Natural Language Processing

Report on the project :

Character gender classification from Video Games dialogues
using Stephanie Rennick and Seán G. Roberts Corpus

https://github.com/seannyD/VideoGameDialogueCorpusPublic/

by **Adam MIR-SADJADI**
https://github.com/ADMR-S/NLP_VG_Gender_Classification

Computer Science Master 1 Student
Schoolyear 2024-2025, Semester 2
Université Côte d'Azur

April 21, 2025

Contents

1 Introduction

1.1 Context

Video Games are not exempt from gender biases. That is what the work [1] of Stephanie Rennick Melanie Clinton, Elena Ioannidou, Liana Oh, Charlotte Clooney, E. T.†, Edward Healy and Seán G. Roberts showed through a thorough analysis of dialogue-based video games scripts they gathered in the *Video Game Dialogue Corpus*[1] [2]. This study showed amongst its results that the majority of games analyzed had more masculine characters than female, this disproportion being the main reason why there are less dialogue lines from female character than from male ones in the different scripts. The proportion of male vs female characters is around 70% to 30%, averaging all games analyzed, with the gap tending to reduce over the years. Nevertheless, you can still find games released in the 6-year span that preceeded the study which contains about only 25% of female characters. Dialogues pronounced by female characters can also tend to present strongly genre-connoted features, such as being more polite or revolve around subjects such as family and relationships.

Taking in consideration this observed bias, we will try to build a classifier that we will train on several game scripts of this same corpus, to determine if a given dialogue line belongs to a female or a male character. First, we will try to determine if a given line is an action or a dialogue, to ensure our classifier can distinguish between two simple categories, before moving on to genre classification of dialogues.

1.2 Materials & Data preparation

To accomplish this task, we tried different models from scikit-learn through a jupiter notebook available on github[2]. The different Classifier models we tried are : Multinomial Naive-Bayes, Multi-Layer Perceptron, Random Forest, Stochastic Gradient Descent and Support Vector Machine.
The Video Game Dialogue Corpus on github did not contain the game scripts directly but provided all necessary resources to build them. By scraping from different sources, it was able to provide us with two source files for several games (some games data were not recoverable as it required owning the games and having it installed which was not our case, and others were linked to online sources which are no longer available).

The games analyzed in our experiment are the following : The Elder Scrolls V : Skyrim, Horizon Zero Dawn (small dataset, only 1 mission for 400 lines), Final Fantasy XIII-2, Hades and Death Stranding. They were choosed because

[1] https://github.com/seannyD/VideoGameDialogueCorpusPublic/
[2] https://github.com/ADMR-S/NLP_VG_Gender_Classification

their scripts were complete enough after building the Corpus, and additionally because they present different gender repartition of their characters and number of dialogues, and I have played all of them (therefore I can confront my expectations with their respective bias and I can't spoil myself from their scripts !).

Figure 1: Visuals for the games used in our experiments

The two resulting files for recoverable games are the data.json file which contains the whole script for its corresponding game and the meta.json file which contains additional information such as the character groups corresponding to each character (male, female, neutral, playerChoice or unknown). An overview of such files is the following :

Listing 1: Data.json Overview (from Skyrim)

```
{ "text" :
    {"Whiterun Gate Guard": "Halt!
        City's closed with dragons
        about. Official business
        only."},
    {"CHOICE": [
        [
        {"Dragonborn": "Riverwood
            calls for the Jarl's
            aid."},
        {"Whiterun Gate Guard": "
            Riverwood's in danger,
            too? You better go on
            in. You'll find the
            Jarl at Dragonsreach,
            atop the hill."}
        ],
        [
        {"Dragonborn": "Stand
            aside, or else.
            (Intimidate)"},
        {"Whiterun Gate Guard": "
            Or else what? You
            think you can stand
            against the entire
            Whiterun city guard?
            The gate's closed."}
        ]
    ]},
    {"ACTION": "[At the
        Dragonsreach castle, the
        prisoner approaches
        Balgruuf, already talking
        with his advisors.]"}
}
```

