Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Εργαστηριο Συστηματων Τεχνητης Νοημοσυνης και Μαθησης

Explaining Multimodal Music Emotion and Genre Recognition

Diploma Thesis

by

Sotirou Theodoros

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Πληροφορικής
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Explaining Multimodal Music Emotion and Genre Recognition

Diploma Thesis

by

Sotirou Theodoros

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 17η Ιουλίου, 2024.

.....	.....	.....
Γεώργιος Στάμου	Αθανάσιος Βουλόδημος	Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2024

.....

ΣΩΤΗΡΟΥ ΘΕΟΔΩΡΟΣ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η Ανάκτηση Πληροφοριών Μουσικής (MIR) είναι ένας τομέας έρευνας που ασχολείται με την εξαγωγή και ανάλυση πληροφοριών από τη μουσική. Μεταξύ άλλων, περιλαμβάνει την παλινδρόμηση/ταξινόμηση μουσικής και συγκεκριμένα την ανίχνευση διάθεσης και την αναγνώριση είδους. Παράλληλα με την ανάπτυξη που παρατηρείται στους τομείς της τεχνητής νοημοσύνης (AI), η MIR έχει επίσης σημειώσει σημαντικές προόδους, συμπεριλαμβανομένης της διαθεσιμότητας εκτεταμένων συνόλων δεδομένων, της ενσωμάτωσης νέων τεχνολογιών και πολυτροπικών προσεγγίσεων καθώς και της ανάπτυξης και εφαρμογής προηγμένων μεθόδων επεξηγησιμότητας.

Σε αυτήν τη διατριβή, εμβαθύνουμε στην επεξήγηση πολυτροπικών μοντέλων για την ταξινόμηση των συναισθημάτων και των ειδών της μουσικής. Πρώτα απ' όλα, αναζητούμε διαθέσιμα σύνολα δεδομένων που παρέχουν πολυτροπικές και πολυ-εργασιακές δυνατότητες. Επιλέγουμε το Music4All [54], που προσφέρει στίχους και ήχο καθώς και μεταδεδομένα συναισθημάτων και ειδών για κάθε τραγούδι, και προχωράμε στην ανάλυση, βελτίωση και ελαφρά επέκταση αυτού του έργου. Συνεχίζουμε χρησιμοποιώντας προεκπαιδευμένες αρχιτεκτονικές transformers, δηλαδή το Robustly Optimized BERT Pretraining Approach (RoBERTa) και το Audio Spectrogram Transformer (AST), για να ταξινομήσουμε μουσικές δημιουργίες σε 9 ξεχωριστές κατηγορίες συναισθημάτων και ειδών, χρησιμοποιώντας τους στίχους, τον ήχο και έναν συνδυασμό των δύο. Τέλος, αναζητούμε μεθόδους για να εξηγήσουμε κάθε μοντέλο και προτείνουμε έναν τρόπο για τη δημιουργία πολυτροπικών επεξηγήσεων από στίχους και ήχο, χρησιμοποιώντας τη δύναμη του LIME [51] και την ηχητική του εφαρμογή audioLIME [25]. Τέλος δημιουργούμε συνολικούς συνδιασμούς [35] των εξηγήσεων LIME, παρέχοντας πληροφορίες για την απόδοση των μοντέλων και την ικανότητά τους να ανιχνεύουν μοτίβα και στοιχεία που είναι διακριτά για κάθε κατηγορία.

Λέξεις-κλειδιά — Ανάκτηση Μουσικής Πληροφορίας, Βαθιά Μάθηση, Πολυτροπικότητα, Ταξινόμηση Μουσικών Ειδών, Ταξινόμηση Συναισθημάτων στη Μουσική, Τοπικές Επεξηγήσεις, Πολυτροπική Επεξηγησιμότητα

Abstract

Music Information Retrieval (MIR) is a field of research concerned with the extraction and analysis of information from music. Among other tasks, it includes music regression/classification and specifically mood detection and genre recognition. Alongside the growth seen in artificial intelligence (AI) fields, MIR has also experienced significant advancements, including the availability of extensive datasets, the integration of new technologies and multimodal approaches as well as the development and application of advanced explainability methods.

In this thesis, we dive into explaining music emotion and genre classification multimodal models. Firstly we look for available datasets that provide multimodal and multi task capabilities. We choose Music4All [54], offering lyrics and audio as well as emotion and genre metadata for each song and proceed by analysing, refining and slightly augmenting this work. We continue by utilizing pretrained transformer architectures, namely Robustly Optimized BERT Pretraining Approach (RoBERTa) and Audio Spectrogram Transformer (AST), so as to classify music creations into 9 distinct emotion and genre categories utilizing their lyrics, their audio and a combination of the two. Finally, we look for methods to explain each model and propose a way to generate multimodal explanations from lyrics and audio, using the power of LIME [51] and its audio implementation auioLIME [25]. Finally we generate global aggregates [35] of LIME explanations, providing insights into the models performance and the models ability to detect themes and elements distinct for each class.

Keywords — Music Information Retrieval, Deep Learning, Multimodality, Music Genre Classification, Music Emotion Classification, Local Explanations, Multimodal Explainability

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να αναφερθώ σε όλα εκείνα τα άτομα, χωρίς την συμβολή των οποίων αυτο το έργο δεν θα ήταν δυνατό. Συγκεκριμένα ευχαριστώ τους Λυμπεράτο Βασίλη και Μενή-Μαστρομηχαλάκη Ορφέα, υποψήφιους διδάκτορες στην ΣΗΜΜΥ του ΕΜΠ, για την στενή τους συνεργασία, την πολύτιμη καθοδήγησή τους και όλα όσα έκαναν για να είναι αυτοί οι μήνες μια ευχάριστη και δημιουργική περίοδος. Ευχαριστώ επίσης και τα υπόλοιπα άτομα στο εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης για τις συζητήσεις οι οποίες επηρέασαν και βοήθησαν στην δημιουργία αυτού του έργου.

Τέλος ευχαριστώ την οικογένεια και τους φίλους μου για την συναισθηματική συνεισφορά και την υποστήριξη τους, όχι μόνο τους τελευταίους μήνες αλλα καθόλη την διάρκεια της ακαδημαικής πορείας. Συγκεκριμένα ευχαριστώ τους Άντρεα, Αντρέα, Αριάδνη, Γιάννο και Κωνσταντίνο που γέμισαν αυτά τα χρόνια με ευχάριστες αναμνήσεις και γέλιο.

Σωτήρου Θεόδωρος, Ιούλιος 2024

Contents

List of Figures

Chapter 1

Εκτεταμένη Περίληψη στα Ελληνικά

1.1 Υπόβαθρο

1.1.1 Έννοιες και δουλειά αναφορικα με την μηχανική μάθηση, την μουσική και την επεξηγησημότητα

Μηχανική και Βαθιά Μάθηση Η Μηχανηκή Μάθηση (ΜΜ) είναι ένας κλάδος της τεχνητής νοημοσύνης που εστιάζει στην ανάπτυξη αλγορίθμων και στατιστικών μοντέλων που επιτρέπουν στους υπολογιστές να εκτελούν εργασίες χωρίς να ακολουθούν σαφείς οδηγίες. Γενικά, τα μοντέλα ΜΜ χρησιμοποιούνται για να κάνουν προβλέψεις δοσμένων κάποιοων δεδομένων εισόδου. Μια συνάρτηση σφάλματος χρησιμοποιείται για να αξιολογήσει την κάθε πρόβλεψη. Το μοντέλο βελτιστοποιείται προσαρμόζοντας κάποια βάρη του ώστε να ελαττώσει την ασυμφωνία μεταξύ κάποιων τιμών του συνόλου εκπαίδευσης και των προβλέψεών του. Αυτή η διαδικασία επαναλαμβάνεται έως να πληρούνται κάποιες προυποθέσεις (π.χ η ακρίβεια να φτάσει μια τιμή). Τέλος η ΜΜ μπορεί να χωριστεί σε εποπτευόμενη (supervised), μη εποπτευόμενη (unsupervised) και ημιεποπτευόμενη (semi-supervised). Η εποπτευόμενη ΜΜ χρησιμοποιεί επισημασμένα σύνολα δεδομένων για να εκπειδεύσει τα αντίστοιχα μοντέλα, ενώ η μη εποπτευόμενη εκπαιδεύεται ώστε να ανιχνεύει μοτίβα και ομαδοποιήσεις χωρίς την ανάγκη για ανθρώπινη εποπτεία. Η ημιεποπτευόμενη συνδιάζει τα άλλα δύο είδη ΜΜ.

Η Βαθιά Μάθηση (ΒΜ) αποτελέι υποσύνολο την μηχανικής μάθησης. Χρησιμοποιεί νευρωνικά δίκτυα με πολλά στρώματα (εξού και το "βαθιά") για να προσεγγίσει τον τρόπο που λειτουργεί ο ανθρώπινος εγκέφαλος και να αναγνωρίζει, να κατηγοριοποιεί και να περιγράφει με ακρίβεια αντικείμενα των δεδομένων εισόδου. Αυτή η μεθοδολογία έχει εφαρμοστεί με επιτυχία σε τομείς όπως η όραση υπολογιστών, η επεξεργασία φυσικής γλώσσας και η αναγνώριση ομιλίας.

Transformers

Οι transformers είναι ένας τύπος αρχιτεκτονικής νευρωνικών δικτύων που επαναστάτησε την επεξεργασία φυσικής γλώσσας (NLP). Σε αντίθεση με τα παραδοσιακά μοντέλα, δεν βασίζονται σε μια διαδοχική προσέγγιση, αλλά μπορούν να αναλύσουν όλα τα μέρη μιας πρότασης ταυτόχρονα. Στην καρδιά ενός transformer βρίσκεται η αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή. Κάθε στρώμα μέσα σε αυτά τα τμήματα εκτελεί συγκεκριμένα καθήκοντα. Το υπόστρωμα self-attention του κωδικοποιητή αναλύει τις σχέσεις μεταξύ των λέξεων στην εισερχόμενη ακολουθία, ενώ το δίκτυο feed-forward συλλαμβάνει πιο περίπλοκα μοτίβα. Οι υπολειμματικές συνδέσεις και η κανονικοποίηση στρώματος βοηθούν στην αποτελεσματική εκπαίδευση. Το masked self-attention του αποκωδικοποιητή αποτρέπει τη διαρροή πληροφοριών κατά την εκπαίδευση, και η προσοχή κωδικοποιητή-αποκωδικοποιητή του επιτρέπει να λαμβάνει υπόψη την κωδικοποιημένη εισερχόμενη πληροφορία κατά τη δημιουργία της εξόδου. Αυτή η στρωματοποιημένη αλληλεπίδραση μεταξύ της κατανόησης της εισόδου και της χρήσης αυτού του πλαισίου για την παραγωγή εξόδου είναι αυτό που κάνει τους transformers τόσο ισχυρούς για εργασίες επεξεργασίας φυσικής γλώσσας.

Η εργασία "Attention Is All You Need" των Ashish Vaswani και συνεργατ(ρι)ών [17] εισάγει το Transformer μοντέλο. Αυτή η επιλογή σχεδίασης επιτρέπει αυξημένη παραλληλοποίηση κατά την εκπαίδευση και μειώνει τον χρόνο που απαιτείται για την εκπαίδευση των μοντέλων. Το μοντέλο Transformer επιδεικνύει ανώτερη απόδοση σε εργασίες μηχανικής μετάφρασης, επιτυγχάνοντας αποτελέσματα αιχμής τόσο στη μετάφραση από Αγγλικά σε Γερμανικά όσο και από Αγγλικά σε Γαλλικά, με σημαντικά χαμηλότερα κόστη εκπαίδευσης σε σύγκριση με τα υπάρχοντα μοντέλα. Ακολουθώντας αυτή την εργασία, το RoBERTa[38] (A Robustly Optimized BERT Approach) βασίζεται στις γλωσσικές αναπαραστάσεις του BERT, εφαρμόζοντας σημαντικές μεθοδολογικές αλλαγές στη διαδικασία προεκπαίδευσης, οι οποίες βελτιώνουν σημαντικά την απόδοση σε διάφορα benchmarks. Οι τροποποιήσεις περιλαμβάνουν την εκπαίδευση του μοντέλου για μεγαλύτερο χρονικό διάστημα, με περισσότερα δεδομένα, σε μεγαλύτερες παρτίδες και χωρίς τον στόχο της πρόβλεψης της επόμενης πρότασης, με αποτέλεσμα τη βελτιωμένη απόδοση και αποδοτικότητα του μοντέλου.

Η επιτυχία των transformers έχει οδηγήσει ακόμη και στην ανάπτυξη των Vision Transformers (ViT), φέρνοντας παρόμοια δύναμη στην ανάλυση εικόνας, αντιμετωπίζοντας τους παρόμοια με τις ακολουθίες λέξεων. Αυτή η ιδέα παρουσιάστηκε στην εργασία "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" [18]. Οι ViTs λειτουργούν τμηματοποιώντας τις εικόνες σε σταθερού μεγέθους τμήματα, επεξεργαζόμενοι αυτά τα τμήματα ως tokens αναλογικά με τα κειμενικά δεδομένα στο NLP. Αυτή η μέθοδος επιτρέπει στον transformer να εφαρμόζει τον ισχυρό μηχανισμό αυτοπροσοχής απευθείας στα τμήματα, συλλαμβάνοντας περίπλοκες χωρικές ιεραρχίες και εξαρτήσεις μεταξύ διαφορετικών μερών μιας εικόνας. Με την εκπαίδευση σε μεγάλα σύνολα δεδομένων και την αξιοποίηση της μεταφοράς μάθησης, οι ViTs έχουν επιδείξει ανταγωνιστική

απόδοση με τα πιο προηγμένα δίκτυα συνελικτικών νευρώνων (CNN), σηματοδοτώντας μια σημαντική αλλαγή στον τρόπο με τον οποίο τα μοντέλα μηχανικής μάθησης αντιλαμβάνονται και κατανοούν τα οπτικά δεδομένα. Βασισμένο στην τεχνολογία των ViTs είναι το Audio Spectrogram Transformer[22]. Πρωτοπορεί στη χρήση ενός καθαρά βασισμένου σε προσοχή μοντέλου, χωρίς συνελικτικές στρώσεις, που εφαρμόζεται απευθείας σε ηχητικά φασματογράμματα για εργασίες ταξινόμησης. Αυτή η προσέγγιση επιτρέπει στο AST να καταγράψει σύνθετα μοτίβα στα ηχητικά δεδομένα, επιτυγχάνοντας αποτελέσματα αιχμής σε διάφορα πρότυπα ταξινόμησης ήχου. Το Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection (HTS-AT) [13] αντιμετωπίζει τις προκλήσεις κλιμάκωσης των ηχητικών transformers εφαρμόζοντας μια ιεραρχική δομή που μειώνει σημαντικά το μέγεθος του μοντέλου και τη διάρκεια εκπαίδευσης. Ωστόσο, το Causal Audio Transformer[37] (CAT) βασίζεται σε αυτήν την επιτυχία, εισάγοντας μια εξειδικευμένη προσέγγιση για την ταξινόμηση ήχου με αξιοποίηση της εξαγωγής χαρακτηριστικών Multi-Resolution Multi-Feature (MRMF) και ενός ακουστικού μπλοκ προσοχής, σχεδιασμένο για να βελτιστοποιεί την επεξεργασία ηχητικών σημάτων. Το CAT ενσωματώνει μια αιτιακή μονάδα που στοχεύει στη βελτίωση της γενικευσιμότητας του μοντέλου, της ερμηνευσιμότητας και στη μείωση της υπερεκπαίδευσης μέσω της χρήσης αντεπιδραστικής συλλογιστικής.

Τέλος, πρέπει να αναφέρουμε το Contrastive Language-Audio Pretraining model (CLAP) [19], ένα πολυτροπικό και θεμελιώδες μοντέλο για πληροφορίες μουσικής. Το CLAP μαθαίνει ηχητικές έννοιες από την εποπτεία φυσικής γλώσσας, χρησιμοποιώντας δύο κωδικοποιητές και την αντιθετική μάθηση για να συνδέσει τη γλώσσα και τον ήχο, δημιουργώντας έναν κοινό πολυτροπικό χώρο. Αυτή η προσέγγιση επιτρέπει τις προβλέψεις Zero-Shot, που σημαίνει ότι μπορεί να προβλέψει χωρίς να έχει εκπαιδευτεί ρητά σε συγκεκριμένες ετικέτες κατηγορίας, και γενικεύει σε πολλαπλούς τομείς και εργασίες. Το μοντέλο εκπαιδεύεται με 128k ζεύγη ήχου-κειμένου και δοκιμάζεται σε 16 υποχρεώσεις, επιδεικνύοντας σημαντικές βελτιώσεις στην ακρίβεια ταξινόμησης και ευελιξία στην πρόβλεψη κατηγορίας κατά την περίοδο επιβεβαίωσης, ειδικά σε ρυθμίσεις Zero-Shot.

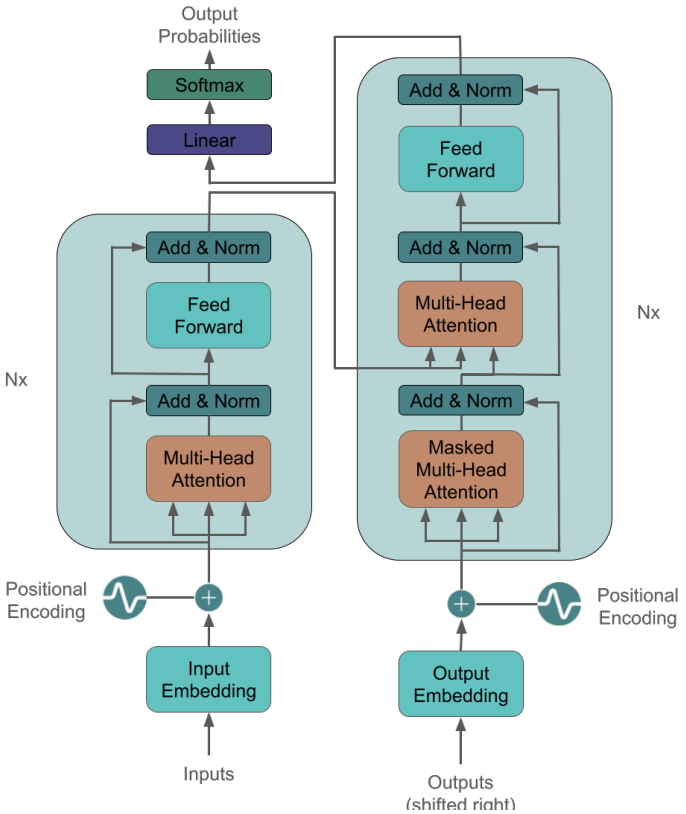


Figure 1.1.1: Η αρχιτεκτονική ενός κωδικοποιητή-αποκωδικοποιητή όπως εμφανίζεται στην δουλειά "Attention Is All You Need" [17].

Πολυτροπικότητα Η τροπικότητα αναφέρεται στον τρόπο με τον οποίο συμβαίνει ή βιώνεται κάτι και συνήθως συνδέεται με τις ανθρώπινες αισθήσεις [7]. Η πολυτροπικότητα στο πλαίσιο της μηχανικής μάθησης και της

ανάλυσης δεδομένων αναφέρεται στην ενοποίηση και επεξεργασία πληροφοριών από πολλαπλούς τύπους δεδομένων ή πηγών όπως κείμενο, εικόνες, ήχος σε σύνολα δεδομένων ή αλγόριθμους. Όπως αναφέρεται στην εργασία [47], οι τύποι πολυτροπικότητας περιλαμβάνουν διάφορες βασικές παραλλαγές, όπως πολυτροπική είσοδο, όπου πολλαπλές μορφές δεδομένων χρησιμοποιούνται για την τροφοδοσία μοντέλων, πολυτροπική έξοδο, όπου τα μοντέλα παράγουν πληροφορίες σε διάφορα μορφότυπα, μετάφραση από μια τροπικότητα σε άλλη, που παραδειγματίζει τον μετασχηματισμό δεδομένων από μια μορφή σε μια κατανοητή αντίστοιχη σε άλλη και τον συνδυασμό διαφορετικών τροπικοτήτων σε μια ενιαία αναπαράσταση, που απεικονίζει την ενσωμάτωση διαφόρων ροών δεδομένων.

Επεξηγησιμότητα

Η επεξηγησιμότητα στην τεχνητή νοημοσύνη περιλαμβάνει μια ποικιλία μεθόδων και τεχνικών, καθεμία εκ των οποίων εξυπηρετεί διαφορετικές πτυχές της διαφάνειας και της κατανόησης. Μερικές κατηγορίες είναι οι εξής [52].

- **Στόχος της επεξήγησης**: Ο στόχος της επεξήγησης μπορεί να κατηγοριοποιηθεί σε σχετιζόμενες με την ενδοσκόπηση και τη δικαιολόγηση. Η ενδοσκόπηση περιλαμβάνει την ανάλυση των εσωτερικών διαδικασιών και χαρακτηριστικών του μοντέλου για την απόκτηση μιας βαθύτερης κατανόησης της λειτουργίας του και των πιθανών προκαταλήψεών του. Η δικαιολόγηση, ωστόσο, εστιάζει στη διασαφήνιση των εξόδων του μοντέλου με τρόπο που να δικαιολογεί τις αποφάσεις του στους τελικούς χρήστες.

- **Αποκλειστικότητα των επεξηγήσεων**: Οι επεξηγήσεις μπορούν να εξεταστούν υπό το πρίσμα των τοπικών έναντι των συνολικών επεξηγήσεων. Η τοπική επεξηγησιμότητα εστιάζει σε συγκεκριμένες περιπτώσεις, προσφέροντας λεπτομερείς πληροφορίες για τη διαδικασία λήψης αποφάσεων για μεμονωμένες προβλέψεις, ενώ η συνολική επεξηγησιμότητα αποσκοπεί στην αποκάλυψη της συνολικής λογικής και συμπεριφοράς του μοντέλου, παρέχοντας μια ευρύτερη κατανόηση του τρόπου λειτουργίας του μοντέλου σε όλες τις περιπτώσεις.

- **Εξάρτηση από το μοντέλο**: Οι μέθοδοι επεξηγησιμότητας μπορούν να διακριθούν σε εξειδικευμένες για το μοντέλο και ανεξάρτητες από το μοντέλο. Οι εξειδικευμένες για το μοντέλο μέθοδοι είναι προσαρμοσμένες στις ιδιαιτερότητες ενός συγκεκριμένου τύπου μοντέλου, αξιοποιώντας τους εσωτερικούς μηχανισμούς του για να διευκρινίσουν πώς λαμβάνονται οι αποφάσεις. Αντίθετα, οι ανεξάρτητες από το μοντέλο προσεγγίσεις σχεδιάζονται για να είναι καθολικά εφαρμόσιμες, παρέχοντας επεξηγήσεις ανεξαρτήτως πρόσβασης στην εσωτερική αρχιτεκτονική του μοντέλου.

- **Χρόνος εφαρμογής της επεξηγησιμότητας**: Τέλος, οι μέθοδοι επεξηγησιμότητας κατηγοριοποιούνται με βάση το πότε εφαρμόζονται, διακρίνοντας μεταξύ μεταγενέστερης (post-hoc) και προγενέστερης (ante-hoc) προσέγγισης. Η μεταγενέστερη επεξηγησιμότητα εφαρμόζεται μετά την εκπαίδευση του μοντέλου, κρίσιμη για πολύπλοκα μοντέλα όπου η ενδογενής ερμηνευσιμότητα είναι δύσκολη. Αντίθετα, η προγενέστερη (ή ενδογενής) επεξηγησιμότητα περιλαμβάνει την ενσωμάτωση της επεξηγησιμότητας σε ένα μοντέλο από την αρχή (π.χ. δέντρα αποφάσεων).

