenge from the same corpus. The crux with this data set is that the labels are unavailable for the researchers working on the challenge, except when an official submission is done to the challenge team. This is also explained in detail in the challenge paper and on the challenge website. Per team working on the challenge, five official submissions are permitted. Because my research was more of the exploratory kind, looking at the effects of different models, feature sets, PCA and feature level fusion, exactly one official submission was done. For this submission the following configurations were chosen, as these yielded the highest scores on the development data set:

For arousal: auDeep-60, with PCA applied, using the SVM classifier.
For valence: Frozen-bert-pos-fuse-rnnatt, with no PCA applied, using the GBM classifier.

The arousal submission yielded a score of 42.70% UAR. Unfortunately this score is not high enough to beat the challenge baseline, this baseline reports a score of 49.80% UAR. For valence the submission yielded as score of 42.80% UAR. Again this score is lower than the score for the challenge baseline, which reports 49.00% UAR.

For further research it is important to note that the challenge team has shared their results on the baseline feature sets in the challenge paper (B. W. Schuller et al., 2020), this is also where the challenge baseline score can be found. When doing further experimentation with these feature sets it is therefore paramount to look for results that also yield good results on the unlabelled data set.

The following conclusions from this thesis are important for the other researchers working on this sub-challenge. Firstly, there are some feature sets where the application of PCA had a positive effect on the score, and some where there was a negative effect on the score noticed. Also, feature level fusion had a negative effect on score of the arousal classification task but a positive effect on the score of the valence classification task. Secondly, for the arousal task the ComParE, AuDeep and DeepSpectrum features give the best results. For the valence task the ComParE, BoAW and DeepSpectrum features give the best results. This leads to the belief that the ComParE and DeepSpectrum feature sets lead to relatively good scores on both classification tasks. The AuDeep feature sets perform best for arousal, better than ComParE and DeepSpectrum. The BoAW feature sets perform best for valence., also better than ComParE and DeepSpectrum. Finally, the SVM classifier seems to be working best for the arousal task when single feature sets are used, and also yielded the best results with fused features on the valence classification task. This applies to both the acoustic and linguistic features. An advantage of using the SVM classifier is that the hyper-parameters are easier to tune, because the model is computationally cheaper. I will go into detail about hyper-parameter tuning next.

There were large differences noticeable for different hyper-parameters when using the GBM classifier. This could imply that more extensive experimentation with more values for the hyper-parameters could lead to higher scores. The SVM classifier scores were also very dependant on the C parameter, but this was easier to tune because of the fact that there is just one parameter to tune. Additionally to the tuning of the hyper-parameters for the machine learning models, more experimentation could be done with different amounts of variance kept after PCA transformation. I chose the value of 95% variance retained because it still keeps the the trend of the data intact while reducing the feature space significantly. Lower or maybe higher percentages of variance retained could lead to different results.

Finally, these experiments were done with the four linguistic feature sets supplied. I think a lot could be gained from experimentation with different linguistic feature representations to achieve higher classification scores. Other corpora with arousal and valence labels could possibly make the machine learning models more robust to over-fitting. These corpora would not come from elderly people, as the corpus used in this research is the first one released for this purpose, but could still provide more stability and higher percentage scores.

To conclude this thesis, this research has shown that there is still a lot of work that needs to be done to be able to accurately classify elderly emotion. In some cases the approaches tried in this research have shown promising results, but it is too early to make any conclusions. It has become clear that using PCA and doing feature level fusion alone is not enough to beat the challenge baseline system.

References

Amiriparian, S., Freitag, M., Cummins, N., & Schuller, B. (2017). Sequence to sequence autoencoders for unsupervised representation learning from audio. In *Proc. of the dcase 2017 workshop.*

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., … Schuller, B. W. (2017). Snore sound classification using image-based deep spectrum features. In *Interspeech.*

Breiman, L. (1997). *Arcing the edge* (Tech. Rep.). Technical Report 486, Statistics Department, University of California.