Listing 2: Meta.json Overview (from Skyrim)

```
{
   ...
   "characterGroups": {
      "playerChoice":
          ["Dragonborn"],
      "male": [
         "Brynjolf",
         "Ulfric",
         ...
      ],
      "female": [
         "Alea",
         "Serana",
         ...
      ],
      "neutral": [
         "Imperial Soldiers",
         "Kids",
         ...
      ]
   }
   ...
}
```

As we can see, an entry in the data.json file can have three different types : Action, Choice, or just a line of dialogue. Choices are made by the player and can contain other choices or actions. They symbolize an embranchment in a dialogue or an action, and are sometimes exclusive. We considered each line in the script, as if the player was able to make every decision at once, to get a full overview of the game's script. Therefore, the first processing of the data applied in our experiment consists in reading the whole script for a given or several games and to build a dataset made from every action or dialogue which appears in it. We also get rid of unexpressive actions (such as {"ACTION" : "—

"} which occured often) or unappropriate formatting which might have led to a high bias in our trained model (every skyrim action was written between braces, so we got rid of them, for example). Finally, some scripts presented labels others than Action, Choice or dialogue which were ignored in our processing (usually they contained additional information about a given line and were not relevant to our experimentation). "SYSTEM" entries were nevertheless considered as actions, as they had a similar syntax and semantic in the games they were observed. Some of the used games scripts don't have actions at all or very little, such as Hades or DeathStranding, so they were only really used for the gender classification. Moreover, the DeathStranding meta.json file didn't contain the character groups (it was still a WIP), so I added them manually, putting the characters I couldn't assign clearly to a group in the "unknown" category.

The characteristics of our datasets are the following :

Game_ID	**Entries**	**Actions**	**Dialogues**	**Male Char.**	**Female Char.**	**Genre**
Skyrim	8892	1217	7675	3598	1580	RPG
Horizon Zero Dawn	329	116	213	96	113	Adventure
Final Fantasy XIII-2	7006	4292	2714	2538	1747	Fantasy
Hades	23571	0	23571	16148	4143	Roguelike
Death Stranding	1578	16	1562	1056	394	Narrative

Table 1: Games dataset compositions

2 Methods

Our approach was to try several gridSearches fit with different pipelines on the Action/Character classification to determine which model would be the most suited for our gender-classification task. This is based on the assumption that if our classification is successful on the Action/Character classification, it would seem that our parameters enable the model to detect dissimilarities between NL extracts. If such dissimilarities exist between the male and female dialogue sets, then our model, once trained on this new classification task, might detect it.

Considering the repartition was inequal between the classes we classified, we used an F1 Score rather than trying to maximize accuracy, ensuring our model would try to optimize precision and recall and thus hopefully not predict the majority class each time (giving it an accuracy equal to the proportion there is of this class). We also monitored the splitting of the sets between training and testing through statistics to ensure there was a proportional repartition in each of them and to be aware of our datasets composition. This equal repartition might give an additional hint to our model as for the amount of each classes it has to predict but this characteristic is inherent to our corpus and we are in a first time interested in seeing if the model manages to answer well given this information.

We took the Skyrim dataset as a starting point to compare the classifiers, as it has numerous Actions which are not always easily distinguishable from Character dialogues, knowing the male/female classification which came after would be more subtle. We run tests on other datasets along the way to assess our results were not specific to the skyrim dataset. Once we had compared the models we wanted on the Action/Character classification for this game, we took the two most promising ones and ran tests on each of our games datasets with it to see if some of them presented particular results and to get a better understanding of our corpus.