Οι τεχνικές επεξηγησιμότητας ανεξάρτητες από το μοντέλο προσφέρουν μια ευέλικτη προσέγγιση για την κατανόηση των προβλέψεων από διάφορα πολύπλοκα μοντέλα. Συγκεκριμένα, οι Τοπικές Επεξηγήσεις Ανεξαρτήτως Μοντέλου [51] (LIME), εισάγουν μια νέα τεχνική επεξήγησης που στοχεύει στο να κάνει τις προβλέψεις των μοντέλων μηχανικής μάθησης κατανοητές στους ανθρώπους. Το LIME σχεδιάστηκε για να εξηγεί τις προβλέψεις οποιουδήποτε ταξινομητή με έναν ερμηνεύσιμο τρόπο, προσεγγίζοντας το μοντέλο τοπικά με ένα ερμηνεύσιμο μοντέλο. Αυτή η μέθοδος αντιμετωπίζει την πρόκληση των μοντέλων μηχανικής μάθησης να λειτουργούν ως μαύρα κουτιά, όπου οι λόγοι πίσω από τις προβλέψεις τους δεν είναι σαφείς. Βασισμένο στο πλαίσιο του LIME, τα audioLIME [25] και CoughLIME [64] παρέχουν ερμηνεύσιμες και ακροάσιμες εξηγήσεις για τις προβλέψεις ήχου. Συγκεκριμένα, το audioLIME προσφέρει ερμηνεύσιμες, ακροάσιμες εξηγήσεις για συστήματα MIR χρησιμοποιώντας διαχωρισμό πηγών για τη δημιουργία διαταραχών, αντιμετωπίζοντας μια μοναδική πτυχή των ηχητικών δεδομένων που παραβλέπεται από τις παραδοσιακές μεθόδους που βασίζονται σε φασματογράμματα. Από την άλλη πλευρά, το CoughLIME έχει σχεδιαστεί για να παρέχει ηχοποιημένες εξηγήσεις για τις προβλέψεις που γίνονται από ταξινομητές βήχα COVID-19. Προσαρμόζει επίσης το LIME για ηχητικά δεδομένα, εστιάζοντας συγκεκριμένα στους ήχους βήχα που σχετίζονται με τον COVID-19, διαχωρίζοντας τον ήχο σε ερμηνεύσιμα συστατικά. Για μια βαθύτερη κατανόηση της επεξηγήσιμης τεχνητής νοημοσύνης για ηχητικές εργασίες, ανατρέξτε στην ανασκόπηση [4].

Οι ερευνητές έχουν σημειώσει σημαντική πρόοδο στην επεξηγησιμότητα για την επεξεργασία μουσικής πληροφορίας (MIR), με διάφορες τεχνικές να προσφέρουν πληροφορίες για το πώς τα μοντέλα φτάνουν στις αποφάσεις τους. Οι Lyberatos κ.ά. [39] παρουσιάζουν μια ροή εργασίας για την αυτόματη επισήμανση μουσικής, δίνοντας έμφαση στην ερμηνευσιμότητα μέσω αντιληπτικών χαρακτηριστικών, ενσωματώνοντας επεξεργασία σήματος, βαθιά μάθηση και συμβολική γνώση. Η προσέγγιση δείχνει ανταγωνιστική απόδοση με μεθόδους αιχμής στα σύνολα δεδομένων MTG-Jamendo και GTZAN, υπογραμμίζοντας την αξία της ερμηνευσιμότητας παρά τις ενδεχόμενες θυσίες στην απόδοση. Οι Dervakos κ.ά. [16] παρουσιάζουν μια προηγμένη προσέγγιση για την αναγνώριση μουσικών ειδών χρησιμοποιώντας CNNs. Ένα αξιοσημείωτο σημείο της μελέτης είναι η εξερεύνηση τεχνικών εξηγήσιμης τεχνητής νοημοσύνης (XAI) που εφαρμόζονται στην ταξινόμηση μουσικών ειδών. Οι ερευνητές προσαρμόζουν διάφορες μεθόδους επεξηγησιμότητας μετά την εκπαίδευση (post hoc), συμπεριλαμβανομένων των Grad-CAM, LIME και μιας τροποποιημένης Γενετικής Προγραμματισμού για Επεξηγησιμότητα (GPX), για να παρέχουν πληροφορίες για τη διαδικασία λήψης αποφάσεων των CNNs. Οι Chowdhury κ.ά. .[14] εισάγουν ένα μοντέλο βαθιάς μάθησης που προβλέπει τις συναισθηματικές πτυχές της μουσικής βάσει αντιληπτικών χαρακτηριστικών μεσαίου επιπέδου, στοχεύοντας στην επεξηγησιμότητα σε συστήματα MIR. Η έρευνα χρησιμοποιεί ένα δίκτυο τύπου VGG, δείχνοντας ελάχιστη απώλεια απόδοσης κατά την ενσωμάτωση αυτών των αντιληπτικών χαρακτηριστικών, τα οποία λειτουργούν και ερμηνεύσιμα με βάση την μουσική. Το μοντέλο διευκολύνει την κατανόηση των συναισθηματικών προβλέψεων, δικαιολογώντας τη μικρή μείωση της ακρίβειας προς όφελος της επεξηγησιμότητας. Οι Zhang κ.ά. [65] εισάγει το BART-fusion, ένα νέο μοντέλο που δημιουργεί ερμηνείες των στίχων τραγουδιών ενσωματώνοντας ένα προεκπαιδευμένο γλωσσικό μοντέλο μεγάλης κλίμακας με έναν κωδικοποιητή ήχου μέσω ενός μηχανισμού προσοχής διαφόρων μορφών. Αυτή η προσέγγιση επιτρέπει στο μοντέλο να κατανοεί τα τραγούδια τόσο από την πλευρά των στίχων όσο και του ήχου, οδηγώντας σε ακριβείς ερμηνείες. Τα πειραματικά αποτελέσματα δείχνουν ότι η ενσωμάτωση πληροφοριών ήχου βελτιώνει την ικανότητα του μοντέλου να κατανοεί και να δημιουργεί ερμηνείες. Οι Won κ.ά. [63] προτείνουν ένα μοντέλο επισήμανσης μουσικής που χρησιμοποιεί αυτοπροσοχή και CNNs για να ενισχύσει την ερμηνευσιμότητα διατηρώντας παράλληλα ανταγωνιστική απόδοση. Η αρχιτεκτονική τους, σχεδιασμένη να καταγράφει τόσο τα τοπικά χαρακτηριστικά όσο και τις μακροχρόνιες σχέσεις μέσα στα μουσικά κομμάτια, περιλαμβάνει ρηχές συνελικτικές στρώσεις ακολουθούμενες από συσσωρευμένους κωδικοποιητές Transformers. Το μοντέλο υπερέχει σε ερμηνευσιμότητα σε σχέση με τις παραδοσιακές προσεγγίσεις πλήρως συνελικτικών και επαναληπτικών νευρωνικών δικτύων χωρίς να θυσιάζει την ακρίβεια.

Οι Rodis et al.[52] παρέχουν μια ολοκληρωμένη ανασκόπηση της Πολυτροπικής Επεξηγήσιμης Τεχνητής Νοημοσύνης (MXAI), επισημαίνοντας τις μεθοδολογικές προόδους και τις μελλοντικές ερευνητικές κατευθύνσεις του πεδίου. Αυτή η ανασκόπηση αναλύει συστηματικά τις κύριες προβλεπτικές εργασίες, σύνολα δεδομένων και μεθόδους της MXAI.

Συναισθήματα και Είδη στην Μουσική Διαχρονικά η μουσική υπήρξε αναμφισβήτητα ένα μέσο έκφρασης και πρόκλησης συναισθημάτων. Από τα μελαγχολικά στελέχη μιας μελωδίας σε ελλάσονα κλίμακα μέχρι τη χαρούμενη ενέργεια ενός γρήγορου ρυθμού, η μουσική μπορεί να μεταφέρει και να πυροδοτήσει ένα ευρύ φάσμα συναισθημάτων. Οι μελέτες υποδεικνύουν τη σημασία της διάκρισης μεταξύ αντιληπτού και επαγόμενου συναισθήματος στις μουσικές δημιουργίες. Αντιληπτό είναι το συναίσθημα που μεταφέρεται από την ίδια τη μουσική ενώ το επαγόμενο συναίσθημα είναι αυτό που προκαλεί η μουσική στους ακροατές. Οι έρευνες έχουν χρησιμοποιήσει διάφορα συναισθηματικά μοντέλα για να περιγράψουν με μεγαλύτερη ακρίβεια τα μουσικά συναισθήματα [24]. Ένα από τα πρώτα τέτοια έργα είναι το συναισθηματικό δαχτυλίδι του Hevner, το οποίο χρησιμοποιεί 66 επίθετα συναισθημάτων, ταξινομώντας τα σε 8 κατηγορίες [27]. Το πρώτο μοντέλο συναισθημάτων που σχεδιάστηκε για συναισθήματα που προκαλούνται από τη μουσική είναι το Geneva Emotional Music Scales (GEMS), που περιλαμβάνει ένα αρχικό σύνολο 45 ετικετών που μπορούν να ομαδοποιηθούν σε εννέα διαφορετικές διαστάσεις [58]. Ωστόσο, τα κατηγορηματικά μοντέλα συναισθημάτων έχουν πρόσφατα παραγκωνιστεί απο μοντέλα συναισθημάτων διαστάσεων. Τα έργα των Russel και Thayer οργανώνουν τις περιγραφές της διάθεσης σε μοντέλα χαμηλών διαστάσεων. Ειδικότερα, που χρησιμοποιείται πιο συχνά στο MER, το κυκλικό μοντέλο του Russell, απεικονίζει τα συναισθήματα σε έναν δισδιάστατο χώρο, όπου ο βαθμός ευφορίας (που κυμαίνεται από ευχάριστο έως δυσάρεστο) και η διέγερση (που κυμαίνεται από ήρεμο έως διεγερμένο) λειτουργούν ως άξονες. Το μοντέλο απεικονίζεται στο σχήμα 1.1.2 προσαρμοσμένο από αυτό το άρθρο [55]. Άλλα έργα περιλαμβάνουν μια τρίτη διάσταση ή εφαρμόζουν κατάταξη, κατανομές πιθανοτήτων και ζεύγη αντωνύμων για την έκφραση μουσικών συναισθημάτων [24].

Το μουσικό είδος κατηγοριοποιεί κομμάτια μουσικής με βάση κάποια κοινά χαρακτηριστικά όπως η ενορχήστρωση, ο ρυθμός, ο ρυθμός και το πολιτισμικό πλαίσιο. Για παράδειγμα, η Hip Hop χαρακτηρίζεται από μοτίβα

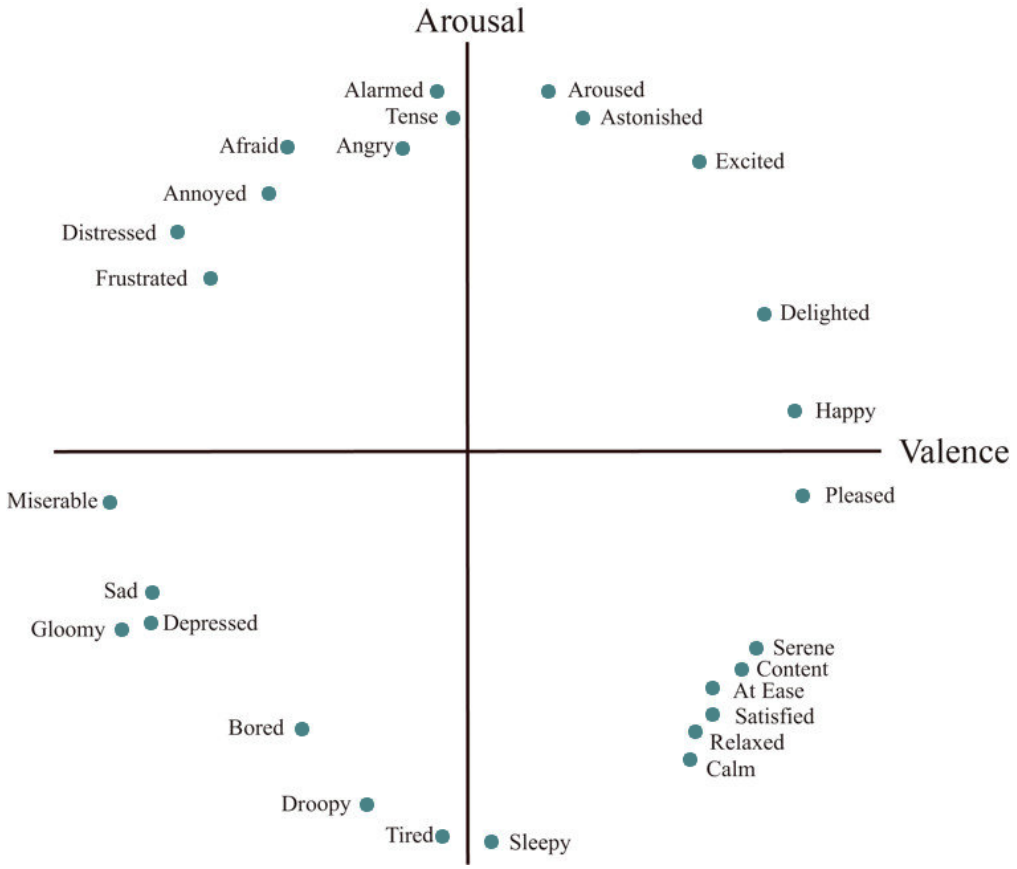


Figure 1.1.2: Το κυκλικό μοντέλου του Russell για τα συναισθήματα (προσαρμοσμένο απο [55]).

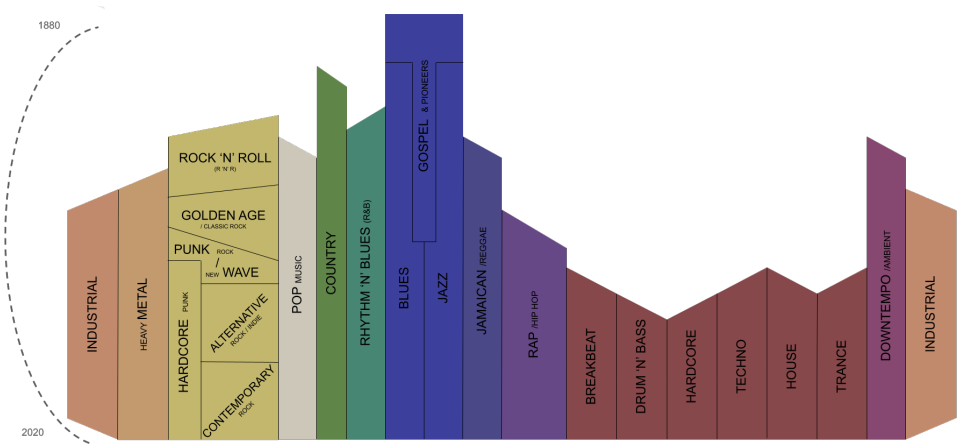


Figure 1.1.3: Η πρόσοψη του διαδραστικού διαγράμματος του Musicmap.info [32]

ραπ και περιέχει θέματα όπως η αστυνομική βία και η καταπίεση, ενώ η ηλεκτρονική μουσική μπορεί να αναγνωριστεί με τη χρήση συνθεσάιζερ. Αν και αυτές οι ομάδες βοηθούν τους ακροατές να πλοηγηθούν στο απέραντο μουσικό τοπίο, μια τέτοια κατηγοριοποίηση συχνά υπόκειται σε διαφορετικές ανθρώπινες εμπειρίες και μερικές φορές είναι επικαλυπτόμενη. Για μια ολοκληρωμένη εξερεύνηση των μουσικών ειδών, των αλληλεπιδράσεων, της ιστορίας και των χαρακτηριστικών τους, το MusicMap.info χρησιμεύει ως πολύτιμος πόρος [32]. Η πρόσοψη του διαδραστικού διαγράμματος που δημιουργήθηκε από τους συντάκτες του πόρου φαίνεται στο 1.1.3. Τα είδη που είναι παρόμοια τοποθετούνται κοντά το ένα με το άλλο ενώ ο κατακόρυφος άξονας μας δίνει μια χρονολογική εκτίμηση της ύπαρξης των ειδών.

1.1.2 Σύνολα Δεδομένων

Το πεδίο της Ανάκτησης Μουσικών Πληροφοριών (MIR) έχει σημειώσει σημαντική ανάπτυξη, με αποτέλεσμα την ύπαρξη μιας ποικιλίας εξειδικευμένων συνόλων δεδομένων που καλύπτουν διάφορες ερευνητικές ανάγκες. Αυτά τα σύνολα δεδομένων, το καθένα μοναδικά κατασκευασμένο, προσφέρουν πληθώρα πόρων. Διαφέρουν πολύ ως προς το περιεχόμενό τους που κυμαίνεται από χαρακτηριστικά ήχου και μεταδεδομένα έως πιο σύνθετους τύπους δεδομένων, όπως σχολιασμούς σε πολιτισμικό πλαίσιο. Για όποιον θέλει να εξερευνήσει το εύρος των διαθέσιμων συνόλων δεδομένων MIR, μια ολοκληρωμένη λίστα με σύντομες περιγραφές του περιεχομένου βρίσκεται στον ιστότοπο της Διεθνούς Κοινότητας για Ανάκτηση Μουσικών Πληροφοριών (ISMIR) website[29].

Στην έρευνά μας, επικεντρωθήκαμε στην επιλογή συνόλων δεδομένων που προσφέρουν μια ποικιλία τροπικοτήτων για μουσικά τραγούδια. Αυτό περιελάμβανε σύνολα δεδομένων που παρείχαν πολλαπλές τροπικότητες εγγενώς, όπως ήχο, στίχους και MIDI, καθώς και εκείνες με μία μόνο τροπικότητα που θα μπορούσε να επαυξηθεί σχετικά εύκολα. Επιπλέον, θα ήταν ιδανικό εάν τα σύνολα δεδομένων παρείχαν μεταδεδομένα που υποστηρίζουν μια σειρά αναλυτικών εργασιών ή διευκολύνουν την εύκολη απόκτηση πρόσθετων μεταδεδομένων. Μια σημαντική πρόκληση που αντιμετωπίσαμε σε αυτήν τη διαδικασία ήταν οι περιορισμοί που δημιουργούσαν η ύπαρξη πνευματικών δικαιωμάτων, οι οποίοι περιόρισαν τη διαθεσιμότητα κατάλληλων συνόλων δεδομένων. Παρά αυτούς τους περιορισμούς, εντοπίσαμε μερικά υποψήφια σύνολα δεδομένων. Ονομαστικά αυτά ήταν το Million Song Dataset (MSD) [9], το GTZAN, το Free Music Archive (FMA) [8], το Database for Emotional Analysis of Music (DEAM) [6], το MTG-Jamendo και τέλος το Music4All [54] το οποίο και επιλέγουμε να χρησιμοποιήσουμε.

1.2 Μεθοδολογία

Οι μέθοδοι που εφαρμόσαμε σε αυτήν την δουλειά συμπεριλαμβάνουν μια αναλυτική μελέτη και κριτική του αρχικού συνόλου δεδομένων, μεθόδους για εκκαθάριση και επαύξησή του, την περιγραφή των μοντέλων που εκπαιδεύσαμε αλλα και τις μεθόδους επεξήγησης που εφαρμόσαμε για να κατανοήσουμε την συμπεριφορά τους. Η υλοποίηση του κώδικα όπως καθώς και διάφορα αποτελέσματα μπορούν να βρεθούν στην σελίδα μας στο

GitHub.

1.2.1 Ανάλυση και επεξεργασία του συνόλου δεδομένων Music4All

Η βάση δεδομένων Music4All [54] (M4A) είναι ένας πόρος που έχει σχεδιαστεί για να υποστηρίζει μια ποικιλία ερευνών στον τομέα του MIR. Περιέχει μια πλούσια συλλογή μεταδεδομένων, ετικετών, πληροφοριών είδους, κλιπ ήχου 30 δευτερολέπτων, στίχους και πολλά άλλα, που συλλέγονται για ένα ευρύ φάσμα μουσικών κομματιών. Η ανάπτυξη της βάσης δεδομένων πραγματοποιήθηκε σε δύο φάσεις: τη φάση του χρήστη και τη φάση του τραγουδιού. Στη φάση του χρήστη, συγκεντρώθηκαν και ανωνυμοποιήθηκαν δεδομένα σχετικά με το ιστορικό ακρόασης των χρηστών, ενώ η φάση του τραγουδιού περιελάμβανε τη συλλογή λεπτομερών δεδομένων τραγουδιού. Το σύνολο δεδομένων έχει εκτεταμένο μέγεθος και περιέχει δεδομένα για περισσότερα από 100.000 τραγούδια. Οι αναγνώστες που ενδιαφέρονται για περισσότερες λεπτομέρειες σχετικά με το σύνολο δεδομένων μπορούν να βρουν περισσότερες πληροφορίες στην εργασία τους [54].

Το σύνολο δεδομένων χαρακτηρίζεται από την ποικιλομορφία αναφορικά με το είδος, προσφέροντας έως και οκτώ ετικέτες είδους ανά κομμάτι και περιλαμβάνοντας περισσότερα από 600 μοναδικά είδη. Σε περιπτώσεις όπου διαφορετικά πεδία ειδών περιείχαν την ίδια τιμή για το ίδιου τραγούδι (π.χ. "genre 1: rock" και "genre 2: rock") διατηρούμε μόνο μία παρουσία του είδους. Η κατανομή των 20 πιο πολυπληθών ετικετών είδους του συνόλου δεδομένων φαίνεται στο Σχήμα 1.2.1. Είναι σημαντικό να σημειωθεί η ανισορροπία του συνόλου δεδομένων, με κυρίαρχη την αναπαράσταση των ειδών Pop και Rock. Αυτή η απόκλιση μπορεί να αποδοθεί στην ευρεία επικράτηση αυτών των ειδών στη μουσική βιομηχανία και στην ευρεία κατηγοριοποίηση τους από μη ειδικούς.

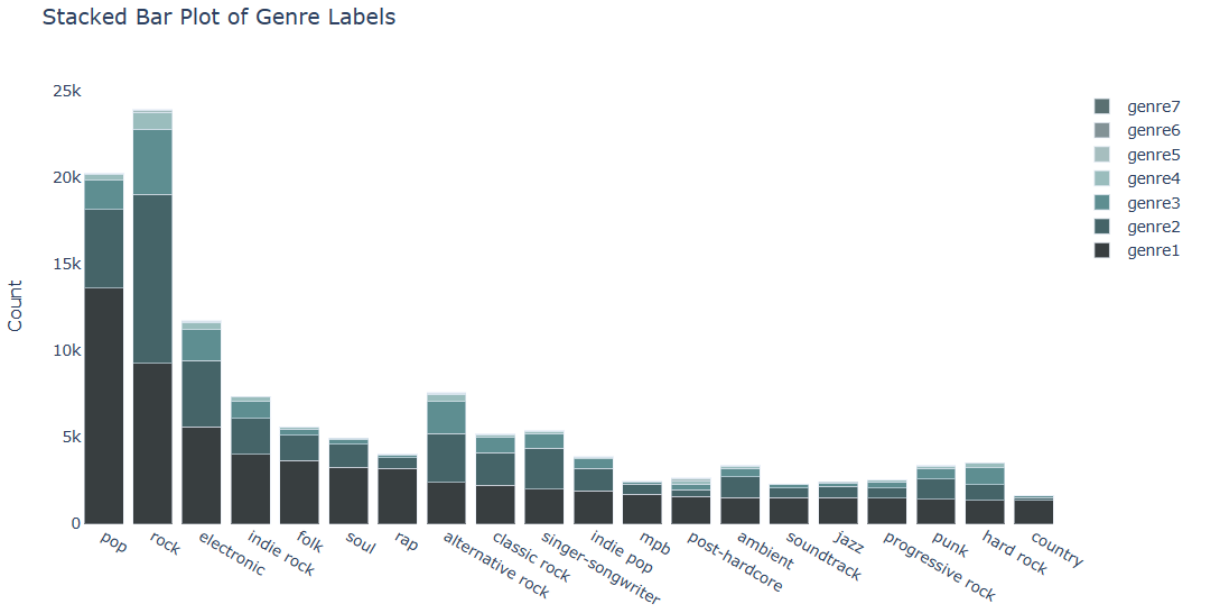
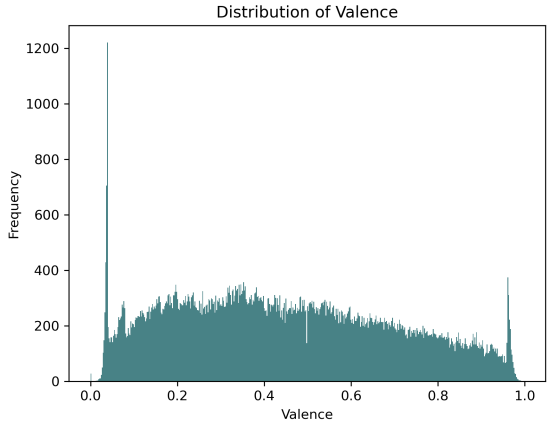


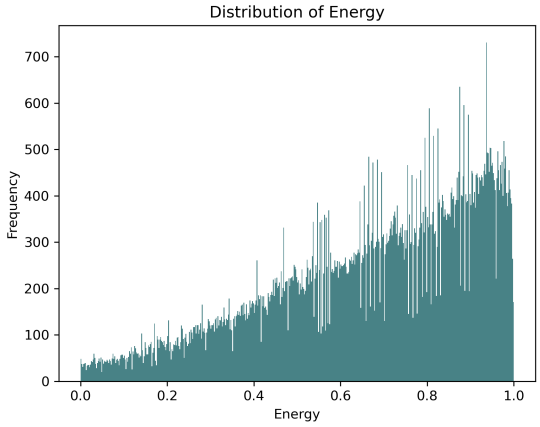
Figure 1.2.1: Εικόνα που περιέχει την κατανομή των ειδών. Κάθε χρώμα αντιπροσωπεύει την κατανομή μιας ετικέτας για κάθε καταχώρηση.

