UPTEC STS 22002
Examensarbete 30 hp
Januari 2022

UPPSALA
UNIVERSITET

Machine learning for identifying how much women and men talk in meetings

Matilda Wellander
Vera Sintorn

Civilingenjörsprogrammet i system i teknik och samhälle

UPPSALA UNIVERSITET

Machine learning for identifying how much women and men talk in meetings

Matilda Wellander
Vera Sintorn

Abstract

For quite some time, it has been discussed that women are underrepresented in company boards. Furthermore, when they are a member of a board, they tend to have lower positions than men, meaning they have less power. One way to start solving the problem is to have more women in company boards and ensure they too have high positions. However, only having more women present might not be a complete solution. They also need space to speak to share their competence, ideas, and thoughts. Although, people tend to perceive women as more talkative than they actually are. For example, if a woman and a man speak the same amount of time, the woman is often perceived as having talked more than the man. To identify this problem, this study aimed to train a machine learning model that takes a recording of a meeting as input and calculates the time women and men spoke in percentage. The training data was based on 1266 episodes from the radio show “Sommar och Vinter i P1” where all episodes contained one speaker, different each time. 633 episodes contained female speakers and 633 contained male speakers, all speakers spoke Swedish in the recordings.

Four different models were trained using different training data, where logistic regression is the best performing algorithm for all four. The four models were evaluated using evaluation data and they showed to not differ significantly in performance. The subsequently chosen model was tested on two recordings with both male and female speakers, where the resulting accuracy was 83.5% and 83.1%. The application developed in this study can help identify the speaking space given to women in the workplace. However, how this tool could be used to achieve a more equal workplace still needs further research.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala

Handledare: Mikael Axelsson & Erik Löthman Ämnesgranskare: David Sumpter

Examinator: Elísabet Andrésdóttir

Populärvetenskaplig sammanfattning

Ojämställdhet mellan kvinnor och män har länge varit ett viktigt ämne i alla delar av samhället. Studier visar att det fortfarande är ett stort glapp i hur könsfördelningen ser ut i bolagsstyrelser då kvinnor endast står för en tredjedel av medlemmarna. Utöver det har kvinnor oftare lägre positioner i styrelsen. För att lösa detta problem är det nödvändigt att få in fler kvinnor i bolagsstyrelser, men det är inte säkert att endast den åtgärden ger kvinnor samma makt som män. För att faktiskt åstadkomma jämställdhet krävs att kvinnor får samma utrymme att prata och förmedla sin kompetens, idéer och tankar. Detta kan vara problematiskt då kvinnor uppfattas prata mer än de faktiskt gör. Om exempelvis en kvinna och en man pratar lika mycket uppfattas kvinnan prata mer än mannen. För att uppmärksamma problemet krävs ett objektivt synsätt som kan vara svårt att få genom människor. Därför behövs ett automatiserat sätt att analysera och mäta talutrymmet som kvinnor och män får.

Ett sätt att analysera data som har blivit mer och mer populärt är maskininlärning. Med hjälp av maskininlärning kan en modell tränas med hjälp av befintliga träningsdata, vilken sedan kan användas för att förutspå framtida värden. Denna studie ämnade att undersöka hur maskininlärning kan användas för att beräkna hur mycket talutrymme kvinnor och män får under möten. Träningsdatat utgjordes av inspelade klipp där varje klipp bestod av en kvinna eller man som talade svenska. Fyra olika träningsset skapades där klippen var 10 till 20 sekunder långa. Ur dessa togs ett antal numeriska värden ut, exempelvis medelfrekvens och standardavvikelse, vilka sedan användes för att träna fyra olika modeller. De fyra modellerna utvärderades med hjälp av utvärderingsdata, vilket visade att det inte fanns en betydande skillnad mellan dem. Den valda modellen användes sedan för att beräkna två inspelade möten där den resulterande noggrannheten var 83,1% och 83,5%. Den här studien visar att det är möjligt att beräkna hur mycket kvinnor och män pratar under ett möte.

Med hjälp av modellen skapades en lokal webapplikation där användaren kan ladda upp en ljudfil i ett ’.wav’-format och få ut ett cirkeldiagram på fördelningen av talutrymme. En viktig aspekt i webapplikationen är dess användningsområde och hur den ska utnyttjas för bästa effekt på en arbetsplats. Det har visat sig att ökad förståelse och kunskap kring ojämställdhet och fördomar kan resultera i ett mer medvetet agerande. Det är dock viktigt att en enskild person inte blir utpekad, utan att applikationen belyser ett gemensamt ansvar för ett strukturellt problem.

Distribution of work

This thesis project was conducted by Matilda Wellander and Vera Sintorn in close collaboration. Some of the programming was divided between the authors where Matilda has focused on programming in R and Vera focused on programming in Python. However, the web app was created using pair programming. Different sections of the report was initially distributed equally, after which reviewing, and rewriting all sections was carried out by both authors.

Förord

Detta examensarbete har utförts på Civilingenjörsprogrammet i system i teknik och samhälle vid Uppsala universitet och markerar slutet på vår utbildning. Examensarbetet har utförts i samarbete med IT-konsultföretaget Consid i Uppsala.

Vi vill särskilt tacka våra handledare på Consid, Mikael Axelsson och Erik Löthman för er stöttning och vägledning under arbetets gång. Vi vill även tacka övrig personal på Consid för allt engagemang och intresse.

Till sist vill vi rikta ett stort tack till vår ämnesgranskare David Sumpter. Din kunskap och entusiasm har varit ovärderlig!

Table of Contents

1. Introduction

Statistics from 2017 show that there were half as many women as men on company boards in Sweden. Additionally, women were less likely to have high positions like chairman of the board, instead having positions such as alternate member of the board (SCB, 2017). In this context, it should be noted that companies with gender equal board of directors show a higher performance, profitability and turnover than companies with a gender gap (Malmström, 2021).

Gender equality means that men and women have the same rights, obligations, and opportunities in all areas of life. Everyone should have the same power to shape their own lives and the community they live in (Ledarna, n.d.). It is reasonable to assume that having a lack of women in leading positions inevitably means they have less power in companies, and, by extension, in society. Additionally, a study made by Make Equal (2018) shows that companies and organisations that do close the gender gap still struggle with inequality.

The power dynamics and opportunities to speak within gender equal groups still favours the men (ibid). Men tend to interrupt more, talk longer and with more authority than women (Kendall and Tannen, 1997; Jacobi and Schweers, 2017). However, when women mimic the men's rhetorical behaviour, they lose likability from their peers (Kendall and Tannen, 1997). Furthermore, when women speak and share their ideas in a group of mixed genders, they more often get overlooked or have their ideas stolen (Tannen, 2017). Even though women do not talk as much as men in professional settings it is often perceived that women talk beyond their remit (Toegel, 2021). Due to this perception of women, it is hard to measure the space taken by women in meetings just by asking the meeting participants.