We tried to improve our gridSearch parameters over the iterations by analyzing which were the most choosed parameters, and which changes seemed to have a positive impact on our models results. We always kept in mind that the inequal repartition of our classes (action/character and male/female) as well as the different style of writing between characters and games might lead to various biases or overfitting on a given set, so we also tried our most promising models on unseen games dataset (training for example a model on 4 games and testing it on the 5th one). For each prediction, we saved the cases where the model was wrong to try and understand if it might be missing something or if the failed predictions are as hard to classify from a human point of view.

The beginning of the pipeline is constituted of a CountVectorizer and a TF-Idf Transformer. We tried other configurations such as preprocessing the text manually with spacy, using a Hashing Vectorizer... but none gave promising results so they were abandonned. Finally, as some gridSearches fit took way too long for the scale of our project (more than 2000 minutes and still processing for some MLP Classifier configurations), we added an option to reduce the size of the dataset of 80% when necessary.
We used 80% of the given dataset for training, and 20% for testing (so if it has been reduced by 80% in the first time we train on 80% of 20% of the original data size).

3 Results & Discussion

The majority of our results are available in the output folder of the project (we saved the classification reports, the confusion matrix, the best parameters, the cross-validation performance, the loss curve when there is one, the wrong cases, etc...)

3.1 Action/Character Dialogue Classification

Using the Skyrim dataset, here are the results we obtained for each Classifier type :

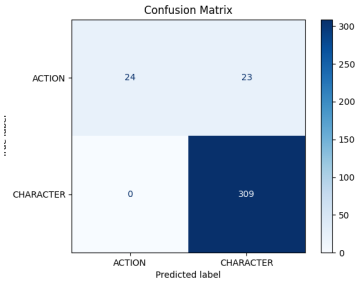

Figure 2: Skyrim MLP

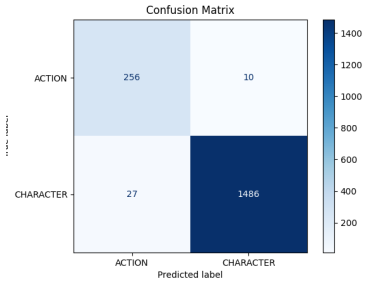

Figure 3: Skyrim Multinomial NB

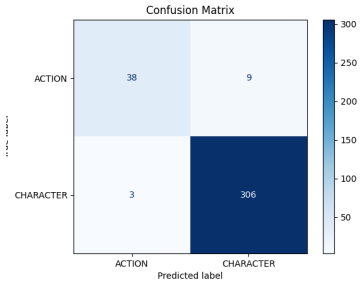

Figure 4: Skyrim SGD Classifier

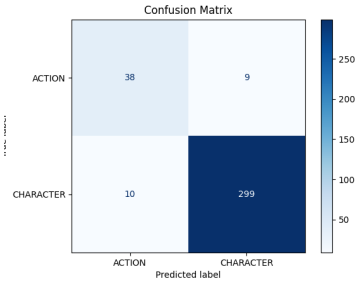

Figure 5: Skyrim Random Forest

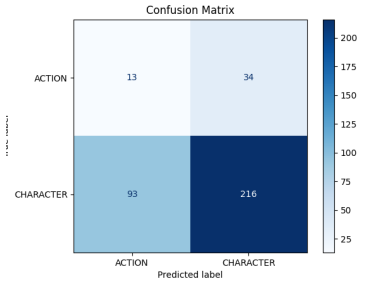

Figure 6: Skyrim SVC

The parameters for each of these models (and the ones presentend later) can be found in the output/parameters/ folder, listing them all here wouldn't be appropriate considering the amount of models we tried. These results show that the best classifiers for the first classification are, in order (with their F1 Score) : MultinomialNB (0.98), SGDClassifier (0.97), RandomForest (0.95), MLPClassifier (0.94) and finally SVC (0.64). However, these scores on SGDClassifier and the RandomForest were obtained later than our first iteration, and the best models at the time were the MultinomialNB and the MLPClassifier. Therefore, we continued with these two models but the SGDClassifier as well as the RandomForest might perform better than the models we built if well optimized.