Οι τιμές βαθμού ευφορίας και ενέργειας που περιέχονται στο σύνολο δεδομένων, που προέρχονται από το Spotify, είναι συνεχείς, κυμαίνονται από 0 έως 1, και παρουσιάζουν ξεχωριστά μοτίβα κατανομής. Οι βαθμολογίες βαθμού ευφορίας τείνουν να ακολουθούν μια σχεδόν κανονική κατανομή με μια μικρή κλίση προς λιγότερο χαρούμενα τραγούδια (περιπτώσεις με βαθμού ευφορίας $< 0{,}5$), όπως φαίνεται στο Σχήμα 1.2.2a. Αντίθετα, η κατανομή ενέργειας μοιάζει με το σχήμα μιας γραμμικής συνάρτησης, παρουσιάζεται στο Σχήμα 1.2.2b, και υποδεικνύει λιγότερα κομμάτια χαμηλής ενέργειας και μια πληθώρα αυτών με υψηλή ενέργεια. Ένα εξαγωνικό διάγραμμα Binning απεικονίζεται επίσης στο σχήμα 1.2.2c, που αντιπροσωπεύει την κατανομή των τιμών σθένους και ενέργειας σε έναν δισδιάστατο χώρο. Αυτή η γραφική παράσταση αποκαλύπτει ότι τα τραγούδια υψηλού σθένους αλλά χαμηλής ενέργειας είναι σπάνια, που υποδεικνύονται από το φωτεινό χρώμα αυτών των εξάγωνων, ενώ τα τραγούδια υψηλής ενέργειας αλλά χαμηλού σθένους είναι πιο συχνά και η ένταση χρώματος του αντίστοιχου

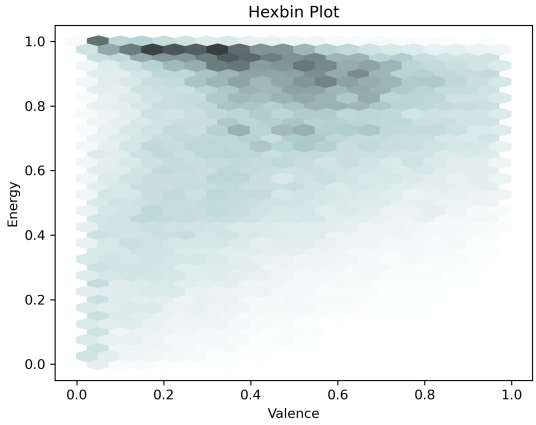
εξαγώνου είναι υψηλότερη, υποδηλώνοντας έλλειψη χαλαρωτικών και ήρεμων μελωδιών. το σύνολο δεδομένων. Αυτή η ανισορροπία μπορεί να εξηγηθεί από τις δημοφιλείς τάσεις της μουσικής, που ευνοούν πιο αισιόδοξα, ενεργητικά κομμάτια που οδηγούν στην υπερεκπροσώπησή τους σε συλλογές και σύνολα δεδομένων.



(a) κατανομή βαθμού ευφορίας



(b) Κατανομή ενέργειας



(c) Εξαγωνικό διάγραμμα Binning

Figure 1.2.2: Κατανομές βαθμού ευφορίας και ενέργειας καθώς και μια γραφική παράσταση εξάγωνων που αντιπροσωπεύει την κατανομή των τιμών σθένους και ενέργειας σε έναν δισδιάστατο χώρο, όπου η χρωματική ένταση κάθε εξαγώνου αντιστοιχεί στη συγκέντρωση τραγουδιών με αυτά τα επίπεδα σθένους και ενέργειας. Όσο πιο σκούρο είναι το χρώμα τόσο μεγαλύτερη είναι η συγκέντρωση τέτοιων τραγουδιών.

Στην προσέγγισή μας για ταξινόμηση των ειδών, αναγνωρίσαμε την ανάγκη να συμπυκνώσουμε το ποικίλο φάσμα των μουσικών ειδών σε πιο γενικές κατηγορίες. Αυτή η απλοποίηση είναι κρίσιμη για την αποτελεσματική ταξινόμηση και ανάλυση. Για να καθοδηγήσουμε την ανακατάταξη μας, χρησιμοποιήσαμε τη χαρτογράφηση είδους που είναι διαθέσιμη στη διεύθυνση Musicmap. Αυτή η πηγή παρείχε μια πληθώρα πληροφοριών καθώς και μια οπτική κατηγοριοποίηση των ειδών που αποδείχθηκαν χρήσιμα για την ενοποίηση διαφόρων μουσικών ειδών σε εννέα ευρείες κατηγορίες. Αυτές οι τάξεις επιλέχθηκαν προσεκτικά για να καλύπτουν το ευρύ φάσμα των μουσικών στυλ, διατηρώντας παράλληλα διακριτές και σημαντικές κατηγορίες για την κατάταξή μας. Οι ετικέτες αυτών των τάξεων είναι:

- Rock, που περιλαμβάνει rock 'n' roll, golden age rock, classic rock και contemporary rock.
- Pop, μια ευρεία κατηγορία που περιλαμβάνει δημοφιλή στυλ μουσικής από την δεκαετία του 1950.
- Hip Hop, καλύπτβντας hip hop όπως και rap songs.
- Alternative Rock, συμπεριλαμβανομένου indie, alternative και άλλα στυλ που διαφέρονυ απο την main-

stream rock.

- Heavy Music, αυτή η ετικέτα χρησιμοποιείται για metal, hardcore, and industrial μουσική με metal στοιχεία.
- Punk, αποτελούμενη από punk rock και new wave τραγούδια.
- Electronic, εμπεριέχοντας electronic dance μουσική, downtempo και industrial μουσική με electronic στοιχεία.
- Rhythm Music, αποτελούμενη απο rhythm 'n' blues, blues, gospel, jazz και Jamaican μουσική.
- Folk, που επίσης περιλαμβάνει country μουσική.
- Other, η οποία περιέχει διφορούμενα και άλλα είδη που δεν ανήκουν σε καμία από τις προηγούμενες κατηγορίες.

Όσον αφορά την ταξινόμηση της μουσικής με βάση το συναίσθημα, μετατρέπουμε τις συνεχείς τιμές βαθμού ευφορίας και ενέργειας σε διακριτές συναισθηματικές ετικέτες. Αυτό επιτυγχάνεται μέσω μιας καθορισμένης χαρτογράφησης που ταξινομεί κάθε τραγούδι σε 9 συναισθηματικές καταστάσεις σύμφωνα με το μοντέλο κύκλου του Russel. Συγκεκριμένα, καταχωρήσεις με υψηλό βαθμό ευφορίας (μεγαλύτερο από 0,65) χαρακτηρίζονται "Exciting" εάν η ενέργειά τους είναι επίσης υψηλή, "Relaxing" αν η ενέργειά τους είναι χαμηλή (κάτω από 0,35) και "Happy" διαφορετικά. Για χαμηλό βαθμό ευφορίας (λιγότερο από 0,35) τα κομμάτια φέρουν την ένδειξη "Angry", "Depressing" και "Sad" εάν έχουν υψηλές, χαμηλές ή ενδιάμεσες τιμές ενέργειας αντίστοιχα. Για τις τιμές βαθμό ευφορίας στο μεσαίο εύρος, τα ίχνη θεωρούνται "Tense", "Calm" ή "Neutral" αντίστοιχα.

Αντιμετωπίσαμε κάποιες προκλήσεις με τις ετικέτες είδους και συναισθήματος. Αρχικά αναφορικά με τις ετικέτες είδους, οι δημιουργοί του M4A συλλέγουν δεδομένα σημασμένα απο χρήστες, και οχι ειδικούς, από το last.fm και στην συνέχεια τα φιλτράρουν. Αν και το last.fm παρέχει βάρη για τις ετικέτες, αυτά παραλείπονται από το M4A. Επίσης εφόσον τις ετικέτες της έβαλαν μη ειδικοί είναι πιθανό να είναι θορυβώδεις. Για να αντιμετωπίσουμε αυτο το πρόβλημα βρίσκουμε επίσης τα είδη κάθε καλλιτέχνη απο το SpotifyAPI και θεωρούμε ότι κατάλληλη είναι η πρώτη ετικέτα που εμφανίζεται στο M4A και υπάρχει και στην λίστα ετικετών του καλλιτέχνη. Όσον αφορά τις ετικέτες συναισθήματος, όπως περιμέναμε, τα δεδομένα παρουσιάζουν έντονη ανισορροπία με πολύ λίγα "Relaxing" τραγούδια. Για αυτόν τον λόγο δημιουργήσαμε scripts για επαύξηση του συνόλου μας, παίρνοντας ήχο και μεταδεδομένα απο το SpotifyAPI και στίχους από την διεπαφή του GeniusLyrics.

Δημιουργούμε το τελικό μας σύνολο δεδομένων που περιέχει τους στίχους μαζί με ένα ηχητικό κλιπ διάρκειας 30 δευτερολέπτων για κάθε καταχώρηση, καθώς και ετικέτες διάθεσης και είδους. Αφαιρούμε τραγούδια με την ετικέτα είδους "other". Το σύνολο δεδομένων μας περιλαμβάνει 9 ετικέτες συναισθημάτων για κάθε τραγούδι και 9 ετικέτες είδους. Δημιουργούμε έναν διαχωρισμό των δεδομένων train-val-test, διασφαλίζοντας ότι καλλιτέχνες που εμφανίζονται στο διαχωρισμό train-val δεν παρουσιάζονται επίσης στο test set. Μια σύνοψη της διαδικασίας που ακολουθήθηκε για τη δημιουργία του συνόλου δεδομένων μας φαίνεται στο σχήμα 1.2.3.

1.2.2 Επιλογή χαρακτηριστικών, Μοντέλα παλινδρόμησης και οι αρχιτεκτονικές των τελικών Μοντέλων κατηγοριοποίησης

Για να κατανοήσουμε καλά την απόδοση διαφορετικών αρχιτεκτονικών, υιοθετήσαμε μια προσέγγιση δύο σταδίων. Στο πρώτο στάδιο, εξερευνούμε διάφορα μοντέλα παλινδρόμησης (regression) που επιφορτίζονται με την πρόβλεψη τιμών βαθμού ευφορίας και ενέργειας, ώστε να πληροφορηθούμε για τα δεδομένα και τις σχέσεις χαρακτηριστικών για κάθε τροπικότητα. Προκειμένου να καθοριστούν βασικές επιδόσεις, εφαρμόζουμε δύο θεμελιώδες μοντέλα τους dummy και mean regressors. Ο πρώτος κάνει τυχαίες προβλέψεις ενώ ο δεύτερος μαντεύει την μέση τιμή του train set. Επίσης μελετάμε μοντέλα με στίχους ως είσοδο όπως είναι τα δυο Long Short-Term Memory (LSTM) μοντέλα που προτείνουν οι δημιουργοί του M4A [54] και η εφαρμογή του XLnet σε στίχους κατά αντιστοιχία με την δουλειά των Agrawa κ.ά. [1]. Συνεχίζοντας μελετάμε πολυτροπικές προσεγγίσεις όπως αυτή των Delbouys κ.ά. [15], που συνδιάζουν τα αποτελέσματα ενός συνελικτικού μοντέλου φασμογραμμάταων με αυτά ενός συνδιαστικού μοντέλου (συνέλιξη και LSTM) με είσοδο Word2Vec embeddings στίχων αλλά και την μελέτη των Krols κ.ά. [31] που ενσωματώνουν υψηλού επιπέδου ηχητικά χαρακτηριστικά μαζί με χαρακτηρηστικά στίχων που προκύπτουν από μια διαδικασία Συχνότητας Όρων - Αντίστροφης Συχνότητας Εγγράφων (TF-IDF) σε ενα Mulitlayer Perceptron (MLP). Δοκιμάσαμε και το MuLan, που αποδίδει καλά σε διάφορες

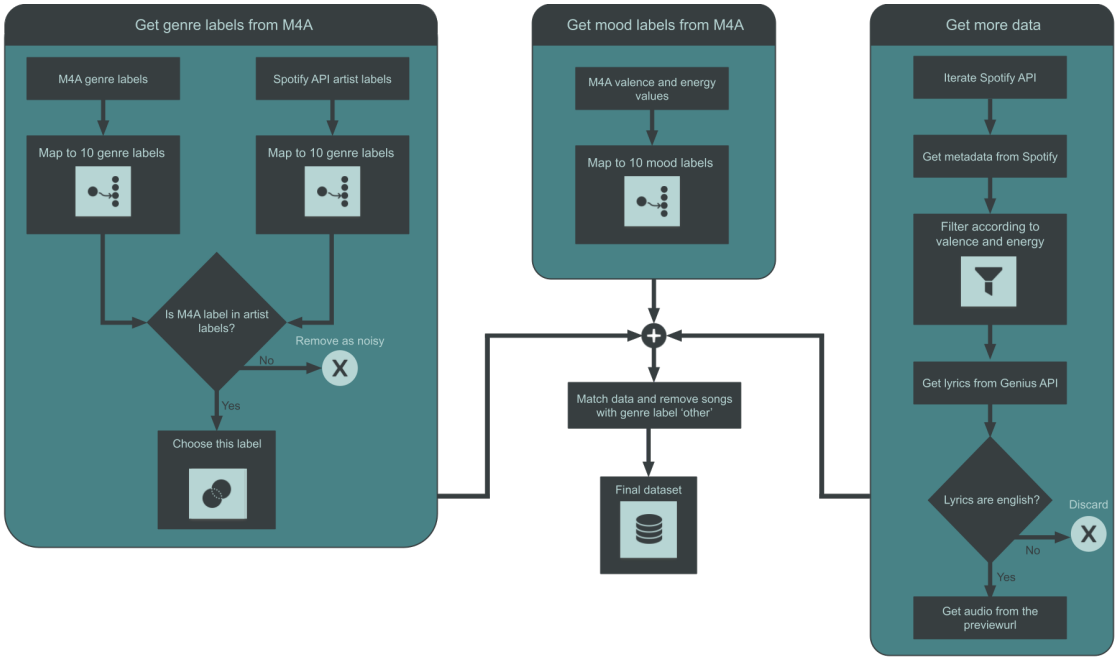


Figure 1.2.3: Μια περίληψη της διαδικασίας που ακολουθήσαμε για την τροποποίηση του M4A.

εργασίες, [28] παρέχοντάς του στίχους αντί για περιγραφές του ήχου. Τέλος συνδιάζομυε ενα Audio Spectrogram Transformer (AST) [22] μοντέλο μετο Robustly optimized BERT approach (RoBERTa) [38]. Αυτή η εξερεύνηση βοήθησε στην επιλογή ενός κατάλληλου μοντέλου για κάθε τροπικότητα, το οποίο στη συνέχεια προσαρμόστηκε για εργασίες ταξινόμησης.

Μετά από αυτή την ενδελεχή έρευνα προχωράμε στο δεύτερο στάδιο, την εκπαίδευση ενός μοντέλου κατηγοριοποίησης για κάθε τροπικότητα και για κάθε εργασία. Συγκεκριμένα, το λυρικό μοντέλο αξιοποιεί τη δύναμη της αρχιτεκτονικής roberta-large για ταξινόμηση κειμένου. Αρχικά προετοιμάζουμε το κείμενο εισαγωγής για tokenization μετατρέποντας όλους τους χαρακτήρες σε πεζούς για να διατηρηθεί η συνέπεια και, στη συνέχεια, αφαιρούμε τα σημεία στίξης για να μειώσουμε την πολυπλοκότητα και τον πιθανό θόρυβο στα δεδομένα. Η συνάρτηση tokenization, διαμορφωμένη με μέγιστο μήκος ακολουθίας 256 διακριτικών, κωδικοποιεί το καθαρό κείμενο σε μια μορφή κατάλληλη για ένα μοντέλο RoBERTa, δημιουργώντας αριθμητικά αναγνωριστικά και μια μάσκα που υποδεικνύει ποια διακριτικά πρέπει να παρακολουθεί το μοντέλο. Προχωράμε στην εκπαίδευση (fine-tuning) του μοντέλου με βάση τις διαφορετικές κατηγορίες ετικετών που υπάρχουν στο σύνολο δεδομένων μας. Ο optimizer επιλογής είναι το AdamW, με ρυθμό εκμάθησης 9e-7. Το μοντέλο εκπαιδεύεται για εννέα εποχές διατηρώντας την κατάσταση του μοντέλου που επιτυγχάνει την υψηλότερη ακρίβεια επικύρωσης. Το μοντέλο ήχου αξιοποιεί τον προεκπαιδευμένο Audio Spectrogram Transformer για ταξινόμηση ήχου που έχει φορτωθεί από το checkpoint "ast-finetuned-audioset-10-10-0.4593" που παρέχεται από το MIT. Χρησιμοποιούμε πρώτα την κλάση ASTFeatureExtractor για να μετατρέψουμε αρχεία ήχου σε φασματογράμματα. Είναι σημαντικό να σημειωθεί ότι μόνο ένα μέρος του αρχικού ήχου χρησιμοποιείται σε αυτή τη διαδικασία. Με τυπικό ρυθμό δειγματοληψίας ήχου 44100 Hz και μέγιστο μήκος 1024 για το φασματογράφημα όπως ορίζεται από αυτήν την κλάση, μόνο περίπου 10,2 δευτερόλεπτα του αρχικού ήχου 30 δευτερολέπτων χρησιμοποιούνται στην πραγματικότητα για τη δημιουργία του φασματογράμματος. Η βελτιστοποίηση των παραμέτρων του μοντέλου ανατίθεται στον optimizer Adam, χρησιμοποιώντας ρυθμό εκμάθησης 6e-6 και εκπαίδευση για συνολικά 5 εποχές. Η αρχιτεκτονική του τελικού μοντέλου μας έχει σχεδιαστεί για να ενσωματώνει και να ταξινομεί πολυτροπικές εισόδους τόσο από πηγές ήχου όσο και από πηγές κειμένου. Το μοντέλο αξιοποιεί τα προεκπαιδευμένα ASTModel, για ήχο, και το RobertaModel, για ανάλυση κειμένου. Οι είσοδοι ήχου και κειμένου υποβάλλονται σε προεπεξεργασία με τον ίδιο τρόπο όπως στις μονότροπες περιπτώσεις. Η συγκεντρωτική έξοδος από το μοντέλο ήχου και το CLS token από το μοντέλο κειμένου επιλέγονται, καθώς παρέχουν μια ολοκληρωμένη περίληψη του περιεχομένου της αντίστοιχης μορφής. Αυτά τα embeddings στη συνέχεια συνδυάζονται σε ένα ενοποιημένο διάνυσμα και επεξεργάζονται από μια κεφαλή ταξινόμησης, που τα κανονικοποιεί και τα προωθεί σε ένα πλήρως συνδεδεμένο

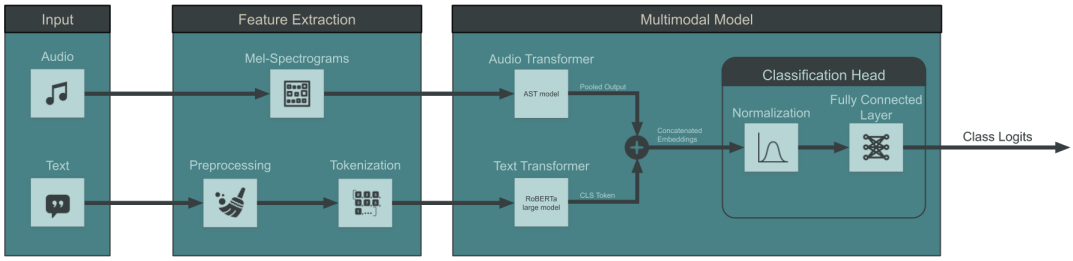


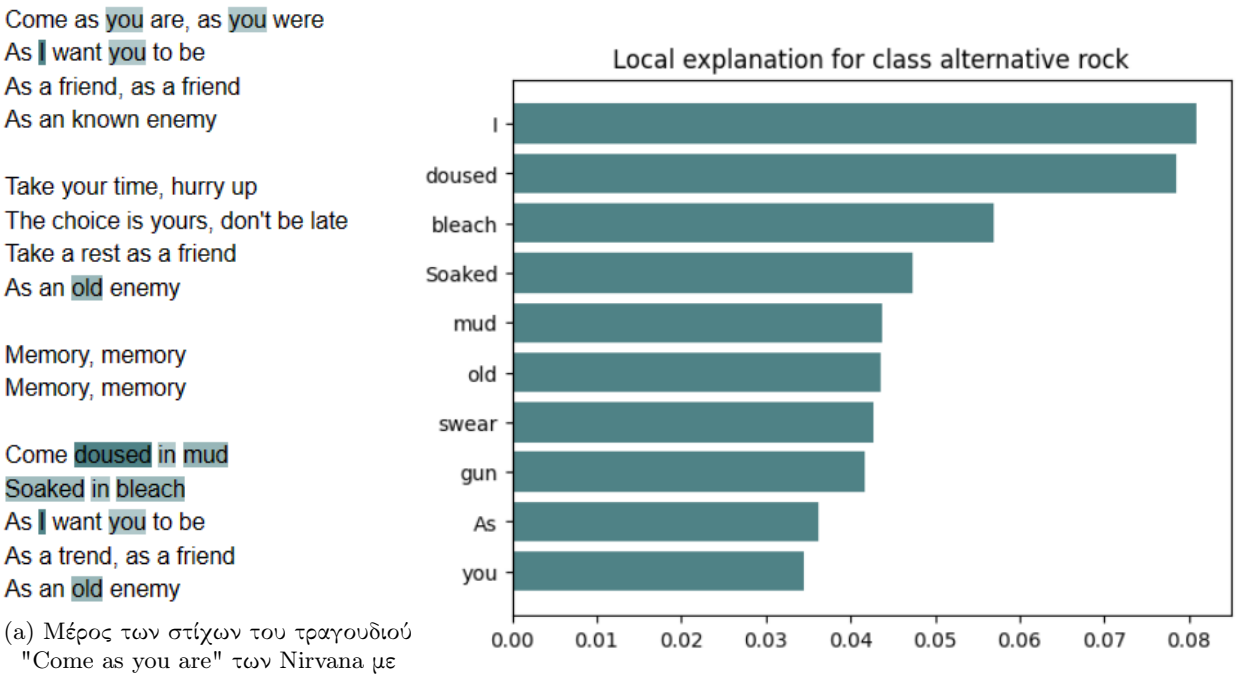
Figure 1.2.4: Η διαδικασία της πολυτροπικής προσέγγισης. Η ακουστική κυματομορφή και οι στίχοι μετατρέπονται πρώτα σε φασματογράμματα mel και tokens αντίστοιχα. Στη συνέχεια, κάθε είσοδος τροπικότητας υποβάλλεται σε επεξεργασία από το αντίστοιχο προεκπαιδευμένο transformer μοντέλο. Η συγγεντρωτική έξοδος των embeddings ήχου που παράγονται απο το ASTmodel και το CLS token απο το RoBERTa συνενώνονται και επεξεργάζονται απο μια κεφαλή ταξινόμησης, παρέχοντας logits για κάθε κατηγορία.

επίπεδο, παράγοντας τα τελικά logits (λογαριθμικές πιθανότητες) ταξινόμησης. Η διαδικασία αυτή μπορεί να φανεί στο 1.2.4.