Devillers, L., & Vidrascu, L. (2006). Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Ninth international conference on spoken language processing.*

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

Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., & Boufaden, N. (2009). Cepstral and long-term features for emotion recognition. In *Tenth annual conference of the international speech communication association.*

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, *98*(45-60), 16.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st acm international conference on multimedia* (pp. 835–838).

Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., & Schuller, B. (2017). audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research*, *18*(1), 6340–6344.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771-780), 1612.

Garssen, J. (2011). *Demografie van de vergrijzing.* Centraal Bureau voor de Statistiek Den Haag.

Gilman, R. C. (1968). The general inquirer: A computer approach to content analysis. philip j. stone , dexter c. dunphy , marshall s. smith , daniel m. ogilvie. *American Journal of Sociology*, *73*(5), 634-635. Retrieved from `https://doi.org/10.1086/224539` doi: 10.1086/224539

Gosztolya, G., Busa-Fekete, R., & Tóth, L. (2013). Detecting autism, emotions and social signals using adaboost..

Grósz, T., Busa-Fekete, R., Gosztolya, G., & Tóth, L. (2015). Assessing the degree of nativeness and parkinson's condition using gaussian processes and deep rectifier neural networks.

Irie, K., Zeyer, A., Schlüter, R., & Ney, H. (2019). Language modeling with deep transformers. *arXiv preprint arXiv:1905.04226*.

Jones, C., & Jonsson, I.-M. (2008). *Using paralinguistic cues in speech to recognise emotions in older car drivers* (C. Peter & R. Beale, Eds.). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from `https://doi.org/10.1007/978-3-540-85099-1_20`

Kaya, H., & Karpov, A. A. (2017). Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold. In *Interspeech* (pp.

3527–3531).

Kenny, P., Parsons, T., Gratch, J., & Rizzo, A. (2008). Virtual humans for assisted health care. In *Proceedings of the 1st international conference on pervasive technologies related to assistive environments* (pp. 1–4).

Luxton, D. D. (2014). Artificial intelligence in psychological practice: Current and future applications and implications. *Professional Psychology: Research and Practice*, *45*(5), 332.

Mohammad, S. M. (2016). 9 - sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion measurement* (p. 201 - 237). Woodhead Publishing. Retrieved from `http://www.sciencedirect.com/science/article/pii/B9780081005088000096` doi: https://doi.org/10.1016/B978-0-08-100508-8.00009-6"

Ortony, A., Collins, A., & Clore, G. L. (1988). *The cognitive structure of emotions / andrew ortony, gerald l. clore, allan collins* (Pbk. ed. ed.) [Book]. Cambridge University Press Cambridge [England] ; New York. Retrieved from `http://www.loc.gov/catdir/toc/cam026/87033757.html`

Osgood, S., & Tannenbaum. (1957). *The measurement of meaning.* University of Illinois Press, Urbana.

Padgett, C., & Cottrell, G. (1996). Representing face images for emotion classification. In (Vol. 9, p. 894-900).

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S., & Setter, J. (1998). Transcription of prosodic and paralinguistic features of emotional speech. *Journal of the International Phonetic Association*, *28*(1-2), 83–94.

Russell, J. (1980, 12). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161-1178. doi: 10.1037/h0077714

Schmitt, M., & Schuller, B. (2017). Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit. *The Journal of Machine Learning Research*, *18*(1), 3370–3374.

Schuller, B. (2011). Voice and speech analysis in search of states and traits. In *Computer analysis of human behavior* (pp. 227–253). Springer.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., & Narayanan, S. (2013). Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, *27*(1), 4–39.

Schuller, B., & Weninger, F. (2012). Ten recent trends in computational paralinguistics. In *Cognitive behavioural systems* (pp. 35–49). Springer.

Schuller, B., Wimmer, M., Mosenlechner, L., Kern, C., Arsic, D., & Rigoll, G. (2008). Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? In *2008 ieee international conference on acoustics, speech and signal processing* (pp. 4501–4504).

Schuller, B. W., Batliner, A., Bergler, C., Messner, E.-M., Hamilton, A., Amiriparian, S., ... others (2020). The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks. *Proceedings INTERSPEECH. Shanghai, China: ISCA.*

Vidrascu, L., & Devillers, L. (2007). Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features. In *Proc. inter. workshop on paralinguistic speech between models and data, paraling.*

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1-4), 43–52.