Keeping the MultinomialNB and the MLPClassifier, here are the results we obtained on two other games from our dataset (FFXIII-2 and Horizon Zero Dawn, considering Hades and Death Stranding don't have relevant actions in their scripts and therefore only have "Characters") :
For MultinomialNB :

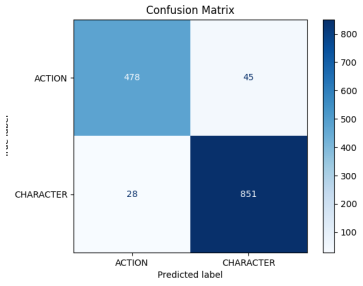

Figure 7: FFXIII-2 MultinomialNB

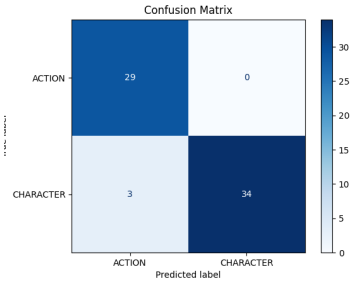

Figure 8: Horizon Zero Dawn MultinomialNB

For Multi-Layer Perceptron :

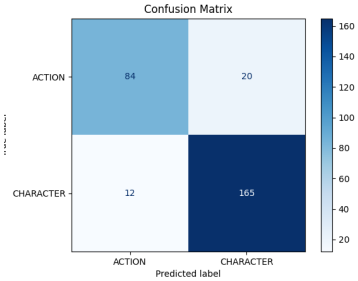

Figure 9: FFXIII-2 MLP

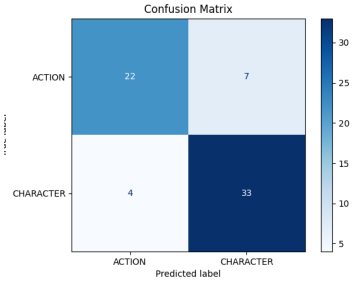

Figure 10: Horizon Zero Dawn MLP

As we can see, both models perform well on every set tested, with the MultinomialNB still showing much better results (0.95 vs 0.89 for MLP on FFXIII-2 data and 0.95 vs 0.83 on Horizon Zero Dawn data). By taking a look at some of the error cases, we obtain the following snippet :

Listing 3: FFXIII-2 MLP error cases

```
Predicted: ACTION, Actual:
    CHARACTER, Dialogue:
  Yeul...
Predicted: CHARACTER, Actual:
    ACTION, Dialogue:
  Player chooses Square - Is
      Lightning in there, too?
Predicted: ACTION, Actual:
    CHARACTER, Dialogue:
  Lightning?
Predicted: ACTION, Actual:
    CHARACTER, Dialogue:
  Final Fantasy XIII-2: The story
      so far...
Predicted: ACTION, Actual:
    CHARACTER, Dialogue:
  Whaddya mean, 'Time Gate'?
Predicted: CHARACTER, Actual:
    ACTION, Dialogue:
  Lebreau is knocked back by an
      attack.
```

Listing 4: FFXIII-2 MultinomialNB error cases

```
Predicted: CHARACTER, Actual:
    ACTION, Dialogue:
  or
Predicted: ACTION, Actual:
    CHARACTER, Dialogue:
  Pathetic.
Predicted: ACTION, Actual:
    CHARACTER, Dialogue:
  (lighting up again) Take it,
      and leave this place! (He
      produces a fragment.) The
      Day of Reckoning has not
      yet come.
Predicted: CHARACTER, Actual:
    ACTION, Dialogue:
  If she keeps going the wrong
      way...
Predicted: ACTION, Actual:
    CHARACTER, Dialogue:
  (nods) The Oracle Drive! (A
      light flashes from the
      drive and Serah's weapon
      lights up. )
```