1.2.3 Μεθοδολογία Επεξηγήσεων των Μοντέλων

Αν και αρκετές τεχνικές όπως το TextFooLer[30] και το MiCE[53] εξετάστηκαν για την επεξήγηση δεδομένων κειμένου, επιλέξαμε να εφαρμόσουμε επεξηγήσεις LIME. Το LIME ξεκινά δημιουργώντας ένα δυαδικό διάνυσμα με ίσο μήκος με το αρχικό σώμα κειμένου, με κάθε ψηφίο να υποδεικνύει την παρουσία ή την απουσία λέξεων. Στη συνέχεια, το Lime διαταράσσει τα δεδομένα εισόδου ενεργοποιώντας ή απενεργοποιώντας τυχαία λέξεις (παρούσες ή απούσες) δημιουργώντας έναν καθορισμένο αριθμό δειγμάτων γύρω από την αρχική είσοδο. Μετά τη δημιουργία αυτών, το LIME χρησιμοποιεί το αρχικό μοντέλο για να προβλέψει τα αποτελέσματα για το καθένα, αντιμετωπίζοντας αυτές τις προβλέψεις ως ετικέτες για την εκπαίδευση ενός απλούστερου, ερμηνεύσιμου μοντέλου, συνήθως μιας γραμμικής παλινδρόμησης. Αυτό το ερμηνεύσιμο μοντέλο έχει σχεδιαστεί για να προσεγγίζει τη διαδικασία λήψης αποφάσεων του πολύπλοκου αρχικού μοντέλου εντός της τοπικότητας του παραδείγματος εισόδου. Δίνοντας έμφαση σε διαταραγμένα δείγματα που είναι πιο κοντά στο αρχικό κείμενο όσον αφορά τον μετασχηματισμένο χώρο χαρακτηριστικών τους, το LIME τους εκχωρεί υψηλότερα βάρη κατά τη διάρκεια αυτής της εκπαίδευσης. Αυτή η διαδικασία διασφαλίζει ότι το μοντέλο επεξήγησης εστιάζει στις πιο σχετικές παραλλαγές των δεδομένων εισόδου, τονίζοντας ποιες λέξεις συμβάλλουν πιο σημαντικά στην πρόβλεψη του μοντέλου. Ένα παράδειγμα φαίνεται στο σχήμα 1.2.5. Το παράδειγμα απεικονίζει ένα μέρος των στίχων "Come as You Are" των Nirvana, με μερικές λέξεις που επηρέασαν το μοντέλο να αποφασίσει την κατηγορία "alternative rock", τονισμένα με μπλε χρώμα. Όσο λιγότερο διαφανές, τόσο μεγαλύτερο είναι το βάρος της λέξης. Εφαρμόζουμε λοιπόν την μέθοδο LIME στα δικά μας δεδομένα για να προσεγγίσουμε τοπικά τον τρόπο λήψης αποφάσεων του RoBERTa.

Στον τομέα της επεξήγησης ήχου, το ταξίδι μας ξεκινά με δυνατότητες που προσφέρει η LIME για την επεξήγηση εικόνων. Σε αυτήν την περίπτωση, αντί να χωρίσει ένα σώμα κειμένου σε λέξεις, το LIME τέμνει τις εικόνες σε superpixel. Στη συνέχεια, δημιουργεί διαταραχές όπου κάθε superpixel ενεργοποιείται ή απενεργοποιείται και σταθμίζεται με βάση την επιρροή του στην έξοδο του μοντέλου, με τα πιο σημαντικά να επισημαίνονται στην αρχική εικόνα. Προσαρμόζοντας αυτή τη μέθοδο στις ανάγκες μας, χρησιμοποιούμε φασματογράφημα, που ουσιαστικά αποτελούν ασπρόμαυρες εικόνες ήχου. Αυτό μας επιτρέπει να εφαρμόσουμε το LIME απευθείας στα ηχητικά μας δεδομένα που αντιπροσωπεύονται σε μορφή φασματογραφήματος. Ένα παράδειγμα φασματογραφήματος καθώς και η εξήγηση όπως δημιουργήθηκε από το LIME φαίνονται στο σχήμα 1.2.6. Για να λάβουμε μια ακροάσιμη εξήγηση από τα φασματογραφήματα, εφαρμόζουμε δύο στρατηγικές: Η πρώτη περιλαμβάνει την προσπάθεια αναδημιουργίας του ήχου από τα επισημασμένα μέρη των φασματογραφημάτων που εξάγονται από το LIME χρησιμοποιώντας έναν αντίστροφο μετασχηματισμό Fourier βραχείας διάρκειας. Η δεύτερη στρατηγική

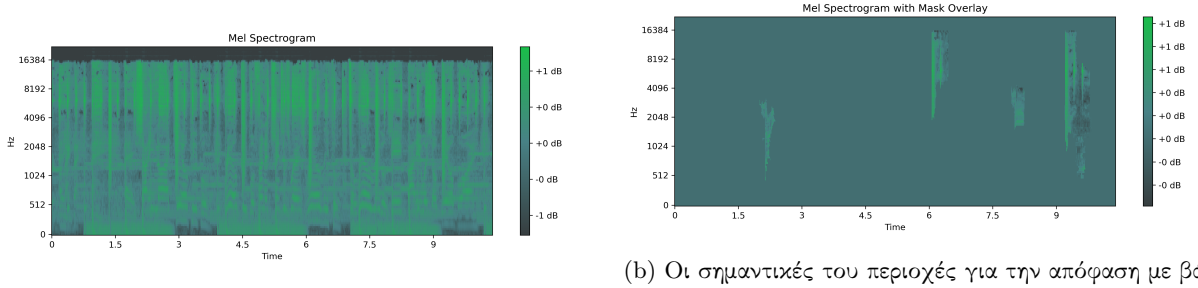


(a) Μέρος των στίχων του τραγουδιού "Come as you are" των Nirvana με σημειωμένες τις σημαντικές λέξεις.

(b) Τα βάρη των σημειωμένων λέξεων με βάση το LIME.

Figure 1.2.5: Ένα παράδειγμα επεξηγήσεων κειμένου LIME με το "Come as You Are" των Nirvana ως είσοδο. Λέξεις που επηρεάζουν το μοντέλο να μαντέψει την κλάση "alternative rock" επισημαίνονται με μπλε χρώμα. Όσο λιγότερο διαφανές, τοσο μεγαλύτερο το βάρος της λέξης όπως φαίνεται στο σχήμα (b).

φιλτράρει τον αρχικό ήχο ώστε το παραγόμενο φασματόγραμμα να μοιάζει με αυτό που δίνει το LIME. Αν και αυτές οι προσεγγίσεις απέδωσαν κάποιες ακροάσιμες εξηγήσεις, τα αποτελέσματά τους, όπως παρουσιάζονται στο επόμενο κεφάλαιο, υποδηλώνουν ότι υπάρχει περιθώριο περαιτέρω βελτίωσης.



(a) Το αρχικό φασματογράφημα

(b) Οι σημαντικές του περιοχές για την απόφαση με βάση το LIME

Figure 1.2.6: Δύο εικόνες που παρουσιάζουν ένα mel-φασματογράφημα και την αντίστοιχή επεξήγηση από το LIME.

Δεδομένου ότι το LIME για επεξηγήσεις εικόνων δεν έδωσε ικανοποιητικά αποτελέσματα, εξερευνήσαμε εναλλακτικές μεθόδους. Η έρευνά μας μάς οδήγησε στο audioLIME [25], μια παραλλαγή LIME ειδικά προσαρμοσμένη για δεδομένα μουσικής. Σε αντίθεση με το παραδοσιακό LIME, το audioLIME διαταράσσει άμεσα τον ίδιο τον ήχο. Η υλοποίηση των συγγραφέων χρησιμοποιεί την τεχνολογία διαχωρισμού πηγών spleeter [26] για να απομονώσει μεμονωμένα στοιχεία ενός τραγουδιού όπως φωνητικά, ντραμς, μπάσο και πιάνο. Επιπλέον, το audioLIME τμηματοποιεί τον ήχο σε χρονικά κομμάτια. Τέλος αξιολογεί τον αντίκτυπο κάθε τμήματος στη διαδικασία λήψης αποφάσεων του μοντέλου. Μια εικόνα που απεικονίζει αυτήν τη διαδικασία μπορεί να φανεί στο σχήμα 1.2.7. Ωστόσο, αναγνωρίζοντας τους περιορισμούς στην ικανότητα του spleeter να αποσυνθέτει με ακρίβεια σύνθετα σήματα ήχου, επιλέξαμε να ενσωματώσουμε μια πιο επιτυχημένη τεχνική παραγοντοποίησης, το open-unmix [57] (UMX), το οποίο είναι ένα βαθύ νευρωνικό δίκτυο σχεδιασμένο για ακριβή διαχωρισμό της

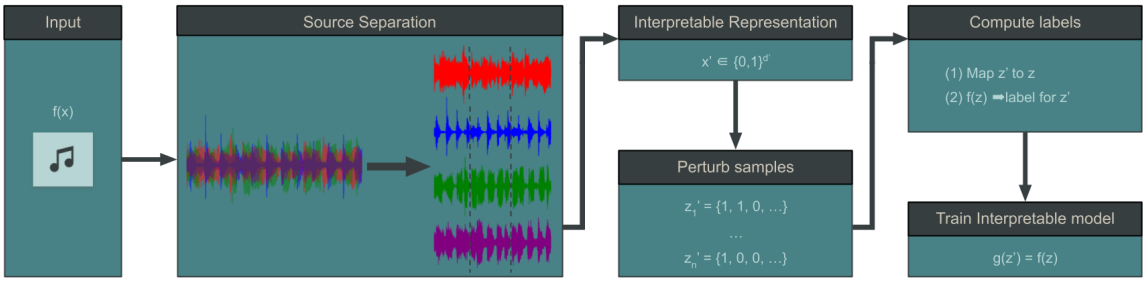


Figure 1.2.7: Η διαδικασία που ακολουθεί το audioLIME [25]. Ακολουθεί στενά την διαδικασία του LIME με την διαφορά οτι χρησιμοποιεί διαχωρισμό πηγής.

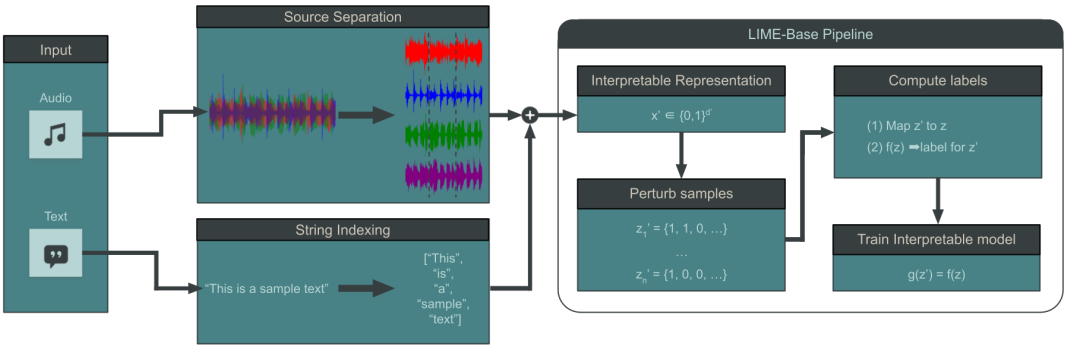


Figure 1.2.8: Η διαδικασία δημιουργίας πολυτροπικών επεξηγήσεων. Όπως φαίνεται στην εικόνα πρώτα χωρίζουμε τον ήχο σε χρονικά διαστήματα και στις επιμέρους πηγές του και το κείμενο σε διακριτά συστατικά όπως λέξεις. Έπειτα δημιουργούμε ερμηνεύσημες αναπαραστάσεις των χαρακτηριστικών και εφαρμόζουμε το LIME-base.

πηγής ήχου.

Όσον αφορά την πολυτροπικότητα και τις επεξηγήσεις, επηρεασμένοι από τα ενδιαφέροντα αποτελέσματα του LIME και του audioLIME στις μονοτροπικές προσεγγίσεις, προτείνουμε έναν τρόπο συνδυασμού τους προκειμένου να εξηγηθεί το πολυτροπικό μοντέλο και να παραχθούν πολυτροπικές εξηγήσεις. Ξεκινάμε δημιουργώντας ένα δυαδικό διάνυσμα που υποδεικνύει την παρουσία ή την απουσία ενός χαρακτηριστικού. Το διάνυσμα έχει μήκος ίσο με το άθροισμα του αριθμού των χαρακτηριστικών κειμένου και του αριθμού των χαρακτηριστικών ήχου. Για παράδειγμα, ας υποθέσουμε ότι έχουμε τον στίχο "Come as you are as you were" ως εισαγωγή κειμένου και ένα ηχητικό κλιπ δύο δευτερολέπτων, που έχει οριστεί να χωριστεί από τον επεξηγητή σε τμήματα ενός δευτερολέπτου, να παραγοντοποιηθεί σε "vocals", "drums" και "other". Ο συνολικός αριθμός χαρακτηριστικών είναι 13, 7 χαρακτηριστικά κειμένου που αντιπροσωπεύουν κάθε λέξη και 6 χαρακτηριστικά ήχου που αντιπροσωπεύουν τα φωνητικά, τα τύμπανα και τα υπόλοιπα κλιπ ήχου 1 δευτερολέπτου. Η διαδικασία δημιουργίας επεξηγήσεων συνεχίζεται παρόμοια με το LIME, δημιουργώντας διαταραχές των εισόδων και υπολογίζοντας τα βάρη κάθε χαρακτηριστικού. Ως αποτέλεσμα, μπορούμε να προσδιορίσουμε ποια χαρακτηριστικά είναι σημαντικά για την απόφαση του πολυτροπικού μοντέλου, να συγκρίνουμε τις τροπικότητες και να έχουμε ακροάσιμες εξηγήσεις καθώς και κειμενικές. Η διαδικασία αυτή απεικονίζεται στην εικόνα 1.2.8

Local explanations often fail to reflect the model's overall behavior. For a more comprehensive understanding of what features are influential across the model, rather than just specific instances we implement Global Aggregations of Local Explanations as outlined in this paper [35]. The authors first mention the global

average class importance for feature j and class c defined as:

$$I_{cj}^{\text{AVG}} = \frac{\sum_{i \in S_c} |W_{ij}|}{\sum_{i \in S_c : W_{ij} \neq 0} 1}$$

where S_c includes all the instances i classified as class c and W_{ij} is the weight of that feature for that specific instance. Another aggregation method described in this work begins by calculating the LIME the importance for a class c for a feature j as

$$I_{cj}^{\text{LIME}} = \sqrt{\sum_{i \in S_c} |W_{ij}|}$$

The authors then calculate the vector of normalize LIME importance per class:

$$p_{cj} = \frac{\sqrt{\sum_{i \in S_c} |W_{ij}|}}{\sum_{c \in L} \sqrt{\sum_{i \in S_c} |W_{ij}|}}$$

where L is the set of all labels. The normalized LIME importance p_j represents the distribution of feature j's importance over all classes $c \in L$. The Shannon entropy of this distribution is defined by:

$$H_j = -\sum_{c \in L} p_{cj} \log(p_{cj})$$

and is used to asses the degree of homogeneity with which the feature attributions of a feature are distributed over multiple classes. Finally in order make out cases where features appear only once in the test set the authors calculate the homogeneity weighted importance for feature j and class c:

$$I_{cj}^{H} = \left(1 - \frac{H_j - H_{\min}}{H_{\max} - H_{\min}}\right) I_{cj}^{\text{LIME}}$$

where H_{min} and H_{max} are the minimum and maximum entropy measured across all features. In short $I_{cj}H$ addresses the issue of common features that may appear significant due to their presence rather than their informative value and corrects for this by using entropy to penalize features that are uniformly distributed across classes, highlighting those that are truly predictive of specific outcomes. It also addresses the assumption that features uniformly affect model outcomes, by adjusting the importance based on the consistency of their influence across classes and therefore penalizing features that show high variability across classes. This method ensures that the global importance reflects genuine, consistent predictive value, especially in complex multiclass scenarios.

Οι τοπικές εξηγήσεις συχνά αποτυγχάνουν να αντικατοπτρίσουν τη συνολική συμπεριφορά του μοντέλου. Για μια πιο ολοκληρωμένη εικόνα, αντί για μεμονωμένες περιπτώσεις, εφαρμόζουμε συνολικούς συνδιασμούς (global aggregates) τοπικών επεξηγήσεων όπως περιγράφεται σε αυτήν τη δουλειά [35]. Οι συγγραφείς αναφέρουν πρώτα την παγκόσμια μέση σημασία κλάσης για το χαρακτηριστικό j και την κλάση c που ορίζεται ως:

$$I_{cj}^{\text{AVG}} = \frac{\sum_{i \in S_c} |W_{ij}|}{\sum_{i \in S_c : W_{ij} \neq 0} 1}$$

όπου το S_c περιλαμβάνει όλες τις περιπτώσεις i που ταξινομούνται ως κλάση c και W_{ij} είναι το βάρος αυτού του χαρακτηριστικού για τη συγκεκριμένη περίπτωση. Μια άλλη μέθοδος συνάθροισης που περιγράφεται σε αυτή την εργασία ξεκινά με τον υπολογισμό της σημασίας του LIME για μια κλάση c για ένα χαρακτηριστικό j ως

$$I_{cj}^{\text{LIME}} = \sqrt{\sum_{i \in S_c} |W_{ij}|}$$

Στη συνέχεια, οι συγγραφείς υπολογίζουν το διάνυσμα της σημασίας του κανονικοποιημένου LIME ανά κατηγορία:

$$p_{cj} = \frac{\sqrt{\sum_{i \in S_c} |W_{ij}|}}{\sum_{c \in L} \sqrt{\sum_{i \in S_c} |W_{ij}|}}$$

όπου L είναι το σύνολο όλων των ετικετών. Η κανονικοποιημένη σημασία LIME p_j αντιπροσωπεύει την κατανομή της σημασίας του χαρακτηριστικού j σε όλες τις κλάσεις $c \in L$. Η εντροπία Shannon αυτής της κατανομής ορίζεται από:

$$H_j = -\sum_{c \in L} p_{cj} \log(p_{cj})$$

και χρησιμοποιείται για να εκτιμήσει τον βαθμό ομοιογένειας με τον οποίο οι αποδόσεις χαρακτηριστικών κατανέμονται σε πολλαπλές κλάσεις. Τέλος, προκειμένου να διακριθούν περιπτώσεις όπου τα χαρακτηριστικά εμφανίζονται μόνο μία φορά στο σύνολο δοκιμής, οι συγγραφείς υπολογίζουν τη σταθμισμένη σημασία ομοιογένειας για το χαρακτηριστικό j και την κλάση c:

$$I_{cj}^{H} = \left(1 - \frac{H_j - H_{\min}}{H_{\max} - H_{\min}}\right) I_{cj}^{\text{LIME}}$$

όπου H_{min} και H_{max} είναι η ελάχιστη και η μέγιστη εντροπία που μετράται για όλα τα χαρακτηριστικά. Εν συντομία, το $I_{cj}H$ αντιμετωπίζει το ζήτημα των κοινών χαρακτηριστικών που μπορεί να φαίνονται σημαντικά λόγω της παρουσίας τους και όχι λόγω της πληροφοριακής τους αξίας και το διορθώνει χρησιμοποιώντας εντροπία για να τιμωρήσει χαρακτηριστικά που είναι ομοιόμορφα κατανεμημένα στις κλάσεις. Αντιμετωπίζει επίσης την υπόθεση ότι τα χαρακτηριστικά επηρεάζουν ομοιόμορφα τα αποτελέσματα του μοντέλου, προσαρμόζοντας τη σημασία με βάση τη συνέπεια της επιρροής τους μεταξύ των τάξεων και επομένως τιμωρώντας χαρακτηριστικά που παρουσιάζουν υψηλή μεταβλητότητα μεταξύ των τάξεων. Αυτή η μέθοδος διασφαλίζει ότι η παγκόσμια σημασία αντανακλά γνήσια, σταθερή προγνωστική αξία, ειδικά σε πολύπλοκα σενάρια πολλαπλών κατηγοριών.

1.3 Αποτελέσματα

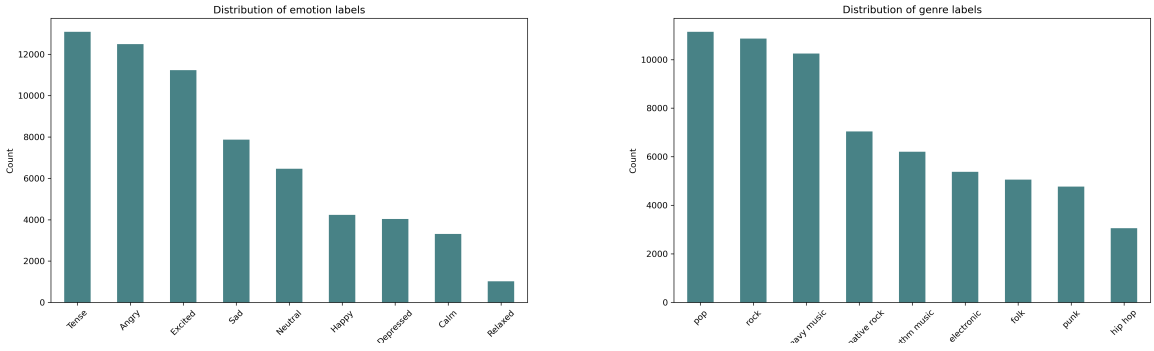
1.3.1 Το τελικό σύνολο δεδομένων

Μετά τις διαδικασίες επαύξησης και ξεκαθάρησης του συνόλου δεδομένων M4A, είμαστε έτοιμοι να παρουσιάσουμε τη σύνοψη του συνόλου δεδομένων που προκύπτει. Ο συνολικός αριθμός εγγραφών είναι 63760, που περιλαμβάνουν τους στίχους καθώς και τα διακριτικά (tokens) του RoBERTa μοντέλου, τον ήχο καθώς και το φασματόγραμμα που δημιουργήθηκε και διάφορα άλλα μεταδεδομένα, συμπεριλαμβανομένης μιας ετικέτας συναισθήματος και μιας ετικέτας είδους για κάθε τραγούδι. Η κατανομή των ετικετών συναισθημάτων και των ετικετών του είδους φαίνεται στο Σχήμα 5.1.1. Σημειώνουμε ότι παρά τις προσπάθειές μας να συλλέξουμε επιπλέον «relaxed» κομμάτια, παρουσιάστηκε δύσκολο να εμπλουτίσουμε το σετ μας με περισσότερα από τα 1020 που συμπεριλάβαμε σε αυτό (498 εκ των οποίων προϋπήρχαν στο σύνολο δεδομένων M4A). Αυτή η έλλειψη θα μπορούσε να αποδοθεί όχι μόνο στο γεγονός ότι τέτοια τραγούδια δεν είναι πολύ δημοφιλή στη δυτική μουσική, αλλά επίσης μπορεί να μην περιέχουν αγγλικούς ή και καθόλου στίχους. Μπορούμε επίσης να δούμε μια παρόμοια αλλά λιγότερο έντονη ανισορροπία στη διανομή του είδους, με υπερπληθή τα «pop» τραγούδια ενώ τα τραγούδια «hip hop» εμφανίζονται λιγότερο συχνά. Από όλες τις καταχωρήσεις, 50660 χρησιμοποιήθηκαν ως σετ εκπαίδευσης και η επαναφορά μοιράστηκε μεταξύ της επικύρωσης και του σετ δοκιμής. Οι καλλιτέχνες που εμφανίζονται στα σετ εκπαίδευσης και επικύρωσης δεν περιλαμβάνονται στο δοκιμαστικό σετ.

1.3.2 Οι επιδόσεις των μοντέλων μας

Αναλύοντας τα αποτελέσματα των προβλημάτων παλινδρόμησης (regression), που απεικονίζονται στον Πίνακα 1.1, κάνουμε τις ακόλουθες παρατηρήσεις. Πρώτον, το μέσο απόλυτο σφάλμα (MAE) του mean regressor είναι περίπου 0,2 για προβλέψεις του βαθμού ευφορίας (valence) και ενέργειας (energy), χρησιμεύει ως βάση για την αξιολόγηση πιο περίπλοκων αρχιτεκτονικών. Είναι αξιοσημείωτο οτι η απλή αρχιτεκτονική MLP έδωσε πολύ επιτυχημένα αποτελέσματα. Από την άλλη πλευρά, παρά τον καινοτόμο σχεδιασμό του MuLaN, η υλοποίησή μας χρησιμοποιώντας τη βιβλιοθήκη musiclm-pytorch οδηγεί σε μέτρια απόδοση. Τέλος, η καλύτερη απόδοση της υλοποίησης RoBERTA και AST, ειδικά για προβλέψεις βαθμού ευφορίας, μας οδήγησε να την υιοθετήσουμε για τις επερχόμενες κατηγοριοποιήσεις διάθεσης και είδους.

Γενικά, τα μοντέλα ήταν καλύτερα στην πρόβλεψη τιμών ενέργειας (energy) σε αντίθεση με τον βαθμό ευφορίας (valence). Αυτή η απόκλιση μπορεί να αποδοθεί στα χαρακτηριστικά κατανομής του συνόλου δεδομένων, όπου τα δεδομένα βαθμού ευφορίας παρουσιάζουν μια πιο ομοιόμορφη κατανομή, ενώ οι τιμές ενέργειας είναι συσσωρευμένες στο 1, όπως αντανακλάται επίσης στις προβλέψεις του mean regressor. Αξίζει να σημειωθεί η ταχεία πρόοδος στον τομέα των transformer της Βαθιάς Μάθησης.



(a) Κατανομή ετικετών συναισθήματος

(b) Κατανομή ετικετών είδους.

Figure 1.3.1: Κατανομές ετικετών συναισθήματος και είδους του τελικού συνόλου δεδομένων.