Helping companies to distinguish this problem could be a step in the direction of more equal workplaces. Studies have shown that implicit bias can be reduced by intervention, which can also increase awareness of bias and discrimination (Devine et al., 2012). However, other studies have showed that being made aware of bias in close relation to an interaction instead can cause a person to be more cautious and inhibited, leading to them appear more prejudice than intended (Vorauer, 2012). A study by Hoefel (1991) shows that being made aware of structural problems allows for a better understanding and respect of different people's perspectives and enables a more benevolent way of thinking. Awareness of discrimination of minorities lead to a better insight of their experiences (Dirth and Branscombe, 2017). When discriminating structures are acknowledged it is common that one single person is targeted as the villain. This is a way for the group to separate from the issue and can lead to false assumptions that the problem is one person and not on a structural level (Hoefel, 1991; Dirth and Branscombe, 2017).

To illustrate the problem of inequality on a structural level, women's speaking space could be measured and analysed. By measuring the differences in opportunities to speak, companies could be made aware of an issue that might be difficult to identify otherwise. One possibility here would be an automized system that can measure the speaking time of men and women in meetings. It is such a system we investigate in this paper.

One increasingly common way to analyse data is to use machine learning (ML). To some people, ML may sound like a futuristic fantasy, but it has now been around for decades. The first ML application to gain popularity was the e-mail spam filter that took over the world in the 1990s. Following that, hundreds of products and features regularly used have taken the same path to enable functions, for example, better recommendations and voice search. The ML program in a spam filter is given examples of spam emails that are flagged by users and examples of no-spam emails. These examples, called *training set*, can then be used by the system to learn which emails to flag (Géron, 2019). Within ML, there are several different algorithms. What determines which algorithm to use is the required output. The different algorithms mainly fall under one of two categories: supervised or unsupervised learning (Bell, 2020). In supervised learning, there is some training data which consist of a set of samples with their input and output (x_i, y_i). x_i is the input vector belonging to data point i, which consists of several values called *features* and y_i is the label or class of the data point. Training the model on enough samples will result in future predictions of y_i from the input x_i (Lindholm et al., 2021).

In order to analyse speaking space, it is of interest to define how female and male voices are identified. Gelfer and Schofield (2000) studied the voices of transsexual women that where categorized by a focus group. They found that the voices that were perceived as female had higher speaking fundamental frequency (SFF) and higher vowel formants than the voices perceived as male. Other studies agree that the main factors that determines the listeners perception of gender of the speaker is SFF and vowel formant frequency (F) (Leung et al., 2021). The mean values of the SFF is 120Hz for male voices and 210 Hz for female (Eriksson and Traunmüller, 1995). In more general terms a male voice has a fundamental frequency between 85-155 Hz and a female voice has between 165-255 Hz (Fitch and Holbrook, 1970). The fundamental frequency can be affected by the speaker's age, length, weight, and medical history.

There are some studies about how voices can be classified as female or male using ML. Kory Becker (2016b) conducted an ML study where different ML algorithms was tested to classify the gender of a speaker using data extracted from voice recordings. This model reached nearly 100% accuracy on the training set and 89% accuracy on the test set with a stacked ensemble method based on Support Vector Machine (SVM), random forest (RF) and XGBoost. Using logistic regression resulted in 72% accuracy on the training set and 71% accuracy on the test set. Raahul, Sapthagiri, Pankaj and Vijayarajan (2017) carried out a similar ML study where five different ML algorithms were compared. The algorithms tested were Linear Discriminant Analysis (LDA), K-Nearest Neighbour (KNN), Classification and Regression Trees (CART) as well as RF and SVM. Eight

different metrics were used to compare the algorithms and the results showed that SVM had the best performance in this classification problem. Another study using the same dataset as previous ones used Multilayer Perceptron Deep Learning, with a supervised learning technique called backpropagation for training the network (Büyükyılmaz and Çıbıkdiken, 2016). The model gave a 96,74% accuracy on the test data set and showed that acoustic properties of voices can be used for detecting the gender of a speaker.

While there are some studies conducted on the topic, the three previously mentioned presented the same training data that was spoken in English (Raahul et al., 2017; Büyükyılmaz and Çıbıkdiken, 2016; Becker, 2016b). The data consists of 3168 data points and is available on Kaggle (Becker, 2016a). Furthermore, differences in intonation show tendencies to impede a correct prediction, especially with voices within an androgynous frequency range (Becker, 2016b). Intonation is unique to each language, causing speakers of different languages to use different tunes (Chamonikolasová, 2017). This could cause issues when using the existing models on languages other than English. The article 'The importance of natural language processing for non-English languages' (Telus International, 2021) addresses the issue of ML mainly catering towards seven languages: English, Chinese, Urdu, Farsi, Arabic, French and Spanish. English is, however, by far the language in which most technological advances has been made. Supporting diverse languages in ML helps secure the global variety of languages and can improve ML technology by introducing new languages before the technology is so advanced that it is impossible to do so (Telus International, 2021). For this reason, this project aims to create a model based on the Swedish language.

1.1 Purpose and Research Question

The purpose of this study is to explore the possibilities to use machine learning to identify the biological gender of a Swedish speaker. The goal is to develop an application that takes a recording of a meeting as input and determines what percentage of the time women spoke and what percentage men spoke.

To reach the goal of this study, the following question will be considered.

- How can machine learning be used in order to identify how much men and women talk in meetings?

1.2 Delimitations

The purpose of this thesis is to identify a structural problem, there will be no focus on remedies to reduce the problem with inequality in the workplace. Additionally, only the genders female and male will be considered in this project.

2. Method

This section will illustrate how the project was carried through and which steps were taken to yield a desired result. Firstly, the data used in the project will be displayed, as well as how it was pre-processed. Secondly, the feature construction will be explained. Thirdly, training, evaluating, and testing the model will be disclosed.

2.1 Data

Data has mainly been collected from Sveriges Radio's website in the form of podcasts in mp3 format. The training data consists of episodes from "Sommar och Vinter i P1" (Sveriges Radio, n.d.). The reason for this is that only one person speaks in each episode, which simplifies the labelling. All of the chosen episodes are in Swedish, which is the spoken language that this project focuses on. 633 recordings classed as female and 633 recordings classed as male were used. The reason for having an equal amount of both genders is that having imbalanced classes can decrease performance due to biases towards the majority class (Sigh Raghuwanshi and Shukla, 2021).

The evaluation data was created using other types of podcasts from Sveriges Radio's website. The purpose of the evaluation data was to evaluate different ML models on data that resemble the test data, and subsequently choosing the one performing best on the evaluation data. Evaluating an ML model on unseen data can be beneficial for determining its performance and ensure that the model is not overfitted to the training data (Lindholm et al., 2021). The podcasts are labelled female only, male only as well as a mix of female and male. To create the best opportunity for a fair evaluation, podcasts with different types of people and voice characteristics were used. The properties and podcasts chosen for the evaluation set are shown in Table 1. It is important to note that each podcast is an arbitrary choice, and the results heavily depend on the exact characteristics of each podcast.

Table 1. Characteristics and podcasts in the evaluation data.