From the error cases, we observe one source of errors lives in the fact that some dialogues are ambiguous. Such a line : "Player chooses Square - Is Lightning in there, too?" or "(lighting up again) Take it, and leave this place! (He produces a fragment.) The Day of Reckoning has not yet come." can be considered both as an action and a dialogue (actually it should be separated in two parts). Therefore, the dataset composition is a source of error in itself. Considering this, our models perform quite well even if we still acknowledge that it is just misbehaving sometimes ("Final Fantasy XIII-2: The story so far..." is a recurring line and the model is wrong several times about it, knowing it contains the game's name in itself and no character ever pronounces it, it is quite surprising that the MLP isn't able to categorize it well). We could have added decision paths to enhance the dataset or give indications to our models but we preferred not to interfere with the learning task, otherwise classifying these dialogues with a human help shouldn't be so difficult. We will not extend furthermore on the error analysis, as it was mainly an iterative process while parametering the pipeline, and it is mainly subjective interpretations.

Nevertheless, our results are particularly promising and we can move on to the gender classification task to see if we still get promising results such as these.

3.2 Male/Female classification from dialogue

Using the skyrim dataset, we also tried every classifier on the gender classification to ensure our models were still ranked in the same order.

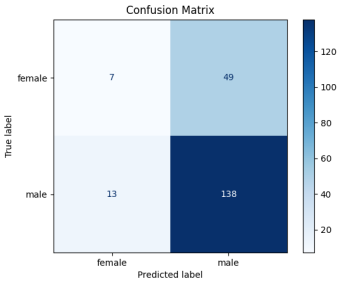

Figure 11: Skyrim MLP

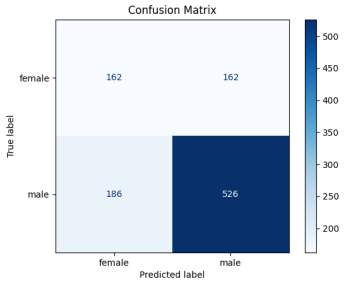

Figure 12: Skyrim Multinomial NB

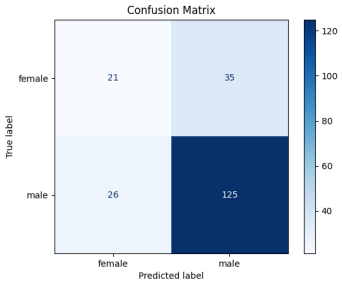

Figure 13: Skyrim SGD Classifier

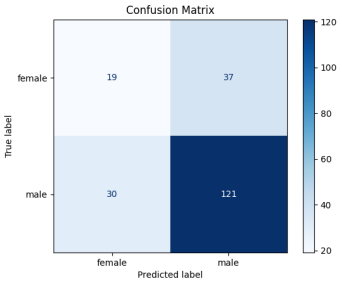

Figure 14: Skyrim Random Forest

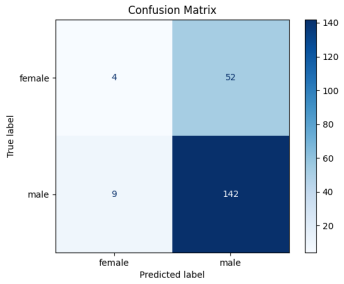

Figure 15: Skyrim SVC

Here, the results are much less successful. The ranking is much harder to establish as an overall good F1 score might be caused by an extremely good recall or an extremely good precision, but not necessarily by both. Considering male dialogue lines are much more represented in the majority of our datasets, some models might achieve higher scores just by guessing male repetitively (it's still better than if we used accuracy but it doesn't mean results are balanced either).
Ranking is now the following for the global score, but sometimes with a terrible score over female classification. Results are written as such - Classifier(overall F1, male F1 - female F1) : SVClassifier(0.71, 0.82 - 0.12), SGDC(0.71, 0.8 - 0.41), MLP (0.7, 0.82 - 0.18), RandomForest (0.68, 0.78 - 0.36) and MultinomialNB(0.66, 0.75 - 0.48). It is much harder to find balanced results, the SVC almost exclusively guessing male and yet achieving the highest score (though not so high considering the proportion of male dialogue lines, which is about the same proportion, 70%). This might underline the fact that there is not much difference between male and female dialogue lines or that our models are not performing well enough.
The multinomialNB, while being the last in the ranking, still performs the best on female classification, with not as much error as the other models. Having obtained such results with the SGDC only recently, we kept going with the MLP and the MultinomialNB for our other games.
Here are the results we obtained on the 4 other games :