Μοντέλο	Μέσο Απόλυτο Σφάλμα			Τροπικότητες	
	Βαθμ. Ευφορίας	Ενέργεια	Μέσος	Ήχος	Στίχοι
Dummy	0.314	0.342	0.328	-	-
Mean	0.206	0.194	0.200	-	-
Music4All LSTM	0.210	0.197	0.203	-	Embedding
XLNet	0.183	-	-	-	XLNet emb.
Conv. Net.*	0.158	0.107	0.211	Φασματογράμματα	Word2Vec emb.
MLP	0.157	**0.085**	0.121	Χαρ/κά Υψηλού Επιπέδου	Vader, TF-IDF
MuLaN*	0.195	0.148	0.171	Φασματογράμματα	
AST + RoBERTa	**0.140**	0.099	**0.119**	Φασματογράμματα	RoBERTa emb.

Table 1.1: Σύγκριση αποδόσεων των μοντέλνω καθώς και χρήση τροπικότητας. Τα μοτνέλα σημασμένα με * εκπαιδεύτηκαν σε υποσύνολο των διαθέσιμων δεδομένων λόγω υπολογιστικού κόστους.

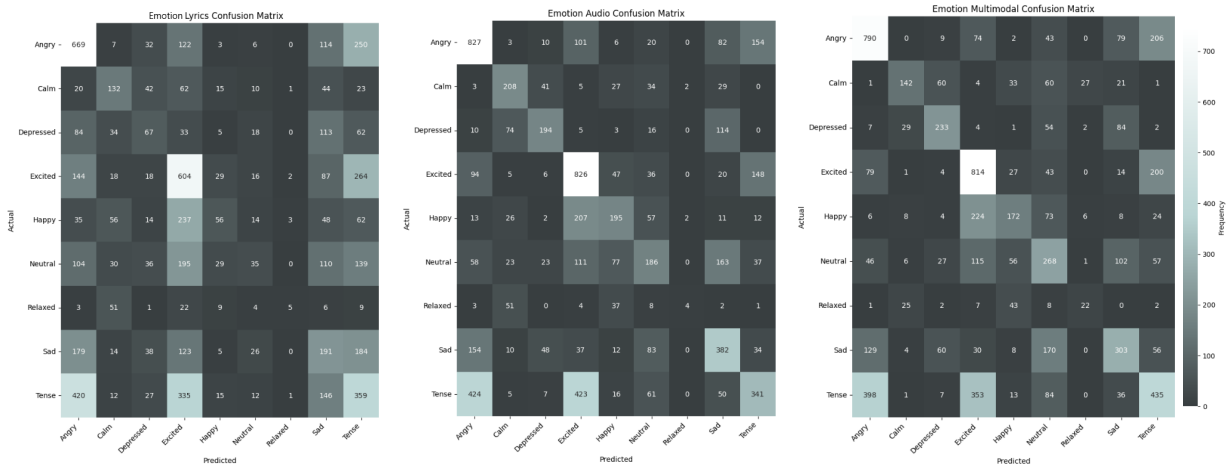
Επηρεασμένοι απο τα παραπάνω εκπαιδεύουμε και αξιολογούμε τα μοντέλα ταξινόμησής. Μια περίληψη της απόδοσης αυτών μπορείτε να δείτε στον Πίνακα 1.2. Για να αποκτήσουμε μια βαθύτερη εικόνα των δυνατοτήτων και των αδυναμιών των μοντέλων μας, παρέχουμε λεπτομερείς πίνακες σύγχυσης και αναφορές ταξινόμησης. Προχωράμε επιδεικνύοντας και αναλύοντας τα μοντέλα που προκύπτουν. Περαιτέρω έρευνα σχετικά με τη συμπεριφορά των μοντέλων μπορεί να βρεθεί στην επόμενη ενότητα, όπου εφαρμόζουμε μεθόδους επεξήγησης για την εξαγωγή ακριβών συμπερασμάτων.

Οι αναφορές ταξινόμησης και οι πίνακες σύγχυσης για τα μοντέλα συναισθημάτων απεικονίζονται στον Πίνακα 1.3 και στο σχήμα 1.3.2 αντίστοιχα. Το μοντέλο στίχων για την πρόβλεψη ετικετών συναισθημάτων είναι το μοντέλο με τη χειρότερη απόδοση από όλα, με ακρίβεια επικύρωσης μόλις 34,03%, υποδηλώνοντας ότι η ανίχνευση πληροφοριών συναισθημάτων από το στιχουργικό πλαίσιο δεν είναι αξιόπιστη και ότι τέτοιες πληροφορίες μπορούν να ανιχνευθούν καλύτερα μέσω της ανάλυσης χαρακτηριστικών ήχου. Αυτό υποδηλώνεται επίσης όχι μόνο από το γεγονός ότι το ηχητικό μοντέλο ξεπερνά σε μεγάλο βαθμό το λυρικό, αλλά και από την παρατήρηση ότι η πολυτροπική προσέγγιση δεν αποφέρει σημαντικά καλύτερα αποτελέσματα από την ηχητική.

Οι αναφορές ταξινόμησης στον πίνακα 1.4 και οι πίνακες σύγχυσης στο σχήμα 1.3.3 απεικονίζουν τα αποτελέσματα των μοντέλων ταξινόμησης είδους. Το λυρικό μοντέλο, αν και έχει σχετικά χαμηλές μετρικές, με ακρίβεια επικύρωσης 46,9% είναι ιδιαίτερα επιτυχής στην πρόβλεψη ορισμένων κατηγοριών, συγκεκριμένα «hip hop» και «heavy music». Αν και το μοντέλο ήχου δεν είναι εξίσου ακριβές στις προβλέψεις περιπτώσεων «hip hop», ξεπερνά το λυρικό μοντέλο σε όλες τις άλλες κατηγορίας με ακρίβεια επικύρωσης 55,3%. Τέλος, το πολυτροπικό

Μοντέλο	Ακρίβεια Επικύρωσης	Ακρίβεια Τεστ	Εποχές (καλύτερη)
Λυρικό Συναισθήματος	34.03%	32.33%	9 (5)
Ηχητικό Συναισθήματος	48.33%	48.29%	5 (3)
Πολυτροπικό Συναισθήματος	49.05%	48.53%	5 (3)
Λυρικό Είδους	46.9%	45.14%	9 (7)
Ηχητικό Είδους	55.63%	53.75%	5 (4)
Πολυτροπικό Είδους	60.33%	57.34%	5 (2)

Table 1.2: Σύνοψη της απόδοσης των μοντέλων



(a) Λυρικό Μοντέλο (b) Ηχητικό Μοντέλο (c) Πολυτροπικό Μοντέλο

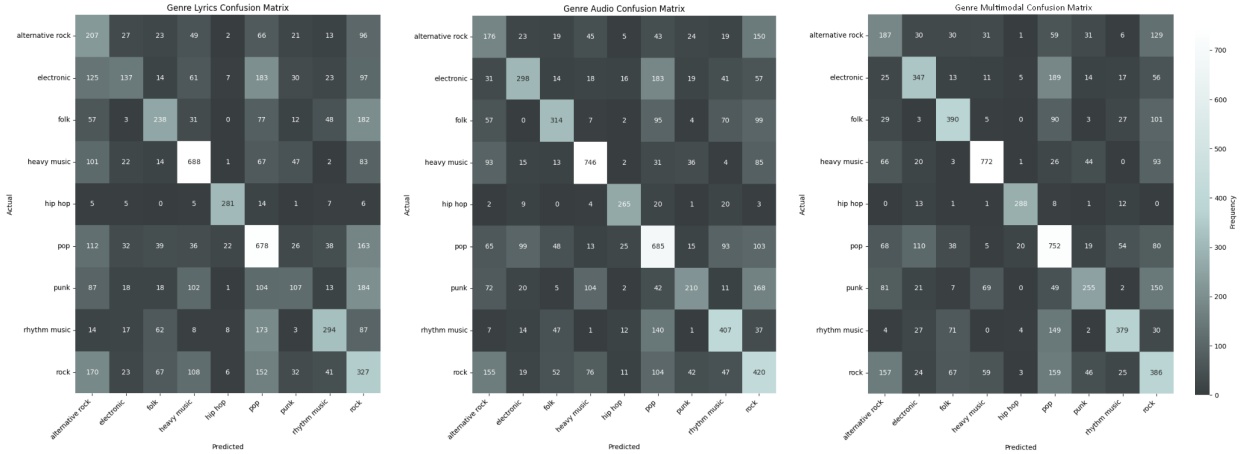
Figure 1.3.2: Οι πίνακες σύγχησης της ταξινόμησης συναισθήματος με βάση τους στίχους, τον ήχο και τον συνδιασμό τους. Φωτεινότερα κουτιά υποδηλώνουν περισσότερες προβλέψεις του μοντέλου. Ιδανικά μονο η διαγώνιος θα έπρεπε να περιέχει τιμές και ο υπόλοιπος πίνακας να είναι κενός.

μοντέλο ξεπερνά και τα δύο αυτά μοντέλα με ακρίβεια επικύρωσης 60,33% αποδεικνύοντας ότι είναι σε θέση να ανιχνεύσει πληροφορίες που δεν είναι ευδιάκριτες από τις μονοτροπικές προσεγγίσεις ξεχωριστά.

Μερικά ενδιαφέροντα μοτίβα προκύπτουν από τις μετρικές των μοντέλων μας με βάση το σκοπό ταξινόμισης. Αρχικά, τα μοντέλα που είναι υπεύθυνα για την πρόβλεψη ετικετών ειδών ξεπερνούν κατά πολύ αυτά για την πρόβλεψη συναισθημάτων. Αυτό το μοτίβο επιβεβαιώνει τη διαίσθησή μας ότι η αποτύπωση της συναισθηματικής πολυπλοκότητας των μουσικών κομματιών μπορεί να είναι δύσκολη λόγω της υποκειμενικής φύσης των ανθρώπινων συναισθημάτων και της δυσκολίας παροχής καθώς και ανίχνευσης γενικευμένων, αντικειμενικών και επομένως ακριβών εκτιμήσεων των συναισθηματικών χαρακτηριστικών που μπορεί να αντιληφθούν οι ακροατές σε ένα κομμάτι [21]. Από την άλλη πλευρά, ο εντοπισμός θεμάτων που επικρατούν σε ορισμένα είδη και ως εκ τούτου η σωστή εκμάθηση πρόβλεψης ετικετών ειδών μπορεί να είναι πιο επιτυχημένη. Θα πρέπει επίσης να σημειώσουμε ότι τα μοντέλα συναισθημάτων υστερούσαν ίσως και λόγω της πιο έντονης ανισορροπίας των δεδομένων. Επιπλέον, οι προσπάθειές μας δείχνουν ότι και για τις δύο εργασίες τα λυρικά μοντέλα είχαν τη χειρότερη απόδοση και τα πολυτροπικά μοντέλα πέτυχαν τα καλύτερα αποτελέσματα. Αυτή η συμπεριφορά μας οδηγεί στο συμπέρασμα ότι οι πληροφορίες που σχετίζονται με εργασίες ταξινόμησης μουσικής, ειδικά για την ταξινόμηση συναισθημάτων, υπάρχουν κυρίως στον τομέα του ήχου. Η απόδοση του συνδυασμού των τρόπων κειμένου και ήχου ήταν η βέλτιστη αποδεικνύοντας τις δυνατότητες της πολυτροπικότητας στη βελτίωση των αποτελεσματων ταξινόμησης μουσικής.

Table 1.3: Αναφορές ταξινόμησης συναισθήματος των τριών μοντέλων

Συναίσθημα	Λυρικό Ακρίβεια, Ανάκληση, F1	Ηχητικό Ακρίβεια, Ανάκληση, F1	Πολυτροπικό Ακρίβεια, Ανάκληση, F1	Υποστ.
Angry	0.40, 0.56, 0.47	0.52, 0.69, 0.59	0.54, 0.66, 0.59	1203
Calm	0.37, 0.38, 0.38	0.51, 0.60, 0.55	0.66, 0.41, 0.50	349
Depressed	0.24, 0.16, 0.19	0.59, 0.47, 0.52	0.57, 0.56, 0.57	416
Excited	0.35, 0.51, 0.41	0.48, 0.70, 0.57	0.50, 0.69, 0.58	1182
Happy	0.34, 0.11, 0.16	0.46, 0.37, 0.41	0.48, 0.33, 0.39	525
Neutral	0.25, 0.05, 0.09	0.37, 0.27, 0.32	0.33, 0.40, 0.36	678
Relaxed	0.42, 0.05, 0.08	0.50, 0.04, 0.07	0.38, 0.20, 0.26	110
Sad	0.22, 0.25, 0.24	0.45, 0.50, 0.47	0.47, 0.40, 0.43	760
Tense	0.27, 0.27, 0.27	0.47, 0.26, 0.33	0.44, 0.33, 0.38	1327



(a) Λυρικό Μοντέλο (b) Ηχητικό Μοντέλο (c) Πολυτροπικό Μοντέλο

Figure 1.3.3: Οι πίνακες σύγχησης της ταξινόμησης είδους με βάση τους στίχους, τον ήχο και τον συνδιασμό τους. Φωτεινότερα κουτιά υποδηλώνουν περισσότερες προβλέψεις του μοντέλου. Ιδανικά μονο η διαγώνιος θα έπρεπε να περιέχει τιμές και ο υπόλοιπος πίνακας να είναι κενός.

1.3.3 Επεξήγηση των μοντέλων

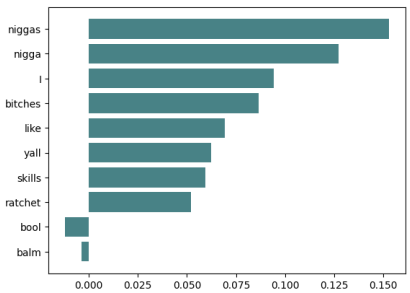
Όπως προαναφέρθηκε θα χρησιμοποιήσουμε το LIME, το audioLIME και το συνδιαστικό MusicLIME για να μελετήσουμε την συμπεριφορά των μοντέλων. Επιπλέον, υπολογίζουμε και τους συνολικούς συνδυασμούς τοπικών επεξηγήσεων με δύο μεθόδους.

Ξεκινώντας με το λυρικό μοντέλο πρόβλεψης ειδών μουσικής, παρουσιάζοντας τόσο τοπικές εξηγήσεις για επιλεγμένα παραδείγματα για κάθε κλάση όσο και τους συνολικούς συνδυασμούς παρατηρούμε κάποια ενδιαφέροντα μοτίβα. Αρχικά, η "hip hop" κατηγορία ξεχωρίζει έντονα, με την ανάλυσή μας να αναδεικνύει την ικανότητα του μοντέλου να εντοπίσει θεματολογία παρούσα σε στίχους τέτοιων τραγουδιών. Αυτή εμπεριέχει την πάλη, κυρίως αφροαμερικάνων, με λέξεις όπως "prison" και "dead", την ζωή στον δρόμο, π.χ. οι λέξεις "hood" και "block" άλλα κυρίως το μοντέλο επικεντρώνεται σε βωμολοχίες, προσβλητικό και σεξιστικό λόγο. Όσον αφορά την κλάση "heavy music" μοτίβα βίας, με λέξεις σαν "blood", "wrath", "destruction" και "killer", συναισθηματικού πόνου, με λέξεις όπως "fear", "misery" και "grief" και θανάτου και θρησκευτικού χαρακτήρα, με λέξεις όπως "dead", "grave", "evil" και "cursed" φαίνεται να επηρεάζουν το μοντέλο. Για την κατηγορία "rhythm music" το μοντέλο φαίνεται να ανιχνεύει κυρίως το Ρασταφαριανό όνομα για τον Θεό "Jah", που έχει έντονη παρουσία στην τζαμαϊκανή μουσική καθώς και άλλες λέξεις θρησκευτικού και γεωγραφικού περιεχομένου και λέξεις που ανήκουν στην Καραϊβική διάλεκτο. Το μοντέλο επίσης αποδίδει αξιοπρεπώς κατά τις προβλέψεις του

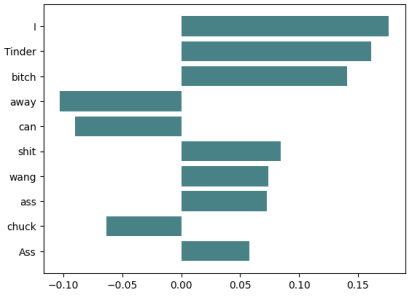
Table 1.4: Classification Reports for Three Models

Συναίσθημα	Λυρικό Ακρίβεια, Ανάκληση, F1	Ηχητικό Ακρίβεια, Ανάκληση, F1	Πολυτροπικό Ακρίβεια, Ανάκληση, F1	Υποστ.
Alternative Rock	0.24, 0.41, 0.30	0.27, 0.35, 0.30	0.30, 0.37, 0.33	504
Electronic	0.48, 0.20, 0.29	0.60, 0.44, 0.51	0.58, 0.51, 0.55	677
Folk	0.50, 0.37, 0.42	0.61, 0.48, 0.54	0.63, 0.60, 0.62	648
Heavy Music	0.63, 0.67, 0.65	0.74, 0.73, 0.73	0.81, 0.75, 0.78	1025
Hip Hop	0.86, 0.87, 0.86	0.78, 0.82, 0.80	0.89, 0.89, 0.89	324
Pop	0.45, 0.59, 0.51	0.51, 0.60, 0.55	0.51, 0.66, 0.57	1146
Punk	0.38, 0.17, 0.23	0.60, 0.33, 0.43	0.61, 0.40, 0.49	634
Rhythm Music	0.61, 0.44, 0.51	0.57, 0.61, 0.59	0.73, 0.57, 0.64	666
Rock	0.27, 0.35, 0.30	0.37, 0.45, 0.41	0.38, 0.42, 0.40	926

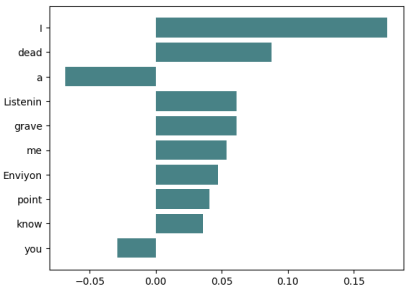
για την κατηγορία "pop". Εντοπίζει ρομαντικά θέματα, όπως υποδεικνύεται απο τα βάρη των λέξεων "kiss", "girlfriend", "sexy" κ.α., θέματα αναφορικά με τον χορό και την διασκέδασή, παραδείγματος χάρη "club", "dance" και "disco" αλλά και λέξεις που υποδηλώνουν συναισθηματικές εμπειρίες και μόδα. Για την "folk" κατηγορία σημαντικότερες για την απόφαση του μοντέλου φαίνεται να είναι λέξεις απο αναφέρονται στην φύση, σε προσωπικές σχέσεις και στην ζωή στην (μικρή) πόλη. Τέλος το μοντέλο δυσκολεύεται να εντοπίσει τραγούδια που ανήκουν στις κατηγορίες "alternative rock", "rock", "electronic" και "punk" συγχαίοντάς τες είτε μεταξύ τους είτε με άλλες κατηγορίες. Καταλήγουμε ότι ένας βαθμός σύγχυσης είναι αναμενόμενος τόσο επειδή κάποιο τραγούδι μπορεί να ανήκει σε πολλά είδη όσο και επειδή διαφορετικά είδη μπορεί να περιέχουν κοινά θέματα αλλά να διαφέρουν σημαντικά στον ρυθμό, τις κλίμακες που χρησιμοποιούν και σε άλλα μουσικά χαρακτηριστικά που δεν είναι παρόντα σε μια ανάλυση στίχων. Ένα παράδειγμα τοπικών εξηγήσεων για την κατηγορία "hip hop" για τρία τραγούδια βρίσκονται στην εικόνα 1.3.4 ενώ οι 5 σημαντικότερες λέξεις για κάθε κατηγορία με βάση τους συνολικούς συνδυασμούς βρίσκονται στην εικόνα 1.3.5. Στις εικόνες 1.3.6 δίνουμε δύο παραδείγματα κλάσεων των οποίων τα θέματα είναι εύκολα διαχωρίσιμα και κλάσεων που δεν ξεχωρίζουν τόσο καλά.



(a) Bool, Balm & Bollective του YG, ορθά προβλεπόμενο ως "hip hop" με αρκετή σιγουριά.



(b) Poontang Boomerang των Steel Panther, που ανήκει στην κατηγορία "heavy music" άλλα η πρόβλεψη ήταν "hip hop".



(c) Lucid Dreams του Juice WRLD, που προβλέθηκε ως "pop" αλλα ανήκει στην κατηγορία "hip hop".

Figure 1.3.4: Τοπικές εξηγήσεις για την κατηγορία "hip hop" για τρία δείγματα του τεστ συνόλου: (a) αληθώς θετικό (b) ψευδώς θετικό και (c) ψευδώς αρνητικό. Τα γραφήματα δείχνουν ποιες λέξεις συνεισφέρουν περισσότερο στο να προβλέψει το μοντέλο "hip hop".

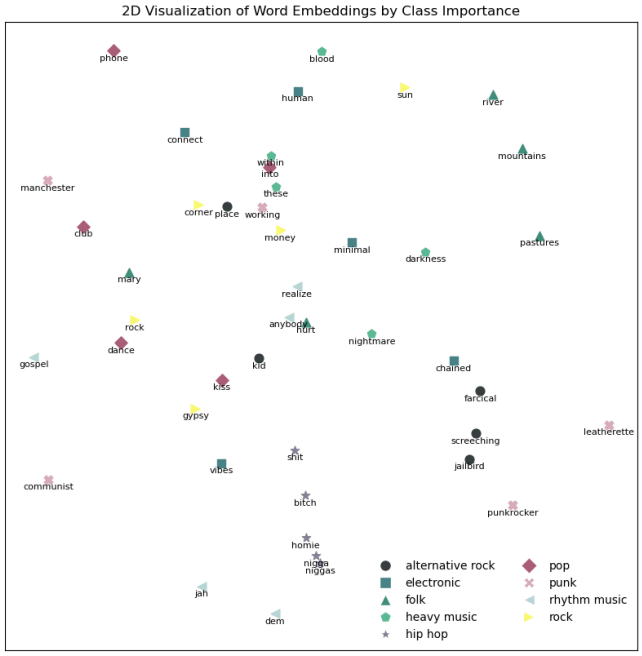
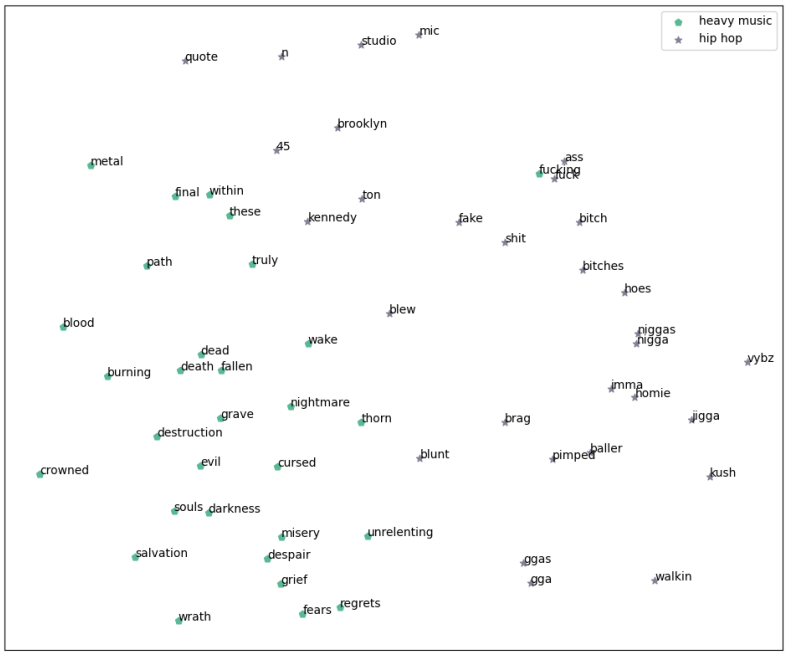
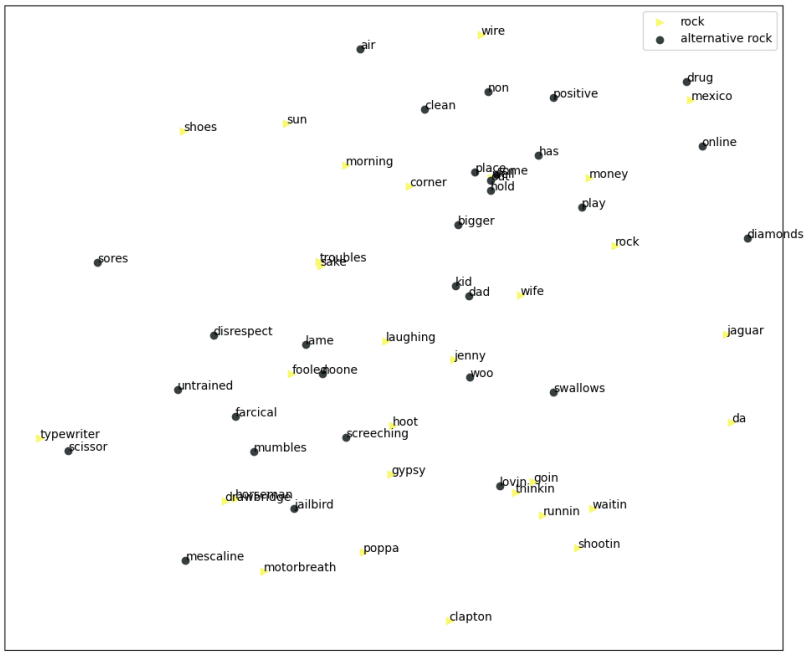


Figure 1.3.5: Οι 5 σημαντικότερες λέξεις γαι κάθε κατηγορία με βάση τους συνολικούς συνδιασμούς.