Characteristics	Podcasts
Only women	recA: Cecilia Gyllenhammar: Det privilegierade livet har en baksida (Hahr and Gyllenhammar, 2021)
Only women, one with a perceived low pitch voice	recB: USApodden: Trump testar gränserna (Stenholm et al., 2021)
Only men	recC: Snedtänkt med Kalle Lind: Om den kommersiella radion (Lind, 2021)

One man with a perceived high pitch voice	recD: Radionovellen Borlängecampagne av Gustav Tegby i uppläsning av Sven Björklund (Björklund, 2020)
Both men and women	recE: Nordegren & Epstein: Kan katten få lika hög status som hunden? (Epstein et al., 2021)
	recF: Filosofiska rummet: Den hemliga donationen (Mogensen et al., 2021)
	recG: Katarina Hahr möter: Sigge Eklund: Jag har känt mycket avund (Hahr and Eklund, 2021)
	recH: Morgonpasset i P3 – Gästen: Krogkungen PG Nilsson – 40 år i branschen, grytprotesten och så beter du dig på krogen (Akolor, Druid, Lundberg, and Nilsson, 2021)
	recI: Morgonpasset i P3 – Gästen: Ninja Thyberg – Pleasure, porrindustrin och premiärbakfylla (Akolor, Druid, Lundberg, and Thyberg, 2021)

After the evaluation, a dataset with test data was used to test the application with the chosen model. The test data consists of a recording of a meeting (Wellander and Borgström, 2021) (test1), as well as a podcast by Nour El-Refai and Henrik Schyffert (El-Refai and Schyffert, 2021) (test2). The two contain female and male speakers in a natural conversation.

2.2 Pre-processing

In this section, the ways in which training, evaluation and test data was pre-processed are presented.

2.2.1 Training set

After downloading each podcast in an mp3 format, they needed to be cut into smaller sections for further pre-processing in R. The reason for this was to cut out parts with intro and other sounds, like music or background noise, but also to shorten the audio for easier processing in R. The initial audio length to be taken from each podcast was chosen to be 20 seconds. Listening to a handful of the podcasts resulted in the decision to cut out the 20 sec of each podcast at 40 sec to 60 sec, since a majority had the wanted characteristics at that time stamp. This was done using the software mp3splt-gtk. After the initial batch cut of all podcasts, each section was listened to, to ensure that there was only the speaker's voice in the section. For sections that had music, background noise, or the speaker doing a different voice to imitate someone else were recut to only contain the speaker's voice.

Each podcast was only used once, with one clip. When all sections were completed, they were all converted to a '.wav' format[1].

Four different training sets were created to train four different ML models, enabling evaluation, and subsequently choosing the one with best performance on an evaluation set. To create the different training sets, silent parts were removed from the audio sections[2], after which shorter sections of the training set was created[3]. The reason for creating different training sets was to see if there was any difference in accuracy on the evaluation set and subsequently if the length of a section had played a role in future prediction. Table 2 shows the training sets being used to train the ML models.

Table 2. Descriptions of the four training sets.

Name	Description
Training set 1 (TS1)	Sections of 20 seconds, silence not removed.
Training set 2 (TS2)	Sections of 20 seconds, silence removed (i.e., some sections are shorter than 20 seconds).
Training set 3 (TS3)	Sections of 15 seconds, silence removed.
Training set 4 (TS4)	Sections of 10 seconds, silence removed.

2.2.2 Test and evaluation set

The evaluation and test sets were created in the same way as the training set. Firstly, parts of the audio that contained music or sounds that were not the speakers' voices, were manually cut out. Secondly, it was converted to a '.wav' format and silent parts are cut out automatically. Thirdly, the audio was broken down into sections of 10 seconds. The reasoning behind the length of each section was that it should be long enough for features to be extracted, i.e., to have enough data in each section to calculate the acoustic properties. At the same time, it should minimise the risk of having both a female and male speaking in the same sections, which could complicate a correct prediction.

[1] See function 'convert2Wav' in script 'meeting2csv' in https://github.com/vsintorn/VoiceGenderClassification
[2] See script 'soundprocessbatch.py' in the link above.
[3] See function 'shortenSound' in script 'shortenSound.R' in the link above.

2.2.3 Feature construction

Becker (2016c) has made a function in R called specan3, which is similar to the R warbleR (version 1.1.2) library function specan (RDocumentation, n.d.). The function takes a batch of sound files as input and measures 22 acoustic parameters on each sound file (Becker, 2016c), see Table 3. These can then be put into a csv file for further use. specan3 has been used for feature construction on the training data as well as the evaluation and test data.

Table 3. Outputs of the function specan3.

Header	Explanation
duration	Length of signal (removed before training).
meanfreq	Mean frequency in kHz.
sd	Standard deviation of frequency.
median	Median frequency in kHz.
Q25	First quantile in kHz.
Q75	Third quantile in kHz.
IQR	Interquantile range in kHz.
skew	Skewness
kurt	Kurtosis
sp.ent	Spectral entropy
sfm	Spectral flatness
mode	Mode frequency
centroid	Frequency centroid
peakf	Frequency with the highest energy (removed before training).
meanfun	Average fundamental frequency.
minfun	Minimum fundamental frequency.
maxfun	Maximum fundamental frequency.
meandom	Average dominant frequency.

mindom	Minimum dominant frequency.
maxdom	Maximum dominant frequency.
dfrange	Range of dominant frequency.
modindx	Modulation index. The accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range.

2.3 Training

In this section, the methods for training the models with be presented.

2.3.1 Machine learning methods

In this project the tool used for training and creating the ML model is a Microsoft Azure platform called Azure Machine Learning. Azure ML is a part of the Azure cloud service that provides compute resources, a user interface and ML tools. The user interface Azure ML provides is called Azure ML studio and this is where the different runs and models were monitored, compute resources were managed and Automated ML (AutoML) runs were initialised (Baccam et al., 2021).

When starting an AutoML run in azure, it automatically runs different algorithms and scaling methods that are evaluated and compared to each other. AutoML is built on Python and uses the Python library Scikit-learn. The AutoML run results in many different models to choose from, all with corresponding evaluation metrics and feature importance graphs (Baccam et al., 2021). It is then possible to choose which model to use, most likely the one with the highest score of the chosen metric. The algorithms tested by AutoML is logistic regression, XGBoost classifier, LightGBM, support vector machine, k nearest neighbour and extreme random trees. We do not detail these methods here, but they are covered in more detail in, for example, http://smlbook.org.

In this project four models were created by running AutoML with the four different versions of training data, as mentioned in section 2.2.1. All AutoML runs resulted in logistic regression being the algorithm that presented the highest performance. Logistic regression is a classification method that is efficient for binary classification problems which are linearly separable (Subasi, 2020). It is often usen in ML due to its simplicity (Joshi, 2017). The input may include one or many independent variables (Stoltzfus, 2011) and the output is dependent on the input (Joshi, 2017).