For MultinomialNB :

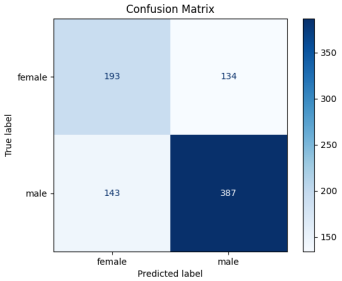

Figure 16: FFXIII-2

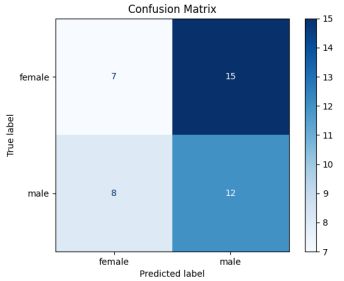

Figure 17: Horizon Zero Dawn

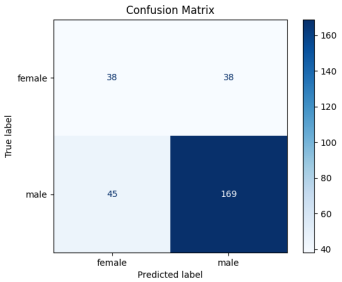

Figure 18: Death Stranding

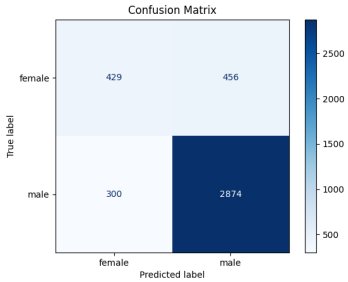

Figure 19: Hades

Figure 20: Comparison of MultinomialNB classifier on several games

For MLP Classifier :

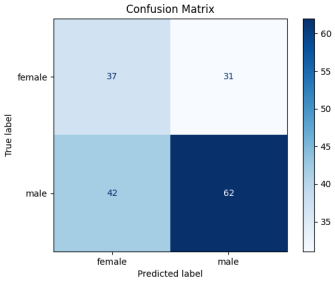

Figure 21: FFXIII-2

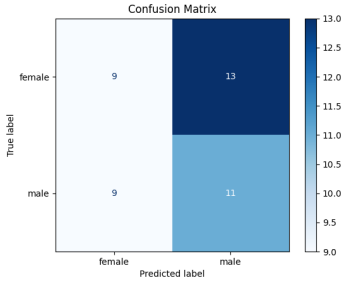

Figure 22: Horizon Zero Dawn

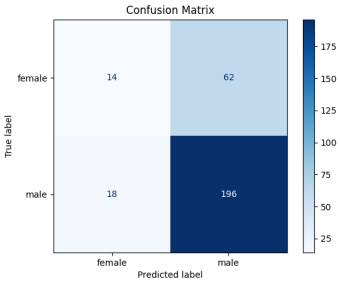

Figure 23: Death Stranding

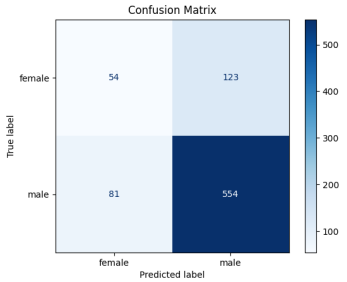

Figure 24: Hades

Figure 25: Comparison of MLP classifier on several games

Depending on the game, the classification might perform better on both models, while being overall unsuccessful. For example, we already see on the MultinomialNB confusion matrix that the FFXIII-2 dataset allows for a better classification than on the other games. The rankings are the following for both models :

MultinomialNB : Hades(0.81, 0.88 - 0.53), Death Stranding(0.71, 0.8 - 0.48), FFXIII-2(0.68, 0.74 - 0.58) , Skyrim(0.66, 0.75 - 0.48), Horizon Zero Dawn(0.45, 0.51 - 0.38).