(a) Οι 30 πιο σημαντικές λέξεις για τις κατηγορίες "hip hop" και "heavy music" με βάση το μοντέλο. Οι θεματολογίες είναι εύκολα διαχωρίσιμες.



(b) Οι 30 πιο σημαντικές λέξεις για τις κατηγορίες "alternative rock" και "rock" με βάση το μοντέλο. Οι θεματολογίες δεν είναι εύκολα διαχωρίσιμες.

Figure 1.3.6: Το t-SNE γράφημα των GloVe embeddings για τις 30 πιο σημαντικές λέξεις ορισμένων κλάσεων. Στην εικόνα (a) παρουσιάζουμε τις λέξεις των κλάσεων "hip hop" και "heavy music" ενώ στην εικόνα (b) τις λέξεις των κλάσεων "alternative rock" and "rock".

Αν και οι επεξηγήσεις του είδους παρέχουν ενδιαφέροντα αποτελέσματα, αυτό από την ανάλυση του λυρικού μοντέλου συναισθημάτων δεν ήταν παρόμοια. Με ανάλογο τρόπο όπως πριν, δημιουργούμε τοπικές επεξηγήσεις και συνολικούς συνδυασμούς να εξιχνιάσουμε τους λόγους της κακής απόδοσης του. Παρατηρούμε ότι το μοντέλο αναγνωρίζει σκοτεινά και μακάβρια θέματα, με λέξεις όπως "dead", "destroy και "bleed" και τα αποδίδει στην κατηγορία "Angry. Αυτή η συμπεριφορά δικαιολογεί το σχετικά υψηλό σκορ F1 του μοντέλου για αυτήν την κατηγορία. Παρομοίως, αλλά με μικρότερη ακρίβεια, το μοντέλο αναγνωρίζει θέματα πάρτι και διασκέδασης και τα αποδίδει στην τάξη "Excited" και θρησκευτικά θέματα, αποδίδοντάς τα στην τάξη "Calm". Ωστόσο, το μοντέλο έχει κακή απόδοση στην αναγνώριση θεμάτων που σχετίζονται με τις άλλες κλάσεις. Τα βάρη των συνολικών συνδυασμών των λέξεων που επηρέασαν το μοντέλο για την πρόβλεψη αυτών των κατηγοριών ήταν χαμηλά, υποδεικνύοντας τη μη αποφασιστικότητα του μοντέλου. Στο σχήμα 1.3.7 παρουσιάζουμε τα 5 χαρακτηριστικά με τη μεγαλύτερη επιρροή για κάθε τάξη. Θα πρέπει να σημειώσουμε ότι ορισμένα χαρακτηριστικά που είναι εκφραστικοί ήχοι όπως "Mmm" ή ονόματα όπως "Bethlehem" δεν έχουν GloVe embedding και επομένως δεν απεικονίζονται στην Εικόνα.

Η αδυναμία του μοντέλου να αποδώσει καλά μπορεί να αποδοθεί σε διάφορους παράγοντες. Πρώτον, η αναγνώριση συναισθημάτων είναι εγγενώς πολύπλοκη και υποκειμενική, εξαρτάται σε μεγάλο βαθμό από τις ανθρώπινες εμπειρίες και ερμηνείες, οι οποίες ποικίλλουν ευρέως μεταξύ των ατόμων [21]. Παρόλο που οι πιο ένοτνες λέξεις μπορεί περιστασιακά να παρέχουν σαφείς ενδείξεις για τις κλάσεις υψηλής ενέργειας "Angry" και "Excited, η μουσική είναι μια πολύπλευρη μορφή τέχνης. Οι στίχοι, αν και σημαντικοί, αντιπροσωπεύουν μόνο μια διάσταση της μουσικής δημιουργίας. Οι κρίσιμες πληροφορίες για την ακριβή ταξινόμηση της μουσικής συχνά βρίσκονται πέρα από το στιχουργικό περιεχόμενο ενώ τα δεδομένα εκπαίδευσης και τεστ σχολιάστηκαν λαμβάνοντας υπόψη τη μουσική και όχι μόνο το λυρικό σώμα των τραγουδιών. Αναμένεται ότι σε περιπτώσεις που οι στίχοι προκαλούν ένα συγκεκριμένο συναίσθημα, αλλά ο ήχος τους εκφράζει ένα διαφορετικό συναίσθημα, προσθέτουν σημαντική σύγχυση στο μοντέλο. Κατά συνέπεια, το μοντέλο δυσκολεύεται να κάνει αποφασιστικές και ακριβείς προβλέψεις, κάτι που είναι εμφανές από την αδυναμία εύρεσης θεματολογιών για κάθε τάξη.

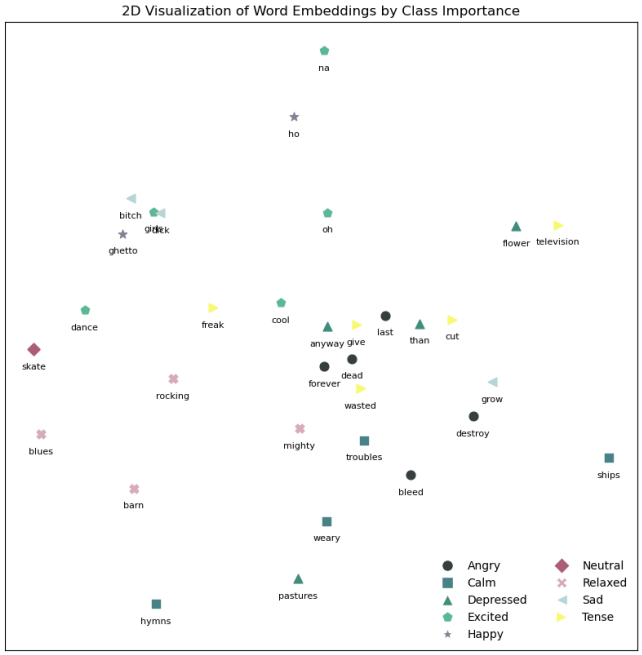
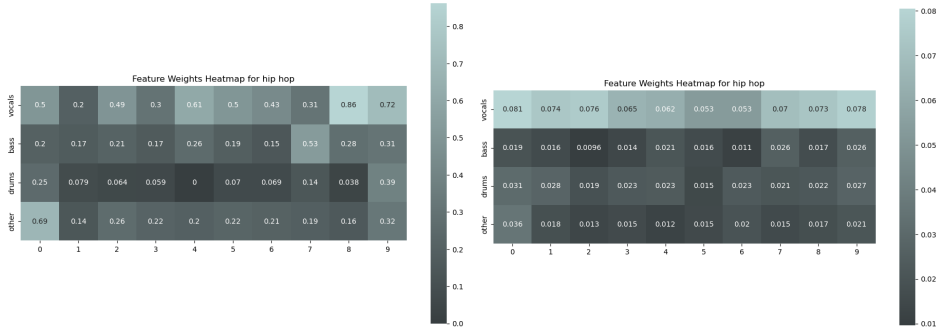


Figure 1.3.7: Οι σημαντικότερες 5 λέξεις με βάση τον συνολικό συνδιασμό τοπικών εξηγήσεων. Λέξεις όπως ονόματα ή επιφωνήματα δεν απεικονίζοντα.

Συνεχίζουμε αναλύοντάς τις εξηγήσεις του ηχητικού μοντέλου που προβλέπει ετικέτες είδους. Ακροάσιμες εξηγήσεις είναι διαθέσιμες στη σελίδα του GitHub. Για να έχουμε μια ολοκληρωμένη εικόνα μελετάμε τόσο τοπικές εξηγήσεις προτότυπων για κάθε κατηγορία όσο και τους συνολικούς συνδυασμούς I_{cj}^{H} και I_{cj}^{AVG}. Ένα παράδειγμα τέτοιων συνολικών εξηγήσεων για την κλάση "hip hop" είναι διαθέσιμο στην εικόνα 1.3.8. Ξεκινώντας λοιπόν από αυτήν την κλάση βλέπουμε ότι τα φωνητικά ωθούν το μοντέλο στο να την προβλέψει, υποδηλώνοντας όχι μόνο την συμπηκνωμένη παρουσία φωνητικών σε τέτοια τραγούδια αλλά και την ικανότητα ανίχνευσης μοτίβα ραπ. Στην "heavy music" φαίνεται οτι φωνητικά και "other" χαρακτηρηστικά που περιέχουν κραυγές και ήχους παραμορφωμένης κιθάρας είναι πιο σημαντικά ενώ αντίστοιχα χαρακτηριστικά, φυσικά με διαφορετικό περιεχόμενο, συνεισφέρουν και στην πρόβλεψη "pop" και "folk" τραγουδιών. Η κλάση "rhythm music" φαίνεται να δίνει έμφαση σε όλα τα χαρακτηριστικά εκτός από τα ντραμς ενώ στην "electronic" τα "other" χαρακτηριστικά, τα οποία συνήθως περιέχουν διακριτές μελωδίες συνθεσάιζερ. Η "punk" κλάση παρουσιάζει σημαντική βελτίωση συγκριτικά με το λυρικό μοντέλο με έμφαση κυρίως στα φωνητικά ενω στην "rock" τα "other" χαρακτηριστικά δείχνουν ότι η ηλεκτρική κιθάρα έχει κύριο ρόλο για την απόφαση του μοντέλου. Αν και το μοντέλο ανιχνεύει στοιχεία της "alternative rock" δυσκολέυεται να τα ξεχωρίσει απο αυτά της "rock" κλάσης. Καταλήγωντας, το ηχητικό μοντέλο ανιχνεύει μουσικά στοιχεία που ξεχωρίζουν κάθε κλάση αλλά μπερδεύει κατηγορίες που είναι κοντά μεταξύ τους. Κάτι τέτοιο είναι αναμενόμενο όπως αναφέρουν και οι συγγραφείς σε αυτήν την δουλειά [43].



(a) Homogeneity importance weights (b) Average weights

Figure 1.3.8: Οι συνολικόι συνδιασμοί τοπικών εξηγήσεων για την κλάση "hip hop". The first heatmap refers to the homogeneity importance weights whereas the second heatmap refers to the average lime weights per feature.

Το μοντέλο ήχου που προβλέπει συναισθήματα είναι σημαντικά ανώτερο σε σύγκριση με το αντίστοιχο στιχουργικό, καθώς ορισμένες κατηγορίες έχουν βελτίωση του σκορ F1 μεγαλύτερη από 0,3 μονάδες. Στους συνολικούς συνδυασμούς της ανάλυσης ηχητικών χαρακτηριστικών, οι κατηγορίες "happy" και "excited" βασίζονται σε μεγάλο βαθμό σε χαρακτηριστικά ντραμς, με την "happy" να δίνει επίσης έμφαση στα φωνητικά, υποδηλώνοντας μια ισχυρή σχέση μεταξύ του ρυθμού και των θετικών συναισθημάτων υψηλής ενέργειας. Για τις "neutral" και τις "calm" κατηγορίες, κυριαρχούν τα χαρακτηριστικά "other", με τα μπάσα χαρακτηριστικά λιγότερο σημαντικά για τα "neutral" συναισθήματα. Θα πρέπει να σημειώσουμε ότι εφόσον το μοντέλο δεν κάνει πολλές "relaxed" προβλέψεις στο δοκιμαστικό σύνολο, δεν έχουμε αρκετά δεδομένα για να προσδιορίσουμε οριστικά τη σημαντικότητα των χαρακτηριστικών. Στα συναισθήματα "angry" και "tense", τα "other" χαρακτηριστικά είναι πιο σημαντικά, αλλά τα "tense" δίνουν έμφαση επίσης στα φωνητικά, υποδεικνύοντας ένα περίπλοκο μείγμα που απαιτείται για να μεταδώσει αρνητικά και έντονα συναισθήματα. Οι κατηγορίες "sad" και "depressed" παρουσιάζουν αποκλίνοντα μοτίβα. Τα ντραμς είναι βασικά για "sad" τραγούδια αλλά αμελητέα για "depressed", αναδεικνύοντας πώς τα επίπεδα ενέργειας επηρεάζουν τη συνάφεια των χαρακτηριστικών αυτών. Τέλος, η κλάση "calm" προσέχει τα φωνητικά και "other" χαρακτηριστικά, με περιστασιακή σημασία του μπάσου και των ντραμς, απεικονίζοντας ένα ποικίλο μείγμα ήχου που βοηθά στην παραγωγή ενός καταπραϋντικού συναισθηματικού εφέ. Τέλος, παρατηρούμε ότι οι παρακείμενες κατηγορίες στον χάρτη συναισθημάτων συχνά θεωρούν παρόμοια χαρακτηριστικά ως σημαντικά, ιδιαίτερα όταν μοιράζονται παρόμοια επίπεδα σθένους ή ενέργειας.

Όσον αφορά τα πολυτροπικά μοντέλα, η δημιουργία συνολικών συνδυασμών τοπικών εξηγήσεων χρησιμοποιώντας την σημασία σταθμισμένης ομοιογένειας (homogeneity-weighted importance) δεν αποτυπώνει με ακρίβεια

την επιρροή του κάθε χαρακτηριστικού (feature). Αυτο συμβαίνει λόγω της ύπαρξης δύο τύπων χαρακτηριστικών: λέξεις και κομμάτια ήχου. Ενώ τα ηχητικά χαρακτηριστικά είναι πάντα τα ίδια 40, οι λέξεις διαφέρουν απο περίπτωση σε περίπτωση. Αυτό σημαίνει οτι τα φωνητικά, τα ντραμσ, τα μπάσα και τα "other" χαρακτηριστικά θα επηρεάσουν με παρόμοιο τρόπο κάθε τάξη αφού έχουν και διαφορετικό περιεχόμενο. Για παράδειγμα, ένα κομμάτι "other" χαρακτηριστικού θα μπορούσε να περιέχει power chords και να ωθεί το μοντέλο στην "heavy music" ή θα μπορούσε να περιέχει ήχους σαξοφώνου και επομένως να ωθεί το μοντέλο στην "rhyhthm music". Αυτό σημαίνει ότι τα χαρακτηριστικά ήχου θα έχουν υψηλή εντροπία αφού δηλαδή παραπέμπουν σε πολλαπλές κλάσεις. Από την άλλη πλευρά, οι λέξεις έχουν χαμηλότερη εντροπία καθώς ορισμένες λέξεις επηρεάζουν μόνο συγκεκριμένες κλάσεις. Αυτό οδηγεί στο να θεωρούνται εσφαλμένα οι λειτουργίες ήχου ως λιγότερο σημαντικές. Επομένως, η δημιουργία συνολικών συνδυασμών τοπικών εξηγήσεων χρησιμοποιώντας την μέση σημασία (average importance) είναι πιο κατάλληλη για την πολυτροπική περίπτωση I_{cj}^{AVG}.

Για να καταλάβουμε αν το πολυτροπικό μοντέλο που προβλέπει ετικέτες είδους καταφέρνει να συνδυάσει τις επιμέρους τροπικότητες αναλύουμε τα αποτελέσματα των μεθόδων επεξήγησης. Σε κατηγορίες όπου το λυρικό μοντέλο ήταν πιο ακριβές από το μοντέλο ήχου, τα πιο σημαντικά χαρακτηριστικά για την απόφαση του πολυτροπικού μοντέλου φαίνεται να είναι τα στιχουργικά χαρακτηριστικά, ενώ σε τάξεις που το λυρικό μοντέλο είχε χαμηλότερη απόδοση, το μοντέλο δίνει προσοχή στα χαρακτηριστικά ήχου. Για παράδειγμα, για την κατηγορία "hip hop", οι συνολικού συνδυασμοί, όπως παρουσιάζονται στο Σχήμα 1.3.9, δείχνουν ξεκάθαρα ότι τα στιχουργικά χαρακτηριστικά έχουν μεγαλύτερη επίδραση. Από την άλλη πλευρά, το μοντέλο εστιάζει περισσότερο σε χαρακτηριστικά ήχου για "punk" τραγούδια, όπως φαίνεται στο Σχήμα 1.3.10. Σε άλλες περιπτώσεις, όπως η κλάση "pop", της οποίας οι συνολικοί συνδυασμοί είναι στο Σχήμα 1.3.11, τα χαρακτηριστικά ήχου και στίχων επηρεάζουν εξίσου το μοντέλο. Γενικά, η πολυτροπική προσέγγιση καταφέρνει να ενσωματώσει και να συνδυάσει πληροφορίες από τα μονοτροπικά μοντέλα για βελτιωμένα αποτελέσματα.

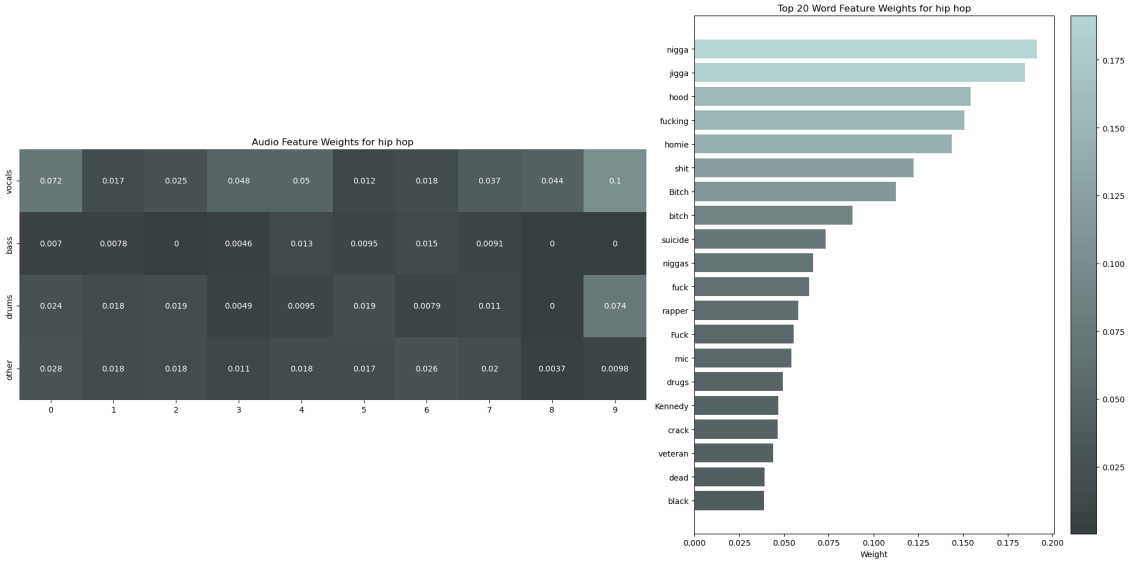


Figure 1.3.9: Οι συνολικοί συνδυασμοί τοπικών εξηγήσεων για τραγούδια ταξινομημένα ως "hip hop" απο το πολυτροπικό μοντέλο. Ο πρώτος χάρτης θερμότητας απεικονίζει τα βάρη των χαρακτηριστικών ήχου, ενώ το ραβδόγραμμα δείχνει τα βάρη των 20 πιο σημαντικών λέξεων.

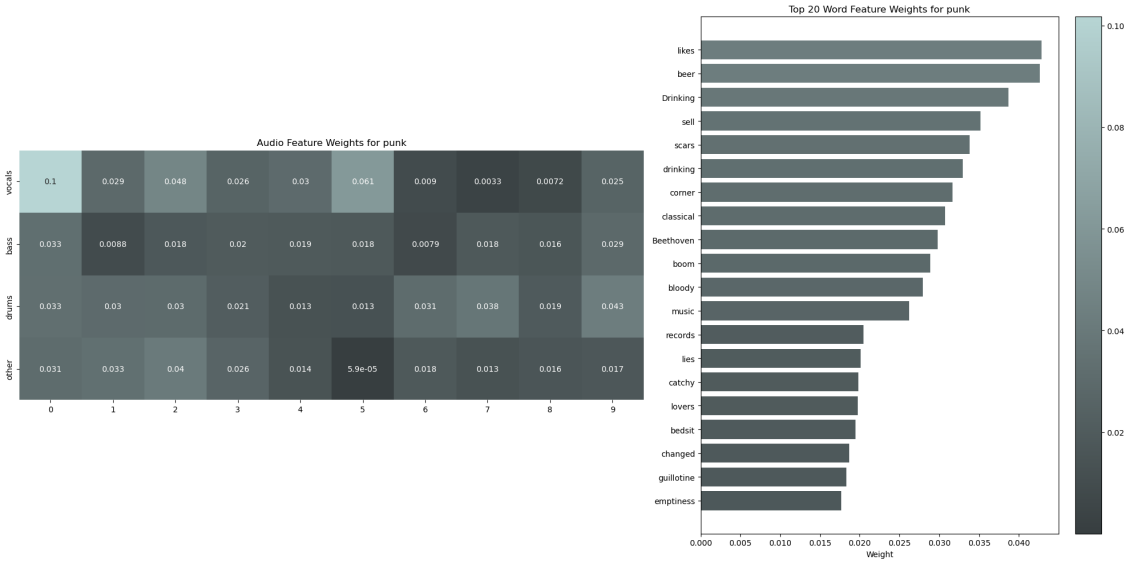


Figure 1.3.10: Οι συνολικοί συνδυασμοί τοπικών εξηγήσεων για τραγούδια ταξινομημένα ως "punk" απο το πολυτροπικό μοντέλο. Ο πρώτος χάρτης θερμότητας απεικονίζει τα βάρη των χαρακτηριστικών ήχου, ενώ το ραβδόγραμμα δείχνει τα βάρη των 20 πιο σημαντικών λέξεων.

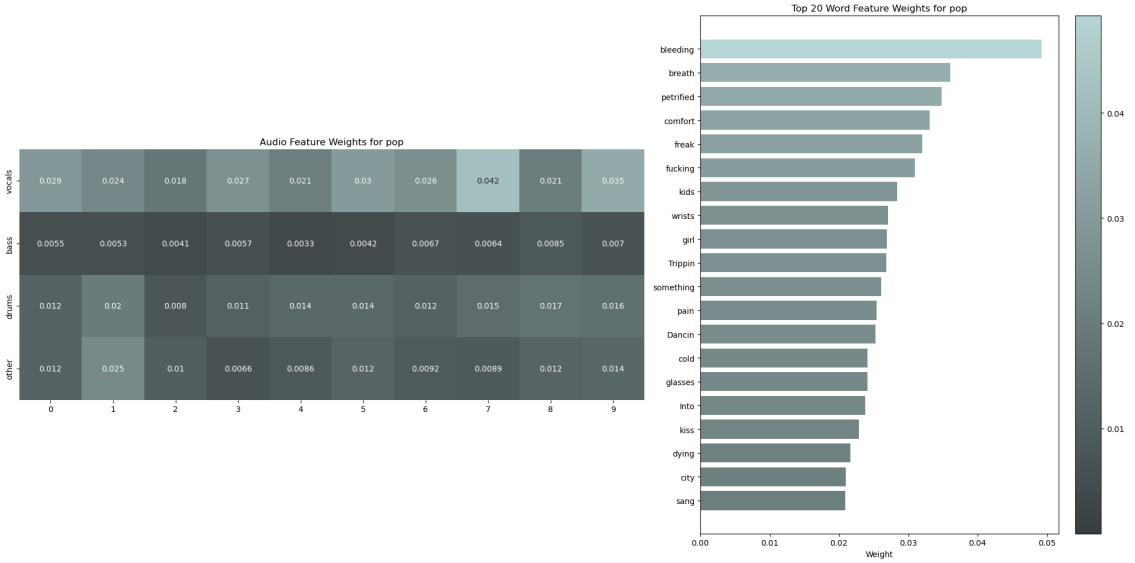


Figure 1.3.11: Οι συνολικοί συνδυασμοί τοπικών εξηγήσεων για τραγούδια ταξινομημένα ως "pop" απο το πολυτροπικό μοντέλο. Ο πρώτος χάρτης θερμότητας απεικονίζει τα βάρη των χαρακτηριστικών ήχου, ενώ το ραβδόγραμμα δείχνει τα βάρη των 20 πιο σημαντικών λέξεων.