When training an ML model, it can be beneficial to balance the importance of the features so that no features ends up incorrectly more dominant than the others. The features of the dataset are of different units and measurements. In order to reduce these differences,

scaling of the data is done (Joshi, 2017). The scalers used to pre-process the data before training the optimal model for each dataset was different for each model. Table 4 shows the different scalers used by the four models in this project.

Table 4. Descriptions of the four scalers.

Scaler	Description
StandardScaleWrapper	Standardizes the features. All values are divided by the standard deviation and the outcome is a distribution with the standard deviation and variance of 1. The mean will be 0. (Hale, 2019)
RobustScaler	Subtracts the median of the features and then divide by the interquartile range. Limits importance of outliers (Hale, 2019).
MinMaxScaler	Scales the data by subtracting the minimum value of each feature and then dividing it by the original range. The shape of the original distribution is not changed and outliers are not reduced.(Hale, 2019)
MaxAbsScaler	Scales the features by maximum absolute value. The maximum absolute values are set to be 1.0 (Baccam et al., 2021)

2.3.2 Validation

The different models that were run by AutoML were automatically validated so the best model can be found and deployed (Microsoft Docs, 2021). For each model, several quality metrics were calculated. Amongst others, these included accuracy and AUC. When initializing the runs that train the models, the primary evaluation metric was chosen to be accuracy. Accuracy is the comparison of true labels and predicted labels (Joshi, 2017). The metrics were calculated using the three-fold cross-validation approach. This is automatically selected in AutoML and is based on the size of the dataset (Microsoft Docs, 2021).

2.3.3 Evaluation

To enable calculating the accuracy of the evaluation set, a solution file was manually created for the podcasts where there were both female and male speakers. The criteria for if a section would get the label ‘female’ or ‘male’ was which gender spoke the majority of the section. It is important to note that the creation of the solutions was done manually and may therefore not be completely accurate at all times. A prediction was made for each podcast using all four models, after which the accuracy of each model was calculated with

the help of the manually created solution[4]. The results were then saved in an Excel spreadsheet for further evaluation. To determine if there was a significant difference between the models, the Friedman test was utilised.

2.3.4 Friedman test

Friedman two-way analysis of variance by ranks can be used when N different cases are observed under k different conditions (Siegel and Castellan, 1988). I.e., it can be used to test the prediction of k different machine learning models on the same set of N subjects. In this case there are four different models and nine subjects to be predicted. The value to analyse is the accuracy each model has on each subject. The null hypothesis of the test is that each model come from the same population, alternatively, populations with the same median (Siegel and Castellan, 1988).

To conduct the test, data are cast in a table with N rows and k columns. Each column represents an ML model, and each row represents a subject. The data in each row is ranked separately, from 1 to k. Each subject in one row will therefore get the rank 1 to 4, depending on their value in relation to the other models in the same row. For the null hypothesis to be true, the distribution of ranks in each column should be unpredictable. Each column should be expected to have about the same frequency of rank 1, 2, 3 and 4, i.e., the sum of ranks in each column could be expected to be N(k + 1)/2 (Siegel and Castellan, 1988), 9(4+1)/2 = 22.5 in this case.

The Friedman test shows if the rank totals (R_j) for each ML model differ significantly from a coincidence. This is tested by computing F_r:

$$F_r = \left[\frac{12}{Nk(k+1)}\sum_{j=1}^{k} R_j^2\right] - 3N(k+1) \quad (1)$$

- N = number of subjects (podcasts)
- k = number of columns (ML models)
- R_j = sum of ranks in the jth column

The critical value for F_r when k=4 and N=9 is equal to 7.82 with 0.05 level of significance. If the calculated value of F_r is more than the critical value, the null hypothesis can be rejected, and the models can be assumed to significantly differ in performance. If not, there is no significant difference between the models. (ibid)

2.4 Testing

After choosing one of the models, it was tested on a test set consisting of one real recorded meeting (one man and two women) and one podcast (one woman and one man). The two recordings were predicted using the chosen model and the results were presented in a pie

[4] See script 'testModels.py' in https://github.com/vsintorn/VoiceGenderClassification.

chart with the percentage of time women and men spoke. Additionally, a solution for both recordings was created to calculate and present the accuracy of the predictions.

3. Results

In this section, the results of the study are presented. Firstly, the performance of each model on the training data is be displayed. Secondly, the models are evaluated using the evaluation data and the Friedman test, after which a final model is chosen. Thirdly, the chosen model will be tested on the test data.

3.1 Training

The results from training the model in AutoML are presented in Table 5.

Table 5. Accuracy and performance of the four different models on the training set.

Training set	Accuracy	Best model
TS1	0,96850	StandardScalerWrapper, LogisticRegression (M1)
TS2	0,92126	RobustScaler, LogisticRegression (M2)
TS3	0,93701	MixMaxScaler, LogisticRegression (M3)
TS4	0,92126	MaxAbsScaler, LogisticRegression (M4)

All four models use the logistic regression algorithm as it provided the highest accuracy. Using logistic regression to classify gender by voice has not been tested in previous studies. Raahul et al. (2017) concluded that a SVM model provided the best accuracy for this ML problem and Becker (2016b) only tested SVM, random forest and XGBoost. In this study, all said algorithms provided lower accuracy than logistic regression when compared in AutoML. All previous studies used the same dataset, larger than the dataset used in this study and based on the English language instead of Swedish. Because of the previous studies not testing logistic regression it is hard to say if logistic regression is the best choice for other datasets as well.

3.2 Evaluation

The results from predicting the evaluation data using the four models are presented in Table 6, followed by the ranks of each model on each subject in Table 7.

Table 6. Accuracy of each recording tested on the four models.

	M1	M2	M3	M4
recA	1.000	0.9936	1.000	1.000
recB	0.7343	0.8112	0.8427	0.8007
recC	0.9862	0.9725	0.9427	0.9771
recD	0.9918	0.9508	0.9344	0.9836
recE	0.9145	0.9294	0.9442	0.9368
recF	0.9429	0.9551	0.9429	0.9510
recG	0.6714	0.4429	0.5643	0.6214
recH	0.9135	0.9027	0.9297	0.8757
recI	0.5706	0.7588	0.6588	0.7412

Table 7. Rank of the accuracies in Table 6.

	M1	M2	M3	M4
recA	2	4	2	2
recB	4	2	1	3
recC	1	3	4	2
recD	1	3	4	2
recE	4	3	1	2
recF	3.5	1	3.5	2
recG	1	4	3	2
recH	2	3	1	4
recI	4	1	3	2
Sum of ranks	22.5	24	22.5	21

Calculate F_r using the sum of ranks and equation (1) from section 2.3.4.

$$F_r = \left[\frac{12}{9 * 4(4+1)}\sum_{j=1}^{4} R_j^2\right] - 3 * 9(4+1) = 0.3$$

The critical value for F_r is 7.82, which is larger than 0.3. I.e., the null hypothesis is not discarded. There is therefore no significant difference in performance between the models.

