Project Final Report Outline

PROJ 201 Project Final Report

Project Title: **Voice Gender Identification of Videos**

Name, Surname & ID of group members: Osman Kantarcıoğlu & 30891

Supervised by: Ahmmad O.M. Saleh

Month, day, year 01,06,2024

Abstract

The hypothesis is detecting the gender of videos are possible. Therefore, the purpose of this project is to identify the gender of videos and find which method is better to identify the gender of voices. Finding which model is better is important because if we find which one is better, we may get more accurate results from our data. We used lots of data and steps for it and they are mentioned in the methods and materials part. In the end, we found that using the first model is better than the second model for our purpose which is detecting the gender of videos.

Introduction

Various coding tools are used for different purposes in today’s world. These tools facilitate people's life. For example, people may use tools to solve math questions or they may also use tools to calculate complex functions. Different problems may be solved by these tools. One problem that is solved by these tools is detecting the gender of the voice. The goal of the project is to detect the gender of the voice video and find the best model for it.

Methods & Materials

Let’s start with how our data looks like at the beginning of the project. Our heading is “STEM”. It involves five different fields. They are “Physics”, “Maths”, “ComputerScience”, “Chemistry” and “Biology”. For each field, we have ten queries, and each query involves twelve videos.

We used coding for detecting the gender of the voices. It involves seven different steps. Name of the steps are getting the ID of the video from our data, extracting the voice of the video, filtering, sampling, detecting the gender of the voice by using the first modal, detecting the gender of the video by using the second model and evaluation.

Preparation, in other words preprocessing, involves the first four steps. In the first step, as mentioned above we have 120 different videos for each field. We are getting the ID of those videos from our database which is listed above.

In the second step, we are extracting the voice from the video because we do need to audio detect the gender of the person on the video.

The third step is filtering. In the filtering part, we filter videos due to various reasons such as the length of the video and accessibility of the videos. We filter to save time and to save our memory.

The last step of preprocessing is sampling. In that step, we are taking samples from audio. For example, our samples are from the first ten seconds of every minute of the videos because we do not want to miss any sound and, we want to save memory and time too. If we take samples from each second, it takes too much time and it may lead to a memory problem. That was the end of our preparation part.

After preprocessing models start. The name of our first model is inaspeechsegmentor. In this model, our model classifies sounds as speech in terms of male and female, noise, and music. The role of the second model is also classifying. For example, if male time is obviously bigger than female time and also no speech time. It is probable that the person who is speaking in the video is male and we have the same logic for also female voice and no speech.

In the last step, which is called evaluation, we are evaluating our results and decide which model is better to detect the gender of the videos which was our goal and the main problem in this project.

Results

Model 1:

|  |  |  |  |
| --- | --- | --- | --- |
| MODEL 1 | PRECISION | RECALL | F1-SCORE |
| MALE | 0.96 | 0.89 | 0.93 |
| FEMALE | 0.90 | 0.93 | 0.91 |
| NEUTRAL | 0.18 | 0.43 | 0.25 |
| NA | 0.30 | 0.44 | 0.36 |

The first model is good at detecting the male voice of the video because it has a high precision rate. In addition, it is also good at detecting the female voice in the video because its precision is also high. However, not as high as detecting the male voice. On the other hand, this model is not good at Neutral and NA because they have lower precision rates as can be seen at the top.

Confusion Matrix For The First Model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL 1 | MALE | FEMALE | NEUTRAL | NA |
| MALE | 380 | 6 | 7 | 32 |
| FEMALE | 0 | 111 | 4 | 5 |
| NEUTRAL | 2 | 1 | 3 | 1 |
| NA | 12 | 5 | 3 | 16 |

If we connect the yellow parts at the table, we get a diagonal line. Values that are not on these diagonal lines are false predictions. Therefore, they are misclassification. As we write above, the first model is good at detecting the voice of the video which is specifically from one gender. And this model is also good at female voice because it has a few misclassifications. Thirdly, the Natural and NA part involves various many mistakes.

Model 2:

|  |  |  |  |
| --- | --- | --- | --- |
| MODEL 2 | PRECISION | RECALL | F1-SCORE |
| MALE | 0.95 | 0.75 | 0.84 |
| FEMALE | 0.71 | 0.85 | 0.77 |
| NEUTRAL | 0.02 | 0.14 | 0.03 |
| NA | 0.30 | 0.44 | 0.36 |

The second model is good at identifying the male voice of the video. It has a high precision rate. Moreover, it is not bad at detecting the female voice in the video. On the other hand, both Neutral and NA have low precision. Especially, the Neutral is the lowest.

Confusion Matrix For The Second Model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL 2 | MALE | FEMALE | NEUTRAL | NA |
| MALE | 319 | 29 | 45 | 32 |
| FEMALE | 5 | 102 | 8 | 5 |
| NEUTRAL | 3 | 2 | 1 | 1 |
| NA | 9 | 11 | 0 | 16 |

If we connect the green parts at the table, we get a diagonal line. Values that are not on these diagonal lines are false predictions. In other words, they are misclassification. As we mentioned above, this model is good at detecting the male voice in the video. As seen above, it made 336 predictions for male voice and 319 of them were correct predictions. It has only 17 misclassifications. Secondly, it is also good for the female voice because it rarely involves misclassification. Thirdly, it is not suitable for the Neutral voice because it only has one correct prediction. Lastly, the second model is also not suitable for the NA voice because it only has a few correct predictions.

Discussion and Conclusion

The first model is better than the second model in terms of accuracy rate. The data is imbalanced. The average of the f1-score should be calculated. The F1 score for the first model is 0.6075 and the f1 score of the second model is 0.5. Therefore, the first model is better than the second model for detecting the gender of the video which is the objective of this project.

The first and second models have similar rates for detecting the male voice. Therefore, both of them can be used for it. However, the first model is better than the second model in terms of detecting the female voice in the video because the first one has high rate than the second one.

Thirdly, the first model is better than the second model for detecting neutral voices. Lastly, the first model and the second model have a second accuracy value for NA voice. To sum up, first one is better than second one for detecting the gender of the voice.