Zhang, Z., Deng, J., & Schuller, B. (2013). Co-training succeeds in computational paralinguistics. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 8505–8509).

Appendix A

Arousal Classification Results

	SVM		**GBM**	
	no PCA	with PCA	no PCA	with PCA
ComParE	40.7	36.53	31.14	31.09
BoAW-125	31.14	38.37	33.09	29.12
BoAW-250	33.96	38.35	30.38	32.97
BoAW-500	40.26	40.67	29.29	28.18
BoAW-1000	39.72	39.81	35.51	29.12
BoAW-2000	36.53	31.22	29.12	28.84
auDeep-30	37.79	35.34	41.54	30.3
auDeep-45	42.76	36.65	28.9	39
auDeep-60	42.85	44.41	40.08	37.56
auDeep-75	41.34	38.11	31.86	36.8
auDeep-fused	35.4	41.29	35.89	33.71
DeepSpectrum	42.09	43.27	32.45	33.26

Table A.1: Single feature results

	no PCA	with PCA
auDeep-45/DeepSpectrum	41,34	41,34
auDeep-45/ComParE	41,89	41,89
auDeep-45/BoAW-500	38,78	35,32
DeepSpectrum/ComParE	40,18	39,76
DeepSpectrum/BoAW-500	34,77	34,77
ComParE/BoAW-500	40,6	39,52
auDeep-45/DeepSpectrum/ComParE	38,95	38,95
auDeep-45/DeepSpectrum/BoAW-500	35,68	35,68
DeepSpectrum/ComParE/BoAW-500	38,43	38,43
auDeep-45/ComParE/BoAW-500	39,61	39,61

Table A.2: Feature level feature results with the SVM classifier

	no PCA	with PCA
auDeep-30/DeepSpectrum	33,86	31,51
auDeep-30/BoAW-125	26,35	28,03
DeepSpectrum/BoAW-125	30,4	32,35
auDeep-30/DeepSpectrum/BoAW-125	27,68	34,35

Table A.3: Feature level feature results with the GBM classifier

Appendix B

Valence Classification Results

	SVM		**GBM**	
	no PCA	with PCA	no PCA	with PCA
ComParE	44,24	38,26	46,35	40,18
BoAW-125	38,56	38,77	37,28	40,65
BoAW-250	40,67	41,03	41,62	42,11
BoAW-500	44,18	39,54	43,2	46,18
BoAW-1000	46,35	41,92	47,04	42,7
BoAW-2000	44,69	43,32	43,82	37,08
auDeep-30	39,45	38,6	39,34	41
auDeep-45	36,69	35,32	36,27	39,19
auDeep-60	36,09	36,92	41,45	36,95
auDeep-75	40,31	40,67	38,59	39,22
auDeep-fused	36,18	37,28	40,82	37,55
DeepSpectrum	43,26	44,01	38,24	37,71

Table B.1: Single feature results

	no PCA	with PCA
BoAW-1000/ComParE	51,32	51,89
BoAW-1000/DeepSpectrum	41,3	43,68
BoAW-1000/auDeep-75	45,4	45,4
ComParE/DeepSpectrum	43,32	43,32
ComParE/auDeep-75	47,63	48,55
DeepSpectrum/auDeep-75	43,41	43,97
BoAW-1000/DeepSpectrum/auDeep-75	40,34	40,34
BoAW-1000/ComParE/DeepSpectrum	51,53	48,02
BoAW-1000/ComParE/auDeep-75	52,6	55,55
ComParE/DeepSpectrum/auDeep-75	51,59	54,54

Table B.2: Feature level feature results with the SVM classifier

	no PCA	with PCA
ComParE/auDeep-60	46,27	44,13
BoAW-1000/auDeep-60	47,54	47,04
BoAW-1000/ComParE	46,12	39,37
ComParE/auDeep-60/BoAW-1000	51,15	42,91

Table B.3: Feature level feature results with the GBM classifier