The results in Table 5 shows that the models performs better when there is only one gender speaking (recA-recD), except for recB where one of the speakers is a female with a low pitch voice. The test sets with mixed genders speaking overall get lower accuracy from the models. The reason for this could be because the sections that is sent to be classified by the models may contain speakers of different genders, whereas the models only were trained using data with one speaker per section.

3.3 Final Model Choice

Since the Friedman test shows that there is no significant difference between the models, there is no clear choice of model. They all have equal performance. The final model choice is therefore the one with a slightly lower sum of ranks than the other models, M4. It is also trained with the TS4, which is the training set most similar to the test data, since the audio sections are 10 seconds long.

3.4 Features

AutoML produces a feature importance graph of each model that is trained. The features with their respective importance in Figure 1 is from the chosen model, M4.

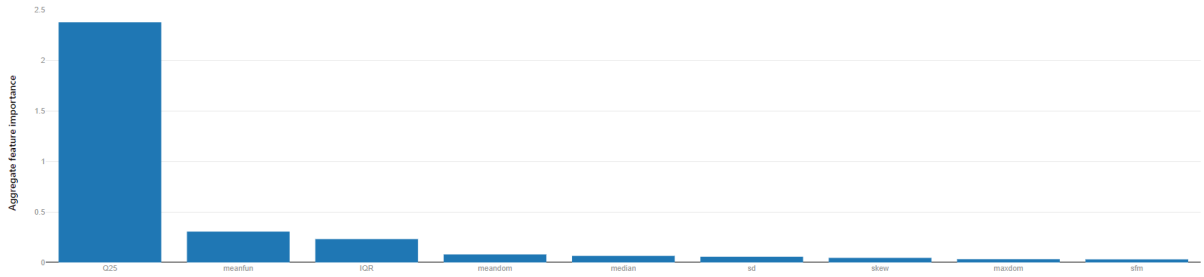

Figure 1. Feature importance graph of the chosen model.

Figure 1 shows that the Q25 (first quantile) is the significantly most important feature when predicting a gender from their voice. Average fundamental frequency (meanfun) comes second with an importance of almost ten times less than Q25. This shows that frequency is an important factor in predicting a person's gender. This agrees with Pausewang Gelfer and Schofield's (2000) findings that voices with a higher speaking fundamental frequency is perceived as female.

Figure 2 shows the dependence plot for the most important feature Q25 for the class ‘female’. It shows that a higher Q25 means a higher chance for a prediction of class ‘female’. This could result in false predictions for males with high pitch voices and females with low pitch voice. This is something that needs to be taken in account when using the model, since it could be perceived as biased towards females with high pitched voices and males with low pitched voices. It is especially true with this model due to it relying mostly on one feature.

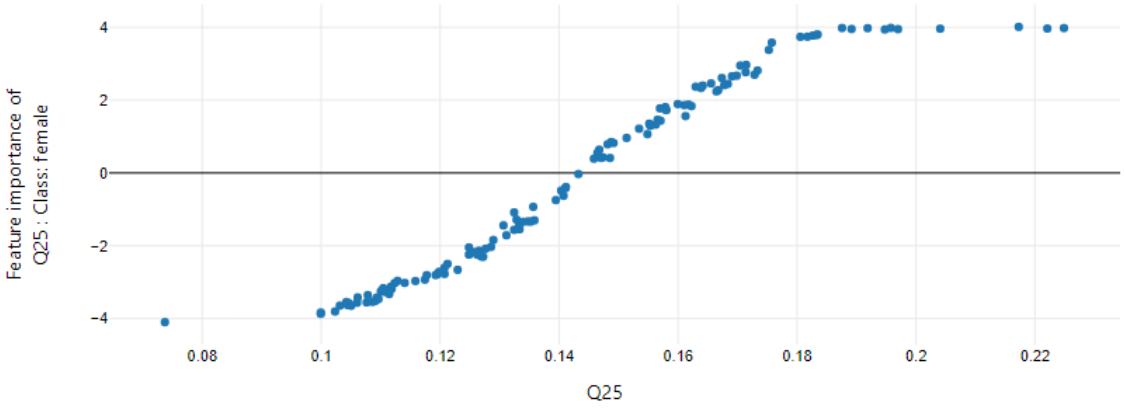

Figure 2. Dependence plot for Q25 on class ‘female’.

Figure 3 shows the dependence plot for the second most important feature average fundamental frequency (meanfun) for the class ‘female’. It shows that a meanfun over circa 135 Hz is most likely classified as female. Looking at the theory of voices, the average man had a meanfun of 85-155 Hz and the average woman 165-255 Hz (Fitch and Holbrook, 1970). This means that the dividing line should be at 160 Hz and not 135 Hz. The results therefore do not completely agree with the theory. However, they do show the same tendencies since both suggest that a higher pitched voice signifies a female and a lower pitched voice a male.

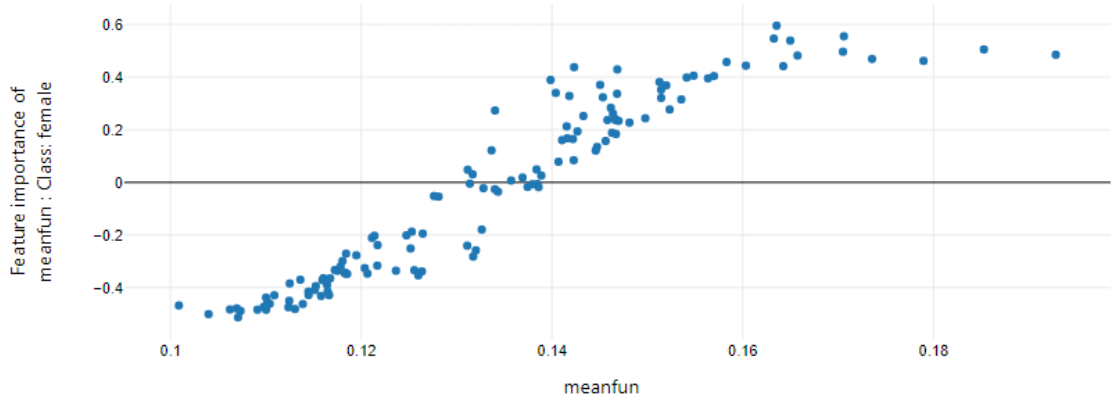

Figure 3. Dependence plot for meanfun on class ‘female’.

3.5 Testing

The results of the test data are displayed in this section, with pie chart of both the prediction and the solution, followed by the accuracy of the prediction.

Figure 4 shows the prediction of how much women and men spoke in test1 and Figure 5 shows the prediction of how much women and men spoke in the same recording. The accuracy was 83.1%.

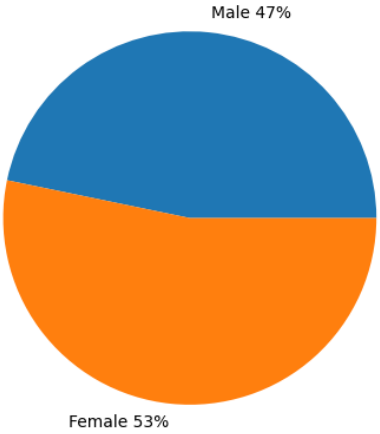

Figure 4. Pie chart of the prediction of how much women and men spoke in the recording test1.