Τέλος για το πολυτροπικό μοντέλο που προβλέπει συναισθήματα παρατηρούμε ότι τα ηχητικά χαρακτηριστικά είναι πιο σημαντικά. Αν και το μοντέλο αποδίδει καλύτερα από τις μονοτροπικές προσεγγίσεις, η διαφορά με το μοντέλου ήχου είναι οριακή. Επίσης βλέπουμε και απο τους συνολικούς συνδυασμούς τοπικών εξηγήσεων οτι το πολυτροπικό μοντέλο θεωρεί τα ηχητικά χαρακτηριστικά πιο μεγάλης επίδρασης για όλες τις κατηγορίες. Για τις "happy", "depressed", "excited", "tense", "calm" και "relaxed" τα 10 χαρακτηριστικά με το μεγαλύτερο βάρος είναι όλα ηχητικά. Στις "sad" και "neutral" μόνο λίγες λέξεις έχουν υψηλό βάρος. Για τα "angry" τραγούδια, φαίνεται ότι ο συνδυασμός χαρακτηριστικών κειμένου και ήχου να είναι εξίσου σημαντικά, με τα ηχητικά να είναι πιο σχετικά, χωρίς όμως η απόδοση του πολυτροπικού μοντέλου να είναι καλύτερη από το μοντέλο ήχου.

1.4 Συζήτηση

Σε αυτή την εργασία εξερευνήσαμε την πολυτροπική κατηγοριοποίηση μουσικής με βάση το είδος και το συναίσθημα. Η έρευνά μας ξεκινά με την διερεύνηση διάφορων υπαρχόντων συνόλων δεδομένων τα οποία προσφέρουν στίχους και ηχητικό περιεχόμενο καθώς και ετικέτες είδους και συναισθήματος. Αφού η βάση δεδομένων ονόματι Music4All καλύπτει τις ανάγκες μας, συνεχίζουμε αναλύοντας και αξιολογώντας τα περιεχόμενά της. Στην συνέχεια, προτείνουμε μεθόδους να τα ομαδοποιήσουμε σε 9 κατηγορίες συναισθήματος με βάση το μοντέλου κύκλου του Ράσελ και σε 9 κατηγορίες είδους. Επιπλέον παρουσιάζουμε τρόπους για επαύξηση των διαθέσιμων δεδομένων λόγω της υπάρχουσας ανισορροπίας σε ετικέτες συναισθήματος και να αποθορυβοποιήσουμε όσον αφορά τις ετικέτες είδους. Αν και καταβάλαμε έντονες προσπάθειες για την επιμέλεια του συνόλου δεδομένων αυτού, το τελικό μας σύνολο είναι και πάλι ανισόρροπο ως προς το συναίσθημα και περιέχει αμφίλογες ετικέτες είδους για μερικές καταχωρήσεις του. Αυτό οφείλεται στην δυσκολία να βρεθούν "χαλαρά" (relaxing) και "ήρεμα" (calm) τραγούδια μέσω της διεπαφής SpotifyAPI άλλα και στο γεγονός ότι πολλά τραγούδια εμπίπτουν σε περισσότερα από ένα είδη [60].

Συνεχίζουμε τις προσπάθειές μας μελετώντας σχετικές εργασίες πάνω στην ανάκτηση μουσικών πληροφοριών, ξεκινώντας από την παλινδρόμηση (regression) συναισθημάτων στην μουσική. Παρόλο που βρίσκουμε πρόσφατες τεχνολογίες αιχμής κατά την διαδικασία αυτή, επιλέγουμε να ακολουθήσουμε μια προσέγγιση μικρορύθμισης προεκπαιδευμένων μοντέλων μετασχηματιστή (tranformer), δημιουργώντας 3 μοντέλα για κάθε διαδικασία ταξινόμησης, ένα μοντέλο κειμένου που χρησιμοποιεί το roberta-large, ένα μοντέλο ήχου με φασματογράμματα ως είσοδο βασισμένο στο AudioSpectrogramTrsnformer και ένα πολυτροπικό μοντέλο συνδυάζοντας τα δύο προηγούμενα. Σύμφωνα με την έρευνά μας, είναι η πρώτη φορά που συνδυάζονται αυτά τα μοντέλα για ταξινόμηση μουσικής. Η τελευταία προσέγγιση ξεπερνά και τις δύο μονότροπες, ενώ το μοντέλο με είσοδο τους στίχους είχε την χειρότερη απόδοση. Γενικά, η ταξινόμηση σύμφωνα με το είδος ήταν πιο επιτυχημένη από την ταξινόμηση σύμφωνα με την ετικέτα συναισθήματος.

Για να κατανοήσουμε τη συμπεριφορά των μοντέλων μας, αναζητούμε μεθόδους επεξήγησης (explainability) στον τομέα κειμένου και ήχου/εικόνας. Διαπιστώνουμε ότι οι Τοπικά Ερμηνεύσιμες Ανεξαρτήτως Μοντέλου Επεξηγήσεις (Local Interpretable Model-Agnostic Explanations) ταιριάζουν στις ανάγκες μας. Επομένως, εφαρμόζουμε επεξηγήσεις LIME στο λυρικό μοντέλο, για να καταγράψουμε ποια θέματα είναι σχετικά για κάθε κλάση. Όσον αφορά το πεδίο του ήχου, αντί να αντιμετωπίζουμε το φασματογράφημα ως εικόνες και να δημιουργούμε επεξηγήσεις για αυτές, εφαρμόζουμε το audioLIME, μια προσέγγιση που βασίζεται στο LIME που διαχωρίζει τον ήχο σε χρονικά τμήματα και κάθε τμήμα σε φωνητικά, ντραμς, μπάσα και άλλα στοιχεία και μας παρέχει ακρόαση εξηγήσεις. Παρουσιάζουμε το MusicLime, έναν τρόπο συνδυασμού των δύο παραπάνω μεθόδων για παραγωγή πολυτροπικών τοπικών επεξηγήσεων με κείμενο και μουσική ως είσοδο. Τέλος, δημιουργούμε καθολικούς συνδυασμούς τοπικών επεξηγήσεων για να έχουμε μια πιο ολοκληρωμένη εικόνα.

Η ανάλυσή αυτή παράγει ενδιαφέροντα αποτελέσματα. Πρώτον, οι μέθοδοι επεξήγησης στίχων καταφέρνουν να συλλάβουν ορισμένα θέματα για κάθε κλάση, ιδιαίτερα αυτά με έντονο λεξιλόγιο, βωμολοχίες και θάνατο. Ωστόσο, η ταξινόμηση της μουσικής μόνο λαμβάνοντας υπόψη το περιεχόμενο των στίχων δεν είναι ιδιαίτερα επιτυχημένη, καθώς η μουσική είναι πολύπλευρη. Επιπλέον, οι ηχητικές επεξηγήσεις δείχνουν ότι το μοντέλο ήχου μπορεί να συλλάβει μουσικά στοιχεία ξεχωριστά για κάθε κλάση, όπως κραυγές που περιέχονται στα φωνητικά της «heavy music». Το μοντέλο που ταξινομεί με βάση το είδος που αξιοποιεί στίχους και ήχο κατάφερε να συνδυάσει και τις δύο μορφές και να επιτύχει βελτιωμένα αποτελέσματα. Για τον τομέα του συναισθήματος, δεδομένου ότι το λυρικό μοντέλο δεν ήταν πολύ ακριβές, το πολυτροπικό μοντέλο ακολουθεί πιστά την προσέγγιση που αξιοποιεί ήχο γεγονός που είναι επίσης εμφανής κατά τη φάση της εξήγησης. Το κύριο εμπόδιο για την καλύτερη απόδοση των μοντέλων φαίνεται να είναι η ασάφεια των ετικετών που υπάρχει και στις δύο εργασίες. Όπως αναγράφουν στην δουλειά τους [21] "ο προσδιορισμός ετικέτας σε ένα μουσικό απόσπασμα με βάση το συναίσθημα μπορεί να είναι μια ασαφής και αμφίθυμη άσκηση λόγω της υποκειμενικής φύσης της". Αν και οι ετικέτες είδους μπορεί να φαίνονται πιο διακριτές από τις ετικέτες συναισθημάτων, πολλά τραγούδια ενσωματώνουν στοιχεία από πολλά είδη και διαφορετικοί ειδικοί μπορεί να χρησιμοποιούν διαφορετικά κριτήρια για την ταξινόμηση του είδους [60].

1.5 Μελλοντικές Κατευθύνσεις

Τα αποτελέσματα αποκαλύπτουν επίσης πολλές διόδους για περαιτέρω έρευνα και πιθανές κατευθύνσεις που μπορούν να βελτιώσουν τη θεμελιώδη δουλειά που παρουσιάζεται εδώ. Αρχικά, προκειμένου να αντιμετωπιστεί το εμπόδιο της ασάφειας των ετικετών, η μελλοντική έρευνα θα μπορούσε να επικεντρωθεί στη δημιουργία ενός συνόλου δεδομένων με σχολιασμούς είδους από ειδικούς. Αν και αυτοί θα μπορούσαν επίσης να έχουν διαφωνίες ως προς το είδος σε πολλές περιπτώσεις, ένα τέτοιο σύνολο δεδομένων θα ήταν σαφώς βελτιωμένο σε σχέση με τα ήδη υπάρχοντα που περιέχουν ετικέτες απο ερασιτέχνες. Επίσης, δεδομένου ότι πολλά τραγούδια μπορεί να εμπίπτουν σε ετικέτες πολλαπλών ειδών, για παράδειγμα να είναι και "pop" και "electronic", ένα λογικό επόμενο βήμα θα μπορούσε να είναι η εφαρμογή των μοντέλων αυτής της μελέτης σε ένα περιβάλλον πολλαπλών ετικετών. Επιπλέον, στην εργασία μας τις αναπαραστάσεις χώρου (embeddings) κάθε τροπικότητας και στη συνέχεια χρησιμοποιούμε μια κεφαλή ταξινόμησης για να κάνουμε προβλέψεις. Προκειμένου να βελτιωθεί η απόδοση, οι μελλοντικές προσπάθειες θα μπορούσαν να μελετήσουν διαφορετικούς τρόπους συνδυασμού των τροπικοτήτων (π.χ. πολλαπλασιασμός ή προσθήκη των αναπαραστάσεων μαζί). Τέλος, τα μοντέλα που χρησιμοποιούνται είναι προεκπαιδευμένα σε δεδομένα εκτός του τομέα μουσικής (π.χ. κείμενα στην wikipedia ή φασματογράμματα απο βίντεο στο youtube). Ένα πολύγλωσσο μοντέλο όπως το BERT προεκπαιδευμένο σε στίχους και ίσως ποίηση και μια αρχιτεκτονική transformer φασματογράμματος προεκπαιδευμένη σε μουσικά κομμάτια θα μπορούσε να δώσει ενδιαφέροντα αποτελέσματα.

Αν και οι μέθοδοι επεξήγησης σε αυτή τη μελέτη έδωσαν επαρκή αποτελέσματα, πρόσθετες μελέτες θα μπορούσαν να αποδειχθούν χρήσιμες. Η δημιουργία τοπικών επεξηγήσεων με το LIME απαιτεί τον ορισμό της μεταβλητής του αριθμού των δειγμάτων. Αυτή η μεταβλητή ελέγχει τον αριθμό των δειγμάτων κοντά στο αρχικό στιγμιότυπο που πρόκειται να δημιουργηθούν. Μέχρι στιγμής, ένας εξαντλητικός αριθμός δειγμάτων είναι απαγορευτικός για μεγάλα μοντέλα με πολλά χαρακτηριστικά εισόδου. Οι επόμενες μελέτες θα μπορούσαν να διερευνήσουν τον επαρκή αριθμό δειγμάτων ή συγκεκριμένους τρόπους για να διαταραχθεί η είσοδος, ώστε οι τοπικές εξηγήσεις να είναι όσο το δυνατόν πιο ακριβείς. Επιπλέον, μπορει να αποδειχθεί καρποφόρο αν οι προσπάθειες επεξήγησης δημιουργούν εξηγήσεις με βάση λυρικές γραμμές αντί για μεμονωμένες λέξεις.

Chapter 2

Introduction

In the dynamic intersection of music and artificial intelligence (AI), music information retrieval (MIR) tasks have emerged as vibrant areas of research. Music classification, music generation, music source separation and other such tasks encompass a wide range of applications and have significantly impacted how we interact with, consume, understand and create music. With the growing volume of music data and the increasing reliance on Deep Learning (DL) models, there is a critical need for interpretability and explainability in these MIR systems.

Our motivation for this thesis is driven by two distinct yet equally compelling aspects of music: its profound emotional implications and its diverse genre categories. The capacity of music to evoke a wide range of emotions and affect human experience deeply underscores the necessity to explore and understand its emotional dimensions. Separately, the rich diversity within music genres, each representing unique stylistic and cultural characteristics, presents a separate challenge that merits detailed investigation. Additionally, the inherent multimodality of music, incorporating both lyrics and audio, adds a layer of complexity to its analysis. This multimodal nature necessitates a sophisticated approach to explain and interpret how these different elements contribute individually and collectively to the experience of music.

This thesis embarks on an exploratory journey into these domains, focusing on the intersection of music classification and regression tasks, multimodality, and Explainable Artificial Intelligence (XAI). Firstly we explore available datasets and methods for obtaining multimodal data, in addition to curating the Music4all[54] dataset, which will be utilized throughout our subsequent experiments. We proceed by examining how incorporating data of different modalities can impact the performance of music mood regression models. Leveraging insights gained in the previous steps, we implemented mood and genre classification using audio, lyrics and their combination as inputs. Lastly, we delve into methods to gain insights into the models' decision-making processes. We introduce MusicLime, a methodology to explain multimodal music classification and regression tasks based on Local Interpretable Model-agnostic Explanations [51] (LIME), offering a novel approach to understanding how different modalities contribute to model decisions in the context of music.

In order to provide a foundation for understanding the motivation and results of our research efforts we structure this thesis as follows:

- Firstly, we begin with an exploration of the background. We provide the theoretical underpinnings of key concepts in machine learning that are foundational to our work as well as a thorough review of related work on music datasets, multimodality, and explainability. In this manner we set the stage for how our research builds upon these existing frameworks.
- We continue with detailing our methodology. This encompasses the meticulous curation of the chosen dataset, the models employed for our regression and classification tasks and the explainability methods for interpreting the models' predictions, gaining insights into the models' functioning.
- In this manner, we set the stage to present our findings in the final section of our work. This encompasses the resulting modified dataset, the outcomes of our models and select examples from the explainability

method employed to dissect and analyze the workings of our models.

Chapter 3

Background

This chapter serves as the backbone for our work. We provide a thorough introduction to the core concepts and algorithms of Machine Learning (ML) and in particular Deep Learning (DL) that are pivotal to our thesis. Focusing specifically on the DL techniques, Explainability (XAI) methods and music genre and emotion concepts utilized in our study, we aim to equip the reader with information required for a comprehensive understanding of the processes followed in our research. Additionally, we dive into an exploration of music datasets, with a particular emphasis on those that are multimodal and contain metadata regarding multiple tasks. This investigation is critical, as it highlights the challenges and opportunities inherent in working with complex, multimodal musical information. Furthermore, we present related work in the domains of music regression and classification, multimodality, and explainability, examining how these studies have influenced and shaped our own research trajectory. This discussion is intended to contextualize our work within the broader academic landscape, demonstrating the contributions of our research for the MIR field.

3.1 Machine Learning and Music Concepts

3.1.1 Machine Learning and Deep Learning

Machine Learning (ML) is a branch of artificial intelligence that focuses on developing algorithms and statistical models that enable computers to perform tasks without following explicit instructions. In general, ML models are used to make predictions, based on some input data. An error function is utilized to evaluate the prediction of the model. The model is optimized by adjusting it's weights to reduce discrepancy between data points of the training set and model predictions. This process can be repeated until certain conditions are met (e.g. the accuracy surpasses a certain threshold value). Machine learning can be categorized into supervised, unsupervised and semi-supervised. Supervised ML uses labeled datasets to train respective models, whereas unsupervised ML models detect patterns and groupings in the data without the need of human supervision. Semi supervised models combine the previous two categories.

Deep learning is a subset of machine learning. It utilizes neural networks with many layers (hence "deep") to approximate the way human brains operate and to accurately recognize, classify and describe objects within the input data. This methodology has been applied successfully in fields such as computer vision, natural language processing and speech recognition, in contrast to traditional algorithms that struggled to perform well.

3.1.2 Transformers

Transformers are a type of deep learning architecture that revolutionized natural language processing (NLP). Presented in the paper "Attention Is All you Need" [61], they don't rely on a sequential approach, but can analyze all parts of a sentence simultaneously. At the heart of a transformer lies its encoder-decoder architecture. The encoder's self-attention sublayer analyzes relationships between words in the input sequence, while the feed-forward network captures more intricate patterns. Residual connections and layer normalization aid in efficient training. The decoder's masked self-attention prevents information peeking during training, and encoder-decoder attention allows it to consider the encoded input when generating the output. This layered interplay has the advantage of requiring less training time than previous architectures setting transformers at the forefront of NLP.

Transformers success has even led to the development of Vision Transformers (ViT). Bringing similar power to image analysis, ViTs treat images similarly to sequences of words. This idea was introduced in the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" [18]. Vits work by segmenting images into fixed-size patches, processing these patches as tokens analogous to textual data in NLP. This method allows the transformer to apply its self-attention mechanism directly to the patches, capturing complex spatial hierarchies and dependencies between different parts of an image. By training on large datasets and leveraging transfer learning, ViTs have demonstrated competitive performance with state-of-the-art convolutional networks, (CNN) marking a significant shift in how machine learning models perceive and understand visual data.

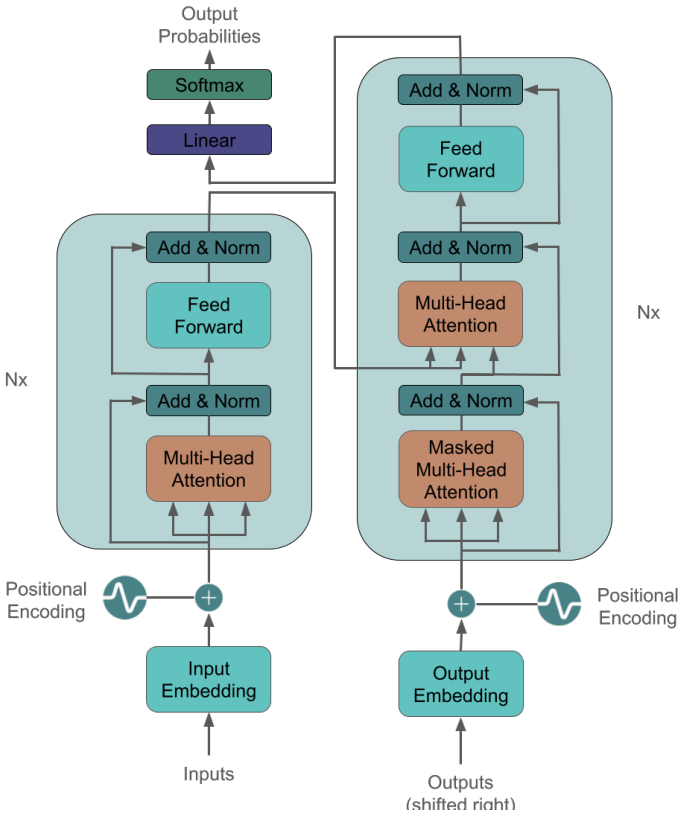


Figure 3.1.1: The transformer architecture as presented in the paper "Attention Is All You Need" [61].

3.1.3 Multimodality

Modality refers to the way something happens or is experienced and is usually connected with the human senses [7]. Multimodality in the context of machine learning and data analysis refers to the integration and processing of information from multiple types of data or sources such as text, images, audio in datasets or algorithms. As mentioned in the paper [47] the types of multimodality encompass several key variations, including multimodal input, where multiple data forms are used to feed into models; multimodal output, where models generate information across different formats; translation from one modality to another, exemplifying the transformation of data from one form to a comprehensible counterpart in another; and the combination of different modalities into a unified representation, illustrating the integration of diverse data streams.

In music information retrieval, integrating different modalities such as lyrical content with audio features can lead to more nuanced emotion recognition and genre classification, underscoring the practical benefits of multimodal learning. While lyrical information is inherently included in the audio domain, the analysis of audio is more demanding computationally and often restricted by the temporal limits of the models. Typically, an audio model might only process up to 30 seconds of a song, while textual analysis can cover the lyrics of potentially the entire song. This disparity highlights the practical limitations of current audio processing technologies, which struggle to handle the full duration and complexity of audio data within resource constraints. Furthermore, most datasets are more likely to provide complete lyrics than extensive audio recordings, as collecting and storing high-quality audio data is more challenging than textual data. This availability bias further complicates the development of robust audio models, making the integration of lyrical text not only beneficial but sometimes necessary to enhance the model's understanding and performance in music-related tasks.

3.1.4 Explainability

The rapid development of AI systems in the last decade has revolutionized numerous fields, however, as AI systems become increasingly integrated into critical decision-making processes, concerns about their trans-

parency and interpretability have surfaced. Explainable AI (XAI) has emerged as a pivotal area of research aimed at addressing these concerns by enhancing the transparency of AI models and algorithms. XAI seeks to provide insights into how AI systems arrive at decisions, making their outputs understandable to humans. The challenge of explainability is not purely technical; it incorporates insights from social sciences [45, 42] to ensure that explanations are meaningful and useful to diverse stakeholders, including end-users, domain experts, and regulators. Explainability in AI encompasses a variety of methods and techniques, each serving different aspects of transparency and understanding, addressing the diverse and sometimes conflicting objectives of XAI [36, 40]. With the multitude of available methods, the evaluation of XAI approaches is crucial and remains an active field of research to ensure their effectiveness and reliability in various contexts [20]. Understanding the capabilities of an XAI method is crucial for its adoption and success. Some of the differentiating characteristics of these XAI methods and techniques that allow us to better comprehend them are the following [52].

- **Aim of explanation**: The aim of the explanation can be categorized into introspection and justification-related. Introspection involves dissecting the model's internal processes and characteristics to gain a deeper understanding of its functioning and potential biases. Justification, however, focuses on elucidating the model's outputs in a manner that justifies its decisions to the end-users.
- **Exclusiveness of explanations**: They can be viewed through the lens of local versus global explanations. Local explainability zeroes in on specific instances, offering detailed insights into the decision-making process for individual predictions, whereas global explainability seeks to unravel the model's overall logic and behavior, providing a broader understanding of how the model operates across all instances.
- **Model dependency**: Explainability methods can be divided into model-specific and model-agnostic. Model-specific methods are tailored to the intricacies of a particular model type, utilizing its internal mechanics to elucidate how decisions are made. On the other hand, model-agnostic approaches are designed to be universally applicable, providing explanations regardless of whether there is access to the model's internal architecture.
- **Timing of applied explainability**: Finally, explainability methods are categorized based on when they are applied, distinguishing between post-hoc and ante-hoc approaches. Post-hoc explainability is employed after the model has been trained, vital for complex models where intrinsic interpretability is challenging. In contrast, ante-hoc (or intrinsic) explainability entails baking explainability into a model from the beginning (i.e. decision trees).

3.1.5 Music Emotion and Genre

Throughout the years, music has undeniably been a medium for expressing and evoking emotions. From the melancholic strains of a minor key melody to the joyous energy of a fast-paced rhythm, music can convey and trigger a wide range of feelings. Studies indicate the importance of distinguishing between perceived and induced emotion in music creations. Perceived emotion means the emotion conveyed by the music itself while induced emotion is the emotion that the music provokes among the audience. Researches have used various emotional models to describe music emotion more accurately [24]. One of the earliest such works is Hevner's affective ring, that uses 66 emotion adjectives, arranging them into 8 categories [27]. The first emotion model designed for music-induced emotion is the Geneva Emotional Music Scales (GEMS), which involves an initial set of 45 labels that can be grouped into nine different dimensions and three higher-order factors [58]. However categorical emotion models have been criticised by scholars so dimensional emotion models are used more recently. Works by Russel and Thayer organize mood descriptors into low-dimensional models. In particular, most commonly used in MER, Russell's circumplex model, depicts emotions in a two-dimensional space, where valence (ranging from pleasant to unpleasant) and arousal (ranging from calm to excited) act as axes. The model is depicted in figure 3.1.2 adapted from this article [55]. Other works include a third dimension in such models or implement ranking, probability distributions and pairs of antonyms to express music emotions [24].