MLP : Hades(0.75, 0.84 - 0.35), Death Stranding(0.72, 0.83 - 0.26), Skyrim(0.7, 0.82 - 0.18), FFXIII-2(0.58, 0.63 - 0.5), Horizon Zero Dawn(0.48, 0.5 - 0.45).

Indeed, the game on which we classify female dialogues correctly the best is FFXIII-2. However, the game on which we get the best overall score is Hades, probably because of its 16 000 male dialogues vs 4 000 female ones. Yet, with the MultinomialNB on Hades, we still get more than 0.5 as F1 score for the female classification, which might indicate the model is able to catch on some dissimilarities between the two classes. Nevertheless, characters style of writing (not taking the gender into account but a given character identity as a whole) might also allow the model to recognize characters it has already seen but not necessarily because they share genre-connoted dialogues. While being unclear, these results are interesting to discuss and would benefit to be put in parallel to a more thorough analysis of the given games bias.
Horizon Zero Dawn, while being the only game with more than 50% of female dialogues, remains the game on which we get the poorest results. This might come from the small size of the dataset, not allowing the model to learn enough (which would indicate there is a slight bias in the other games that it manages to catch), but also from the fact this game's dialogues might be less genre-connoted than others.

When looking at the error cases for example for FFXIII-2 as above, we obtain this time :

Listing 5: FFXIII-2 MLP error cases

```
Predicted: male, Actual: female,
    Dialogue:
  Hmm. I'll supply you both with
      comm devices. Right this
      way, please.
Predicted: male, Actual: female,
    Dialogue:
  (auto-talk) We should talk to
      those people. Maybe they
      can tell us what's going on.
Predicted: male, Actual: female,
    Dialogue:
  This has gone too far. (She
      sucks in a deep breath and
      shouts.) That's enough!
      I've about had it with your
      mischief! You should be
      ashamed of yourselves for
      behaving like this! You
      should know better! Well?
      What do you have to say for
      yourselves?
Predicted: female, Actual: male,
    Dialogue:
  Good. Why don't we go and find
      Lightning?
Predicted: female, Actual: male,
    Dialogue:
  I wonder if he really meant it.
      Maybe you didn't want to
      go, and you were glad to
      have the excuse.
```

Listing 6: FFXIII-2 MultinomialNB error cases

```
Predicted: male, Actual: female,
    Dialogue:
  (auto-talk) We're still inside
      the dreamworld. How do we
      get out?
Predicted: male, Actual: female,
    Dialogue:
  How could you tell?
Predicted: male, Actual: female,
    Dialogue:
  No life or death? You mean like
      Valhalla. That's what Caius
      is after! If he can bring
      down Cocoon, millions of
      people would die, and the
      power of chaos... ... would
      turn this world into
      another Valhalla.
Predicted: female, Actual: male,
    Dialogue:
  You don't believe me?
Predicted: female, Actual: male,
    Dialogue:
  Oh, wow. You grow vegetables?
```

Looking at the error cases, except special cases where there is an indication inside the dialogue such as "(she sucks in a deep breath)" (where the POS gives us enough information to determine the character is a woman), the majority of error cases are ambiguous, even for us. It is therefore normal to get more mitigated results, but we could enhance our classification by focusing on the lines we are able to classify, by determining why our model isn't able to do so.

Lastly, we tried to train a model on 4 games and test it on the last one, to get an idea of how overfitted our model is (to its training data but most of all to it's corresponding game). We trained it on every game except Hades and tried it on this one, being the game with the highest overall score in each model.