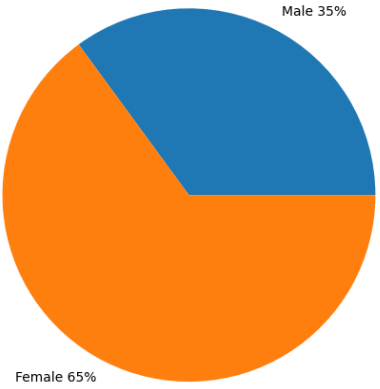

Figure 5. Pie chart of the solution of how much men and women spoke in test1.

Figure 6 shows the prediction of how much women and men spoke in test2 and Figure 7 shows the prediction of how much women and men spoke in the same recording. The accuracy was 83.5%.

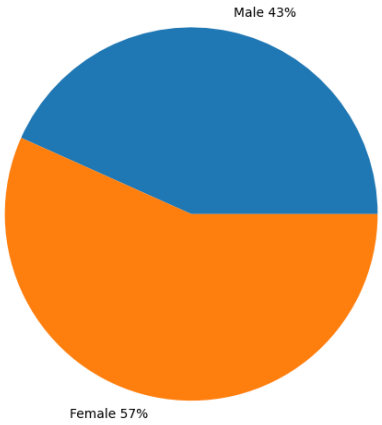

Figure 6. Pie chart of the prediction of how much men and women spoke in test2.

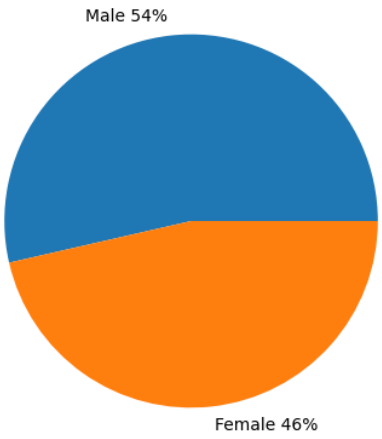

Figure 7. Pie chart of the solution of how much men and women spoke in test2.

4. Application

To facilitate the prediction of speakers in a meeting, a web application was created[5]. The application takes as input a sound file in a ".wav" format and outputs a pie chart of the prediction of how much women and men spoke in the sound file.

First, the silence is removed from the whole file. After that, it is broken down into several sections of 10s each. These are then used to construct features using specan3 (see section 2.2.3). To achieve a prediction of each section, a deployed rest API from the chosen model is called with all the data points. The API returns a prediction of each section, which is then used to create the pie chart. Figure 8 illustrates the flow of the web application.

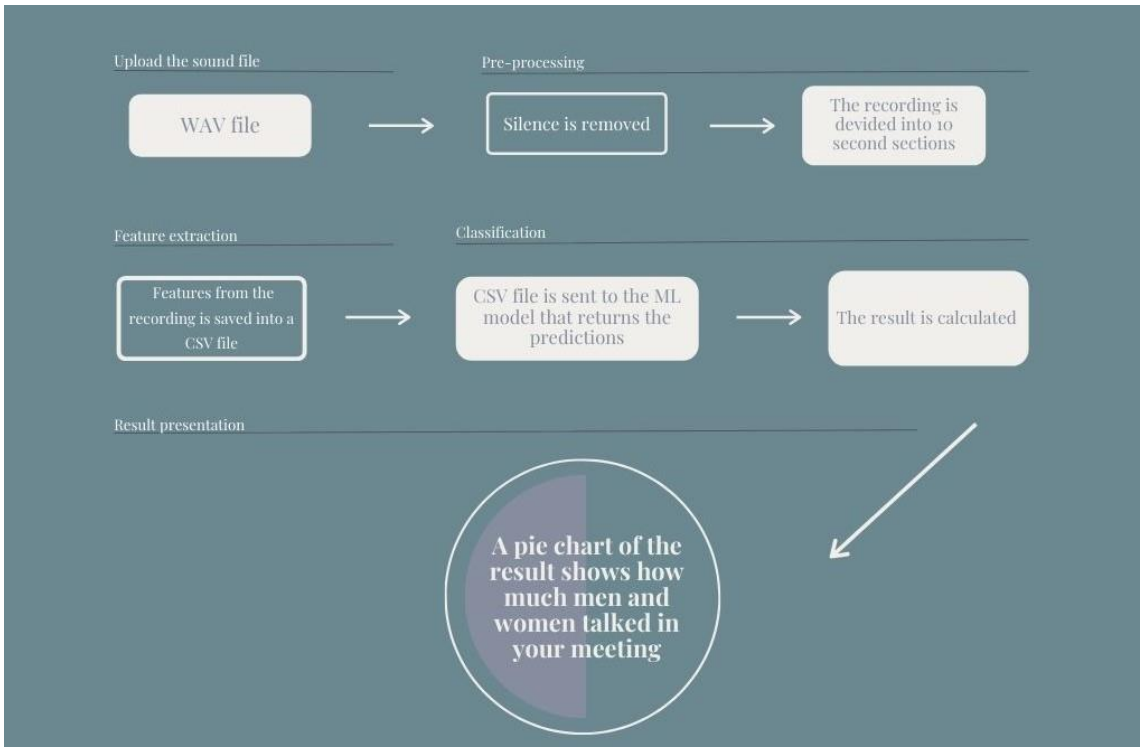

Figure 8. Flow chart of the web application.

[5] See https://github.com/mwellander/VoiceClassificationWebApp.

5. Discussion and Conclusions

The purpose of this project was to research how machine learning can be used to identify how much men and women talk in meetings. To answer this question, a training set was created using 1266 podcasts of people of different genders speaking. Four different ML models were created from four training sets with different characteristics in order to find the one yielding the highest accuracy. The models were trained using Microsoft Azure AutoML and the best algorithm for all four models turned out to be logistic regression with different scalers for each model.

To find the best model out of the four, the models were evaluated using nine podcasts as evaluation data. The results showed that there was no significant difference in performance between the models. Figure 1 shows that there are mainly one or two features that determine the prediction, which means that the feature construction of at least these two features are equal between the different training data sets. Additionally, the methods do not make as much difference and there is no need for a more complex algorithm than logistic regression. This makes it clear that the chosen data and pre-processing can be more important than the algorithms and model training itself. The final model was chosen to be the one with a slightly higher accuracy than the others on the evaluation data. The chosen model had a 92.1% accuracy on the training data and testing it on the two test meetings gave an accuracy of 83.1% and 83.5%.

