Music genre classification based on song lyrics - comparison between different word embedding techniques and classifiers
Proof of Concept report for NLP Course, Winter 2022

Bartłomiej Eljasiak
Warsaw University of Technology
bartlomiej.eljasiak.stud@pw.edu.pl

Aleksandra Nawrocka
Warsaw University of Technology
aleksandra.nawrocka.stud@pw.edu.pl

Dominika Umiastowska
Warsaw University of Technology
dominika.umiastowska.stud@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

Music genre classification (MGC), although a well-known task, still remains challenging in the domain of Music Information Retrieval. We tackle the problem of MGC based solely on lyrics and try to solve it using a solution composed of a state-of-the-art word embedding method tuned for this problem and a separate classification model. Our main contribution is the comparison between different word embedding methods, classification techniques and optimization techniques, which in the domain of MGC is currently lacking. The novelty comes with an additional approach in the form of testing the impact of enriching the words with their sentiment obtained using a separate model.

1 Data preprocessing

In order to prepare our data for usage we had to process it appropriately. This was done in a few steps. The result of the final step needed to be a text which can be directly fed into the embedding algorithm that produces the embeddings. First, we filtered out the non-english lyrics, since the encoders we are going to use were trained on this specific language. In the following step, we removed artefacts and unimportant words from our lyrics, such as commas, punctuation marks or words like *verse* or *chorus*, as they typically were not part of the lyrics, but represented the structure of the songs. In the last step, we removed infrequent classes from one of the datasets. This action did not change the fact that our problem is significantly unbalanced since music genres' counts differ hugely from one another in both datasets.

2 Current state of the project

Our project is divided into two parts. Currently, we are working on the main part, which consists of testing different embedding methods and classifiers for the MGC problem. To this date, we have tested one embedding method (GloVe) with four classifiers: Naive Bayes, linear SVM, XGBoost and CNN. Tests were performed on the MetroLyrics dataset. The results are in Table 1.

Classifier	**Accuracy**	**Balanced accuracy**	**F1 score**
Naive Bayes	15.43%	16.48%	9.66%
Linear SVM	46.18%	20.55%	42.78%
XGBoost	30.39%	13.20%	31.17%
CNN	50.91%	25.84%	48.72%

Table 1: Current results

In our implementation, we make use of the following libraries for embedding methods and classifiers:

- Spark NLP/NLU (GloVe),
- sklearn (Naive Bayes, linear SVM),
- xgboost (XGBoost).

As GloVe produces tables of size $\{\text{number\_of\_words}\} \times 100$ as output and we need vectors of the same length as inputs to our classifiers, we decided to, firstly, resize the embeddings to size 400×100 (by taking first 400 rows or adding rows with zeros) and secondly, flatten the resulting tables.

3 Project proposal reviews discussion

3.1 Strong points analysis

We would like to shortly summarise a few aspects which were positively viewed. Reviewers seemed to be satisfied with our choice of the datasets, they found them e.g. ”interesting”. They were also interested in seeing various experiments that are planned and their results. Our overall score was 4.25 out of 5 which lands somewhere between ”Strong” and ”Exciting”. This ensures us that goal of our project is worth pursuing and as a team, we are on a good track.

3.2 Weak points analysis

Whereas positive opinions bring some amount of satisfaction, one of the key purposes of every review is to find weak points in the work. They may cast a light on committed mistakes or wrong assumptions. We received a fair share of criticism, nevertheless, in our opinion, the process that we presented is still valid, yet it might benefit from a few changes. It was pointed out to us that we should justify the novelty of our work slightly better. One reviewer raised a concern that our classification models might be too simple, and in some cases this might be true, but we do not expect great performance from them all. This work is oriented on the comparison of different approaches and in our opinion it will be highly beneficial to see by how much e.g. Naive Bayes performs worse than CNN. Additionally, we received a couple of comments about the presentation itself, which are irrelevant going further in the project. Overall we are satisfied with the criticism we received, mainly due to the fact that reviewers pointed out several possible improvements and they did not spot any critical mistakes or flaws in our understanding of the matter. This leaves us ensured that the project can be continued without the necessity for any major changes and the obtained results will be valuable to other scholars.

4 Authors' contribution

Self-assessed authors' contribution is in the Table 2.

Author	**Contributions**	**Workload**
Bartłomiej Eljasiak	Exploratory data analysis, part of the report	32%
Aleksandra Nawrocka	Data preprocessing, refactorization of training code, part of the report	33.5%
Dominika Umiastowska	Preparation of models and training code and conducting tests	34.5%

Table 2: Authors' contribution