Here are the results we obtained :

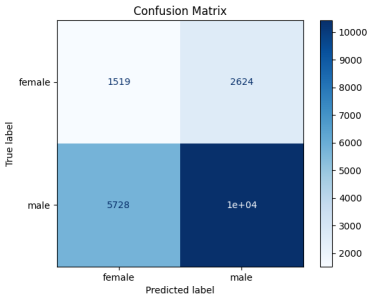

Figure 26: 4 games trained MultinomialNB vs Hades

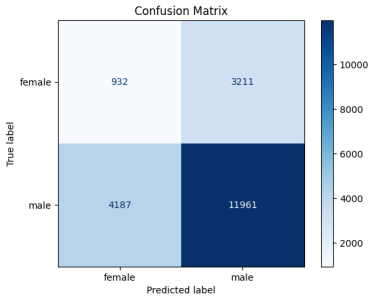

Figure 27: 4 games trained MLPClassifier vs Hades

Results are not good at all considering the female classification, while results on test-sets of the 4 original games were better :

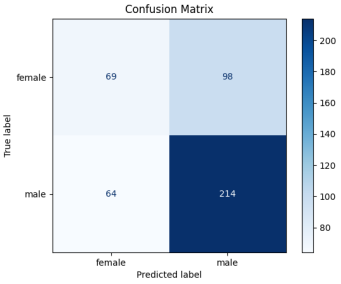

Figure 28: 4 games trained MultinomialNB test results

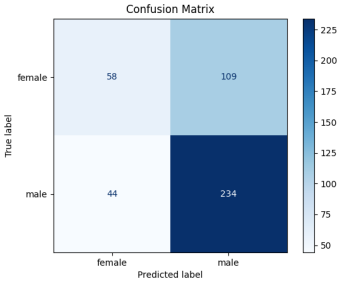

Figure 29: 4 games trained MLPClassifier test results

From this, we are able deduce that our model is only efficiently trained to recognize samples from games it already knows. While not necessarily too overfitted to its training data as a subset, it is definitely only able to predict from samples written in a style it knows. Moreover, for MLP Classification, loss vs validation loss curves which were plotted and are available in the output folder (not included here as the report is already quite heavy) seems to indicate the MLP classifiers perform badly outside of their training set, explaining the better results of the MultinomialNB. There is no real convergence on the gender classification which definitely indicates overfitting on its training set.

4 Conclusion and possible improvements

Our models perform quite well on the Action/Character classification. Indeed, there are strong NL markers allowing to determine a given line is an action, the task being easy for humans and not too hard for our models apparently. Moving on to the gender classification, the task becomes much more subtle for humans as for machines. MultinomialNB performs well overall, and we also focused on MLPClassifiers but obtained poor results from it in the end (considering the amount of time consumed to train them compared to the multinomialNB, there is no match. MultinomialNB performs way better, but running bigger gridSearches might lead to an inversion of those results). SGDClassifiers might have a lot of potential, revealing better metrics than MLP in our last tests. The games analyzed presented different dataset characteristic, are written in different writing styles, and show an interesting range of the video game industry (american and japanese games, from 2010 or recent (2019)). A bigger set might be interesting to analyze, I would've liked to compare games from the same series over time but I didn't get the necessary scripts from scraping. Overall, we principally scratched the surface which allowed us to get a better understanding of the mechanisms at play in NL classification.

5 References

[1] Stephanie Rennick, Melanie Clinton, Elena Ioannidou, Liana Oh, Charlotte Clooney, E. T., Edward Healy, Seán G. Roberts (2023) Gender bias in video game dialogue. Royal Society Open Science 10(5). https://royalsocietypublishing.org/doi/10.1098/rsos.221095

[2] Stephanie Rennick, Seán G. Roberts, (in press) The Video Game Dialogue Corpus. Corpora 19(1). preprint