This study found that how the pre-processing and scaling of the data is done can have a big impact on the accuracy, even in models using the same classification algorithm. Previous studies made by Becker (2016b) that get an accuracy of nearly 100%, have found logistic regression not to give an especially accuracy. However, no other studies have used scaling before training the models. All previous studies that use machine learning for this problem have also used the same dataset for training, whilst this model is trained on a new and unique dataset based on a different spoken language. The dataset in this study (1266 recordings) is also smaller than in previous studies (3168 recordings) which could be a reason for the different results. Even though the data set is smaller, it is beneficial to train models using new data, as well as data in other languages than English. This will lead to a more inclusion in the world of ML.

The model made in this study mostly classify voices with a mean fundamental frequency (SFF) above 135Hz as female. According to previous research about voices the fundamental frequency for females is between 165-255 Hz (Fitch and Holbrook, 1970).This can probably be explained by the fact that the model in this study uses several features to predict genders. The most important feature is the first quantile of the frequency (Q25) when classifying the gender of a voice, this shows that fundamental frequency is not as important when predicting the gender as previous studies indicate.

The accuracy of the test meetings does not yield the same accuracy as on the training data. This might be due to overfitting on the training data, but also since the characteristics of

the training data differ from the test data. In the training data, only one person speaks in each section, but in the test data, there may be more than one person and more than one gender speaking in one section. Looking at the results in the evaluation data shows that the model generally has a higher accuracy on meetings where there are people of only one gender speaking, than if there is more than one. It also has a better performance on podcasts with both female and male speakers where their speaking does not overlap very often. Another reason might be the fact that the solutions for the evaluation and test data are manually created, which may cause the accuracy to differ.

The feature importance graph (see Figure 1) shows that frequency is the most important feature in predicting a person's gender, which can lead to false predictions on people with a frequency within an androgynous range. This could be especially problematic with transgender or non-binary people. A further limitation of the model is that it only has two classes, does not take non-binary voices into account.

An important aspect of this project is how the application could be used in companies and what effect it might have on an organisation. There are some differences in the findings of what effect awareness of bias has on people. Results show that it could both reduce bias in a positive way (Devine et al., 2012), but also that if they are made aware of it right before an interaction it could cause people to inhibit themselves and therefore be perceived as more prejudice (Vorauer, 2012). This means that the use of this application could have the opposite effect if not used with careful consideration. A solution to this could be that the awareness is built over time and not solely in close relation to the meetings where equality is an issue. This would give participants time for reflection and an opportunity to modify their behaviour. It is difficult to say what exact effect this application can have on an organisation, but with support in related studies it is reasonable to assume it could have a positive effect and lead to a reduction in bias.

This application is built with the intent to be used to gather statistics for an organisation and different departments over time. Instead of informing participants about results directly after one meeting, all meetings at the company can be evaluated. This way, the application gives insight on the structural problem and might also provide insights of areas, departments or meeting types that can benefit from more equality driven work. An additional benefit of this approach is that no single person will be considered a scapegoat and the responsibility falls on each person in the organisation. Since studies have shown that only providing information about the structural issues can contribute to a more open-minded culture (Dirth and Branscombe, 2017), it might be enough to present the result from this app in a general way in order to see change. However, it probably needs more work and implementations alongside of this application to make a significant difference, which needs to be further researched before applying this application to a workplace.

As mentioned previously, the application was tested on an episode of 'Nours poddelipodd. Och Henrik.' (El-Refai and Schyffert, 2021) with approval from the creators of the podcast. They were also interested in the results in order to speak about it in their next podcast. The results they were given was the solution and not the prediction, which

was that El-Refai speaks 46% and Schyffert 54% of the time, as seen in Figure 7. When speaking about it in the podcast subsequent to the tested one, El-Refai talks about how she has been very aware of how much they speak to make it equal between them (El-Refai and Schyffert, 2021). This suggests that being aware of inequalities can make a situation more equal, since the distribution of speaking space is fairly equal in this podcast. Additionally, it is slightly planned beforehand, facilitating an equal agenda.

Using the model created in this project, it is possible to of identify how much men and women spoke in a meeting to a certain extent, however it is not always completely accurate. There are possible improvements to be made regarding the data and the pre-processing. To further develop this model, it would be beneficial to create more training data than the 1266 data points used in this project. Additionally, creating training data that more resembles the test data could be advantageous. For example, it could be changed into a regression problem, where the training data consists of both women and men speaking in each section and the label is the percentage of time women spoke in the section. Another interesting aspect of the problem would be to study if there is a difference in the way women and men speak and if it can be identified by an ML model. For example, if one gender tends to speak longer before pausing or interrupt others more often.

References

Akolor, K., Druid, D., Lundberg, V. & Nilsson, P. (2021) *Morgonpasset i P3 - Gästen: Krogkungen PG Nilsson – 40 år i branschen, grytprotesten och så beter du dig på krogen*. 3 November. [online]. Available from: https://sverigesradio.se/avsnitt/sa-beter-du-dig-pa-krogen-krogkungen-pg-nilsson.

Akolor, K., Druid, D., Lundberg, V. & Thyberg, N. (2021) *Morgonpasset i P3 - Gästen: Ninja Thyberg – Pleasure, porrindustrin och premiärbakfylla*. 5 October. [online]. Available from: https://sverigesradio.se/avsnitt/ninja-thyberg-pleasure-porrindustrin-och-premiarbakfylla.

Baccam, N. et al. (2021) What is automated machine learning (AutoML)? Microsoft Docs [online]. Available from: https://docs.microsoft.com/en-us/azure/machine-learning/concept-automated-ml.

Becker, K. (2016a) Gender Recongition by Voice. Kaggle [online]. Available from: https://www.kaggle.com/primaryobjects/voicegender.

Becker, K. (2016b) Identifying the Gender of a Voice using Machine Learning. Primary Objects [online]. Available from: http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/.

Becker, K. (2016c) *primaryobjects/voice-gender/sound.R*. [online]. Available from: https://github.com/primaryobjects/voice-gender/blob/master/sound.R.

Bell, J. (2020) *Machine Learning: Hands-On for Developers and Technical Professionals*. [online]. Available from: https://learning.oreilly.com/library/view/machine-learning-2nd/9781119642145/f01.xhtml.

Björklund, S. (2020) *Ljudböcker från Radioföljetången & Radionovellen: Radionovellen Borlängechampagne av Gustav Tegby i uppläsning av Sven Björklund*. 16 November. [online]. Available from: https://sverigesradio.se/avsnitt/radionovellen-borlangechampagne-av-gustav-tegby-i-upplasning-av-sven-bjorklund.

Büyükyılmaz, M. & Çıbıkdiken, A. (2016) *Voice Gender Recognition Using Deep Learning*.

Chamonikolasová, J. (2017) *Intonation in English Czech Dialogues*.

Devine, P. G. et al. (2012) Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*. 481267–1278.

Dirth, T. P. & Branscombe, N. R. (2017) Disability Models Affect Disability Policy Support through Awareness of Structural Discrimination: Models of Disability. *Journal of social issues*. [Online] 73 (2), 413–442.

El-Refai, N. & Schyffert, H. (2021) *Nours Poddelipodd. Och Henrik.: 63. Trollat och Imiterad*. 24 November.

El-Refai, N. & Schyffert, H. (2021) *Nours Poddelipodd. Och Henrik.: 64. Genusben*. 1 December.

Epstein, L. et al. (2021) *Nordegren & Epstein i P1: Kan katten få lika hög status som hunden?* 28 October. [online]. Available from: https://sverigesradio.se/avsnitt/kan-katten-fa-lika-hog-status-som-hunden.

Eriksson, A. & Traunmüller, H. (1995) *The frequency range of the voice fundamental in the speech of male and female adults*. 11.

Fitch, J. L. & Holbrook, A. (1970) Modal Vocal Fundamental Frequency of Young Adults. *Archives of Otolaryngology - Head and Neck Surgery*. [Online] 92 (4), 379–382.

Géron, A. (2019) *Hands-On Machine Learning with Scikit-Learn, Keras, and TensroFlow, 2nd Edition*.

Hahr, K. & Eklund, S. (2021) *Katarina Hahr möter: Sigge Eklund: Jag har känt mycket avund*. 24 May. [online]. Available from: https://sverigesradio.se/avsnitt/1730126.

Hahr, K. & Gyllenhammar, C. (2021) *Katarina Hahr möter: Cecilia Gyllenhammar: Det privilegierade livet har en baksida*. 16 August. [online]. Available from: https://sverigesradio.se/avsnitt/cecilia-gyllenhammar-det-privilegierade-livet-har-en-baksida.

Hale, J. (2019) Scale, Standardize, or Normalize with Scikit-Learn. towards data science [online]. Available from: https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02 (Accessed 10 November 2021).

Hoefel, R. (1991) Confronting Exclusionary Ideologies in the Classroom: Transforming Toward Inclusion and Diversity. *Transformations: The Journal of Inclusive Scholarship and Pedagogy*. 2 (2), 36–49.

Jacobi, T. & Schweers, D. (2017) Female Supreme Court Justices Are Interrupted More By Male justices And Advocates. Harvard Business Review [online]. Available

from: https://hbr.org/2017/04/female-supreme-court-justices-are-interrupted-more-by-male-justices-and-advocates.

Joshi, P. (2017) *Artificial Intelligence with Python*. Book, Whole. Birmingham: Packt Publishing, Limited. [online]. Available from: https://go.exlibris.link/H1t5f53W.

Kendall, S. & Tannen, D. (1997) *Gender and Language in the Workplace*. [online]. Available from: https://time.com/wp-content/uploads/2017/06/d3375-genderandlanguageintheworkplace.pdf.

Ledarna (n.d.) Vad jämställdhet är. Ledarna [online]. Available from: https://www.ledarna.se/stod-i-chefsrollen/jamstalldhet/jamstalldhet-vad-och-varfor/ (Accessed 7 September 2021).

Lind, K. (2021) *Snedtänkt med Kalle Lind: Om den kommersiella radion*. 28 October. [online]. Available from: https://sverigesradio.se/avsnitt/om-den-kommersiella-radion.

Lindholm, A. et al. (2021) *Machine Learning: A First Course for Engineers and Scientists*. [online]. Available from: http://smlbook.org/book/sml-book-draft-latest.pdf.

Make Equal (2018) *Jämställdhetsanalys av Jönköpings Kommunfullmäktige*.

Malmström, M. (2021) *Könsdiversifiering i bolagsstyrelser och företagsprestation*. 22.

Microsoft Docs (2021) *Configure data splits and cross-validation in automated machine learning*. [online]. Available from: https://docs.microsoft.com/en-us/azure/machine-learning/how-to-configure-cross-validation-data-splits (Accessed 12 November 2021).

Mogensen, L. et al. (2021) *Filosofiska rummet: Den hemliga donationen*. 24 October. [online]. Available from: https://sverigesradio.se/avsnitt/den-hemliga-donationen.

Pausewang Gelfer, M. & Schofield, K. J. (2000) Comparison of Acoustic and Perceptual Measures of Voice in Male-to-Female Transexuals Percived as Female versus Those Percived as Male. *Journal of Voice*. [Online] 11–33.

Raahul, A. et al. (2017) Voice based gender classification using machine learning. *IOP Conference Series: Materials Science and Engineering*. [Online] 263042083.

RDocumentation (n.d.) *specan: Measure acoustic parameters in batches of sound files.* [online]. Available from: https://www.rdocumentation.org/packages/warbleR/versions/1.1.2/topics/specan (Accessed 8 November 2016).

SCB (2017) *Styrelsemedlemmar efter funktion i aktiebolag år 2017*. [online]. Available from: https://www.scb.se/hitta-statistik/statistik-efter-amne/levnadsforhallanden/jamstalldhet/jamstalldhetsstatistik/pong/tabell-och-diagram/pa-tal-om-kvinnor-och-man-2020/lathund-2020-10.-inflytande-och-makt/.

Siegel, S. & Castellan, N. J. (1988) '*Nonparametric statistics for the behavioural sciences*', in 2nd edition p.

Sigh Raghuwanshi, B. & Shukla, S. (2021) Minimum class variance class-specific extreme learning machine for imbalanced classification. *Expert Systems With Applications*.

Stenholm, S. et al. (2021) *USApodden: Trump testar gränserna*. 20 October. [online]. Available from: https://sverigesradio.se/avsnitt/trump-testar-granserna.

Stoltzfus, J. C. (2011) Logistic Regression: A Brief Primer. *Academic Emergency Medicine*. [Online] 18 (10), 1099–1104.

Subasi, A. (2020) 'Machine learning techniques', in *Practical Machine Learning for Data Analysis Using Python*. S.l.: Elsevier Ltd. p. [online]. Available from: https://go.exlibris.link/zQ0tD6xR.

Sveriges Radio (n.d.) *Sommar & Vinter i P1*. [online]. Available from: https://sverigesradio.se/avsnitt?programid=2071.

Tannen, D. (2017) The Truth About How Much Women Talk - and Whether Men Listen. Time [online]. Available from: https://time.com/4837536/do-women-really-talk-more/.

Telus International (2021) The importance of natural language processing for non-English languages. Telus International [online]. Available from: https://www.telusinternational.com/articles/the-importance-of-natural-language-processing-for-non-english-languages?INTCMP=ti_lbai.

Toegel, G. (2021) 'Women talk too much' simply isn't true, data show. *imd.org*. February. [online]. Available from: https://www.imd.org/research-knowledge/articles/women-talk-too-much-simply-isnt-true-data-show/ (Accessed 7 September 2021).

Vorauer, J. D. (2012) Completing the Implicit Association Test Reduces Positive Intergroup Interaction Behavior. *Psychological Science*. 231168–1175.

Wellander, M. & Borgström, J. (2021) *Möte*.

Yeptain Leung et al. (2021) Associations Between Speaking Fundamental Frequency, Vowel Formant Frequencies, and Listener Perceptions of Speaker Gender and

Vocal Femininity--Masculinity. *Journal of Speech, Language & Hearing Research*. [Online] 64 (7), 2600–2622.