Music genre categorizes music based on shared characteristics like instrumentation, tempo, rhythm, and cultural context. For example Hip Hop music is characterized by rap patterns and contains themes such as police brutality and oppression, while electronic music can be identified by the use of synthesizers. Although

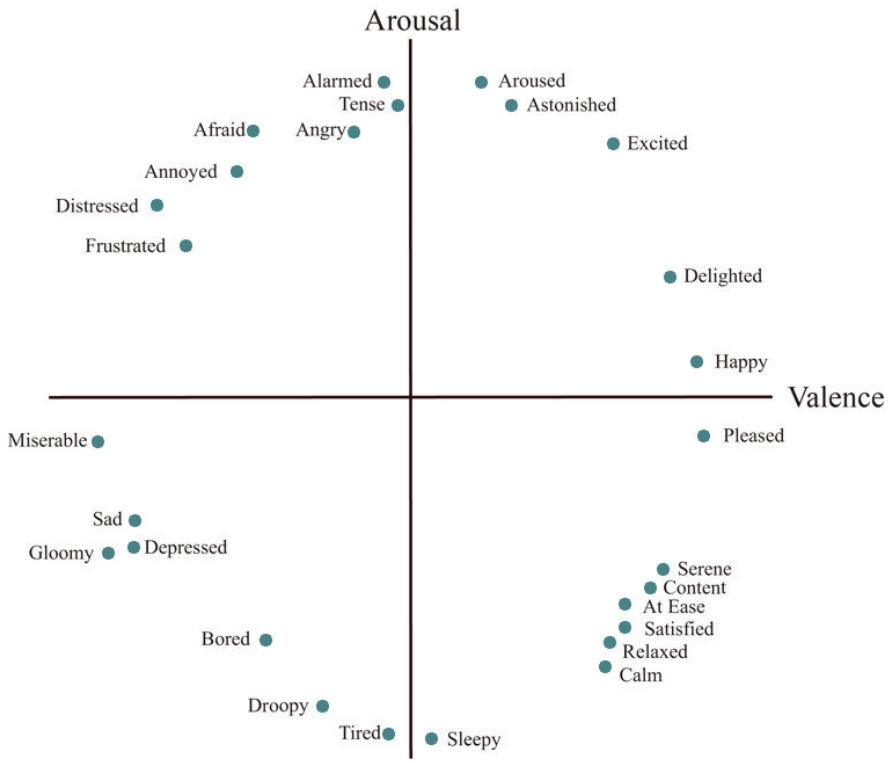


Figure 3.1.2: Russell's circumplex model of emotion (adapted from [55]).

these groupings help listeners navigate the vast musical landscape such a categorization is often subject to different human experiences and sometimes is overlapping. For a comprehensive exploration of music genres, their interrelationships, history and characteristics MusicMap.info serves as a valuable resource [32]. The facade of the interactive diagram created by the authors of the resource can be seen in 3.1.3. Genres that are similar are placed adjacently while the vertical axis gives us an chronological estimate of the genres existence.

3.2 Datasets

In the dynamic field of music information retrieval (MIR), the integration of multiple data modalities offers a nuanced perspective, ideal for understanding and analyzing musical content. The primary objective of this section is to illustrate the process undertaken in selecting an appropriate multimodal dataset, pivotal for advancing our MIR tasks focused on regression and classification. We delve into the criteria that guided our dataset selection .

The MIR field has witnessed significant growth, leading to the development of a diverse range of specialized datasets catering to various research needs. These datasets, each uniquely crafted, offer a wealth of resources for advancing MIR studies. They vary widely in their contents, ranging from audio features and metadata to more complex data types like annotations and cultural context. For anyone looking to explore the breadth of available MIR datasets, a comprehensive list with short descriptions of their contents can be found at the International Society for Music Information Retrieval (ISMIR) website website[29]. This resource serves as a valuable guide for selecting the most suitable dataset for specific MIR objectives.

In our research, we concentrated on selecting datasets that offer a variety of modalities for music songs. This included datasets providing multiple modalities inherently, such as audio, lyrics, and MIDI, as well as those with a single modality that could be augmented to become multimodal. For example, datasets initially offering only audio were considered with the potential of fetching additional data like lyrics to enrich the modality. Additionally it would be ideal if datasets provided metadata supporting a range of analytical tasks or facilitated the easy acquisition or scraping of additional metadata, thereby transforming them into versatile resources for multitask research endeavors. Specifically we focused on datasets providing lyrics and

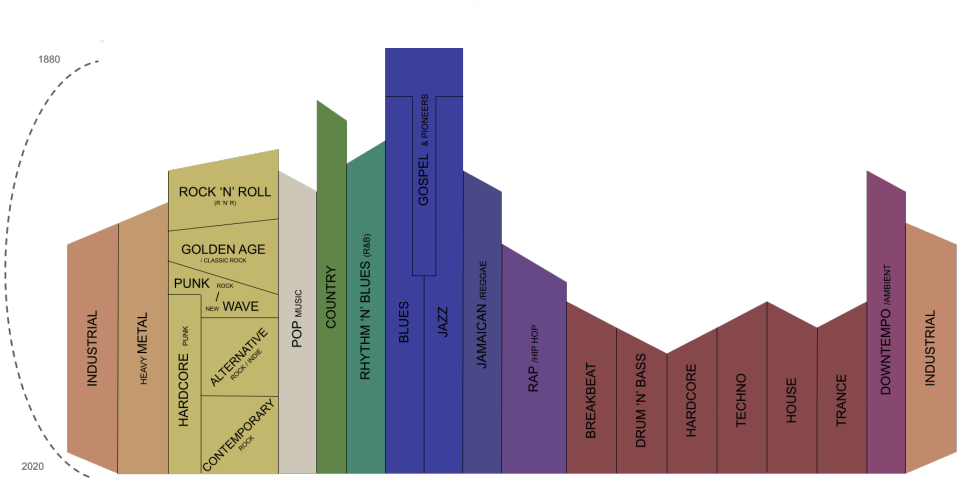


Figure 3.1.3: The facade of the interactive diagram found at [32]

audio along with genre and emotion labels.

A significant challenge we encountered in this selection process was the constraints posed by copyright restrictions, which limited the availability of suitable datasets. That is, most of the available datasets do not include raw audio files but rather provide processed audio features, such as Mel Frequency Cepstral Coefficients (MFCCs). Despite these limitations, we identified several candidate datasets that aligned with our criteria.

- The **Million Song Dataset** (MSD) [9] is a freely available collection of features and metadata for one million contemporary popular music tracks. It provides information about the songs themselves, including audio characteristics extracted through analysis and details like artist and genre. This dataset serves as a valuable resource for researchers in the field of MIR. However, it's important to note that while the dataset offers a wealth of information about the music, it does not include the actual audio recordings themselves.
- The **GTZAN** dataset, since its introduction in 2002, has become an extensively-utilized public dataset for music genre recognition (MGR) within the machine listening research community. Comprising 1,000 half-minute excerpts across ten genres, GTZAN's widespread usage has not been without criticism due to its lack of metadata, repetitions, mislabelings, and distortions as presented by Sturm[59]. Influenced by these alongside the difficulty in acquiring additional metadata for the tracks we decide not to utilize GTZAN.
- The **Free Music Archive** (FMA)[8] is an expansive dataset designed for evaluating several tasks in MIR, featuring over 106,574 tracks from various artists and genres, all under Creative Commons licenses. It offers a hierarchical taxonomy of 161 genres and provides both full-length audio and pre-computed features alongside rich metadata, including track, album, and artist details. However, we did not select the FMA for our research primarily due to the obscurity of its tracks and the absence of lyrics, which are essential for our multi-modal analysis. The challenge of sourcing publicly available lyrics for these less-known tracks further complicated their potential use in our study.
- The **Database for Emotional Analysis of Music** (DEAM) [6] dataset is notable for its dynamic annotations of music, capturing changes in emotional content over time. It includes valence and arousal annotations for 1,802 songs at a 2Hz time resolution, which allows for fine-grained analysis of emotional trajectories within songs. The dataset's contributions are particularly valued for research on dynamic music emotion recognition, enabling the exploration of how emotions evolve throughout a piece of music.
- The **MTG-Jamendo** dataset offers music and corresponding tags for genre, mood, and instruments. Sourced from royalty-free Jamendo music, it includes over 55,000 tracks with labels in various categories. This dataset is valuable for researchers in Music MIR tasks but the audio is provided (in MP3 format), it's important to note that this dataset as well doesn't contain the lyrics.

- The **Music4All** database,includes metadata, tags, genre information, 30-second audio clips, and lyrics, providing a robust foundation for various MIR tasks such as music recommendation, genre classification, and mood classification. With its extensive collection of over 100,000 tracks enriched with detailed metadata and user-generated tags, Music4All stands out for its diversity of musical genres and depth of data. Due to its rich multimodal offerings and the relative ease of data access compared to other datasets, we chose Music4All as the primary dataset for our research endeavors, particularly valuing its potential for supporting various multitask possibilities. We will describe in more detail in later chapters.

3.3 Related Work

Understanding the vast landscape of MIR research is crucial for our work. This section delves into key areas: (1) existing approaches for genre classification and emotion recognition using either one or multiple modalities, (2) the potential of transformers for music and text analysis, and (3) the importance of explainability in machine learning models. By examining these areas, we gain valuable insights and pave the way for our own contribution to MIR.

3.3.1 Unimodal and Multimodal MIR

Unimodal techniques are a crucial foundation for MIR tasks, offering valuable insights and serving as a basis for more complex multimodal methods. The study by Jacek Grekow[23] investigates the application of RNNs for detecting emotions in music segments. Leveraging the Russell's circumplex model, the research explores the prediction of continuous values of emotions such as arousal and valence through trained regression models. By extracting audio features and creating sequential data for learning networks with LSTM units, the paper illustrates the utility of RNNs over traditional methods. Various experiments conducted with data featuring different sets of features and segmentation approaches underscore the importance of data division into sequences and the efficacy of recurrent networks in recognizing emotions in music. Moreover, the study highlights the significant gains achieved by employing pretrained models for processing audio features before training the RNN, demonstrating the advantages of this method in enhancing model performance for both arousal and valence prediction. Agrawa et al. [1] use a transformer-based model employing XLNet (xlnet-base-cased) to identify emotinal connotations of music based on lyrics. The inclusion of lyrics as a primary source for emotion recognition addresses a gap in MIR, where lyrics have been underappreciated despite their strong emotional cues. Their approach outperforms the existing methodologies available at the time of their study. Finally, Dervakos et al.[16] explore the application of CNNs for genre recognition in symbolic music representations like MIDI. The study introduces the Multiple Sequence Resolution Network (MuSeReNet), a CNN architecture tailored for symbolic music data, focusing on optimizing network depth, width, and kernel sizes while maintaining constant trainable parameters and receptive field sizes. The MuSeReNet processes MIDI data at multiple resolutions, improving genre recognition performance on the topMAGD and MASD datasets beyond state-of-the-art methods. The paper also ventures into the domain of XAI.

Research in multimodal MIR has explored various approaches that leverage different information sources beyond single modalities, demonstrating the effectiveness of combining complementary data streams for improved mood classification and genre prediction. Introduced in 2007 the paper[46] presents a method for music genre classification that integrates both textual and audio features using machine learning techniques. Specifically, it employs lyric text analysis alongside audio signal processing to extract features relevant to genre. These features are then combined and fed into classifiers to determine the music's genre. The study showcases how the fusion of text and audio data can capture a broader range of characteristics inherent to different music genres, leading to improved classification accuracy over approaches that rely on a single data type. In 2013, Panda et al.[47] introduces a multi-modal approach to MER, leveraging audio, MIDI, and lyrics information to classify music into emotional clusters. They developed an automatic method to create a multi-modal music emotion dataset using the AllMusic database, organized into five emotion clusters as defined in the MIREX Mood Classification Task. From audio data, 177 standard and 98 melodic features were extracted, while MIDI files contributed 320 features, and lyrics analysis yielded 26 features. The research demonstrates the effectiveness of combining these multi-modal features for MER, showcasing a significant improvement in classification accuracy. Specifically, using only standard audio features achieved an F-measure

performance of 44.3%, whereas the multi-modal approach enhanced results to 61.1%, using a subset of 19 multi-modal features.

In more recetnt work, Delbouys et al.[15] explore, in 2018, multimodal music mood prediction utilizing both the audio signal and lyrics, employing deep learning techniques.The authors developed and tested a bimodal deep learning model against classical models on a dataset of 18,000 songs, finding that their deep learning model outperformed traditional models in arousal prediction and matched their performance in valence prediction. Their audio-focused model utilizes mel-spectrograms with 40 filterbanks and 1024 time frames, processed through two convolutional layers. On the other hand, their lyric analysis model employs Word2Vec embeddings—trained on 1.6 million lyrics and incorporates a convolutional and LSTM layer to effectively capture the semantic nuances of the lyrics. An important aspect of their work is the examination of modality fusion strategies, where they identified that mid-level fusion significantly improves valence prediction accuracy. In 2022, Pyrovolakis et al.[49] focuses on detecting music mood through leveraging both song lyrics and audio signals. They employed deep learning architectures including CNNs for audio features, and transformers (BERT) for lyrics, to classify songs into four mood categories: happy, angry, sad, relaxed. Their study uniquely combines audio and lyrics data, previously processed separately, to enhance mood detection accuracy. The models were trained and evaluated on the MoodyLyrics[12] dataset, which contains 2000 song titles with mood annotations. The multi-modal approach significantly outperformed single-modal models, demonstrating that combining lyrics and audio information provides a more accurate representation of a song's emotional content. In 2021 Pandeya et al.[48] present a novel approach to classifying music by genre and emotion, with multiple labels, through a DNN architecture that processes both music and lyrics information. Their approach categorizes music into 44 coarse and 255 fine categories. By utilizing a novel convolutional technique that separates channel and filter convolutions, their model achieves high accuracy and low computational costs. The system was tested on the Music4All dataset, showing that the integration of audio and lyrics significantly improves the classification and regression tasks. Finally, in 2023 the paper [31] employs 11 high-level audio features retrieved from the Spotify API, including valence and energy, in conjunction with lyrics features like sentiment, TF-IDF, and Anew to predict valence and arousal scores from the Deezer Mood Detection Dataset (DMDD). The methodology involves regression models to assess the predictive power of these features, with a notable emphasis on the multi-modal approach combining both audio and lyrics. The findings underscore that including multi-modal features, especially those involving Spotify's valence and energy, enhances the prediction accuracy for valence compared to utilizing audio features alone.

MuLaN establishes itself as a state-of-the-art method for music emotion recognition. As detailed in the paper [28], MuLan represents an advanced approach in the field of music emotion recognition, positioning it near the pinnacle of current state-of-the-art methodologies. This model merges music recordings and weakly-associated, free from text annotations into a unified embedding space. It takes the form of a two-tower, joint audio-text embedding, trained utilizing contrastive learning to optimize the shared audio-text embedding space. The audio tower processes log mel spectrograms of music, while the text tower processes tokenized text sequences, each producing embeddings that are aligned through the learning process.

A comprehensive overview of multimodal machine learning can be found in the review by Baltrušaitis et al.[7]. The survey categorizes the challenges in multimodal learning into five core areas: representation, translation, alignment, fusion, and co-learning. It critically evaluates recent advances in each category, introduces a new taxonomy to organize the field's developments and identifies future research directions.

3.3.2 Text and Audio Transformers

As mentioned in a previous section the work "Attention Is All You Need" by Ashish Vaswani and colleagues [61] introduces the Transformer model, a novel architecture that significantly diverges from previous sequence transduction models by relying entirely on attention mechanisms. The Bidirectional Encoder Representations from Transformers (BERT) [17] utilizes the encoder from the transformer architecture. This model uses a "masked language model" (MLM) pretraining objective. That is it randomly replaces a percentage of the input tokens (either with the [MASK] token, with another random token or with the same token) and trains itself to predict these tokens. Additionally, BERT also incorporates "Next Sentence Prediction" (NSP), a training task that decides weather given two sentences, the second one is the next sentence of the first one in the original document. BERT is pretrained on large corpora and can be fine-tuned to achieve impressive

results in a variety of NLP tasks.

Following this work RoBERTa[38] (A Robustly Optimized BERT Approach) builds on BERT's language representations by implementing key methodological changes in the pre-training process, which significantly improves performance across various benchmarks. Modifications include training the model longer, with more data, on bigger batches, and without the next sentence prediction objective, resulting in improved model performance and efficiency. Additionaly Generative Pretrained Transformer[11] (GPT) employs the transformer-based architecture optimized for natural language understanding and generation. It leverages unsupervised pre-training on a large corpus of text focused on predicting the next word given all previous words. Followed by fine-tuning on specific tasks, GPT is able to understand and generate human-like text, making it highly effective for a wide range of language tasks. Finaly, Text-to-Text Transfer Transformer[50] (T5) reframes all NLP tasks as a text-to-text problem, where both the input and output are text strings. By using a unified approach, T5 simplifies the processing pipeline for various tasks, including translation, question answering, and summarization. This model is pre-trained on a diverse text corpus using a denoising objective to predict the masked span of text, which is then fine-tuned on task-specific datasets. Given the characteristics and performance of these models for our lyrical modality, we chose RoBERTa due to its more lightweight nature.

Transformers have also extended their success in the audio domain. The Audio Spectrogram Transformer (AST) [22] is a pioneering example, being the first model to apply a purely attention-based mechanism, without convolutional layers, directly to audio spectrograms for classification tasks. AST is distinctively pretrained on the ImageNet dataset, employing a cross-modality transfer learning strategy where vision transformer weights are adapted for audio by averaging the weights for the three input channels to suit the single-channel audio spectrograms. In its operation, AST converts audio into log Mel spectrograms, which are then divided into overlapping 16x16 patches. These patches are each transformed into embeddings with added positional embeddings to maintain sequence information. AST has proven highly effective, setting new performance benchmarks in various audio datasets. Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection (HTS-AT) [13] addresses the scalability challenges of audio transformers by implementing a hierarchical structure that significantly reduces model size and training duration. Incorporating a token-semantic module, HTS-AT not only excels in audio classification tasks, achieving new state-of-the-art results on AudioSet and ESC-50 and matching top performance on Speech Command V2, but also introduces the capability for sound event detection. However, the Causal Audio Transformer[37] (CAT) builds upon this success, introducing a specialized approach for audio classification by leveraging a Multi-Resolution Multi-Feature (MRMF) feature extraction and an acoustic attention block, designed to optimize audio signal processing. CAT integrates a causal module aimed at enhancing model generalizability, interpretability, and reducing overfitting through the use of counterfactual reasoning. This design allows CAT to achieve or surpass state-of-the-art classification performance across several datasets, including ESC50, AudioSet, and UrbanSound8K. While both HTS-AT and CAT offer impressive capabilities, we opted for the AST for its ease of use for our audio modality.

We should finally mention the Contrastive Language-Audio Pretraining model (CLAP) [19], a multimodal and foundational model for music information. CLAP learns audio concepts from natural language supervision, by employing two encoders and contrastive learning to connect language and audio, creating a joint multimodal space. This approach enables Zero-Shot predictions, meaning it can predict without having been explicitly trained on specific class labels, and generalizes across multiple domains and tasks. The model is trained with 128k audio-text pairs and tested on 16 downstream tasks, demonstrating significant improvements in classification accuracy and flexibility in class prediction at inference time, especially in Zero-Shot setups.

3.3.3 Explainability

Model-agnostic explainability techniques offer a versatile approach to understanding predictions from various complex models. Local Interpretable Model-Agnostic Explanations[51] (LIME), is a seminal work that introduced an explanation technique aimed at making the predictions of machine learning models understandable to humans by approximating the model locally with an interpretable model. This method addresses the challenge of machine learning models acting as black boxes, where the reasons behind their predictions are not clear. Additionally, the paper[62] "RELAX: Representation Learning Explainability" introduces RELAX, a framework aimed at providing attribution-based explanations for representations learned through self-

supervised learning, delivering superior explanations compared to gradient-based baselines. Recent works, address the problem of explainability through the utilization of Knowledge Graphs [41, 33, 34, 44], offering a more structured approach to black-box explanations.

Another approach is contrastive explanations and adversarial examples that shed light on the decision-making of natural language processing models. In particular, Minimal Contrastive Editing[53] (MiCE) is introduced as a novel method for generating contrastive explanations of model predictions through minimal edits to input text that result in a specified output change. This approach addresses the gap in existing NLP model explanation methods by leveraging human-like contrastive explanations that are also minimal, focusing on why a specific prediction occurred instead of another. MICE's effectiveness is demonstrated across various tasks including sentiment analysis, topic classification, and question answering. Furthermore, TextFooler[30] is a simple but effective method designed to generate adversarial text. It demonstrates its strength in attacking models across two fundamental natural language processing tasks: text classification and textual entailment. TextFooler successfully attacks target models showcasing three primary advantages: effectiveness in outperforming previous attacks by success rate and perturbation rate, utility preservation by maintaining semantic content, grammaticality, and correct classification by humans, and efficiency by generating adversarial text with computational complexity linear to the text length.

Building upon LIME's framework, audioLIME and CoughLIME provide interpretable and listenable explanations for audio predictions. In particular, audioLIME[25] offers interpretable, listenable explanations for MIR systems by utilizing source separation to create perturbations, addressing a unique aspect of audio data that is overlooked by traditional spectrogram-based methods. On the other hand, CoughLIME[64] is designed to provide sonified explanations for the predictions made by COVID-19 cough classifiers. It also adapts LIME for audio data, specifically focusing on cough sounds associated with COVID-19, by decomposing audio into interpretable components. For a deeper dive into explainable artificial intelligence for audio tasks, refer to the review [4]

Researchers have made significant strides in explainability for MIR, with various techniques offering insights into how models arrive at their decisions. Lyberatos et al.[39] present a workflow for automatic music tagging emphasizing interpretability through perceptual features, integrating signal processing, deep learning, and symbolic knowledge. The approach demonstrates competitive performance with state-of-the-art methods on the MTG-Jamendo and GTZAN datasets, underscoring the value of interpretability despite potential performance trade-offs. Dervakos et al.[16] presents an advanced approach to music genre recognition using CNNs as mentioned in a previous subsection. A notable aspect of the study is its exploration into XAI techniques applied to music genre classification. The researchers adapt various post hoc explainability methods, including Grad-CAM, LIME, and a modified Genetic Programming for Explainability (GPX), to provide insights into the CNN's decision-making process.

Innovative techniques are being developed to balance accuracy and explainability in MIR systems. Chowdhury et al.[14] introduces a deep learning model that predicts music's emotional aspects based on mid-level perceptual features, aiming for explainability in MIR systems. The research employs a VGG-style network, showing minimal performance loss when incorporating these perceptual features, which serve as interpretable, musically meaningful intermediaries. The model facilitates understanding of emotion predictions, justifying the slight reduction in accuracy for the benefit of explainability. This approach illustrates how mid-level features can offer insights into a model's emotion predictions. Zhang et al.[65] introduces BART-fusion, a novel model that generates interpretations of song lyrics by integrating a large-scale pre-trained language model with an audio encoder through a cross-modal attention mechanism. This approach allows the model to understand songs from both lyrics and audio perspectives, leading to precise and fluent interpretations. Experimental results demonstrate that incorporating audio information improves the model's ability to understand and generate interpretations. Won et al.[63] propose a music tagging model utilizing self-attention and CNNs to enhance interpretability while maintaining competitive performance. Their architecture, designed to capture both local characteristics and long-term relationships within music tracks, includes shallow convolutional layers followed by stacked Transformer encoders. The model outperforms traditional fully convolutional and recurrent neural network approaches in interpretability without sacrificing accuracy.

Rodis et al. [52] provide a comprehensive review of Multimodal Explainable Artificial Intelligence (MXAI), elucidating its methodological advances and pointing towards future research directions. Their work systematically categorizes MXAI's main prediction tasks, datasets, and methods, while offering a critical evaluation

of current methods according to their handling of different modalities, the stage of explanation generation, and the types of methodologies applied. This review is particularly invaluable for its thorough analysis of the metrics used to evaluate MXAI methods, highlighting how these techniques not only advance our understanding but also improve the transparency and accountability of AI systems across various applications.

Chapter 4

Methodology

In this chapter, we aim to detail the systematic methods undertaken in our study. We begin with an in-depth analysis of the original dataset, including an examination of the dataset's label distributions, a critical assessment of its structure and content, and a discussion on the necessity and methods for its refinement and augmentation. Subsequently, we describe the models deployed for the regression and classification tasks. This section outlines the architectural choices, the rationale behind these choices, and the specific configurations employed, providing a clear view of how these models are expected to interact with our refined dataset. Lastly, the chapter presents the explainability methods we implemented to scrutinize the decision-making processes of our models. This multifaceted approach ensures a thorough understanding of the methodologies driving our research, setting a solid foundation for the subsequent analysis presented in the results chapter. Code implementation of our processes can be fount at our GitHub repository.

4.1 Music4All dataset

The Music4All (M4A) database is a substantial new resource designed to support a variety of research in the field of MIR. It contains a rich compilation of metadata, tags, genre information, 30-second audio clips, lyrics, and more, collected from a wide range of music pieces. The development of the database was carried out in two phases: the user phase and the song phase. In the user phase, data regarding users' listening histories was gathered and anonymized, while the song phase involved the collection of detailed song data. The dataset is particularly notable for its extensive size, containing data on over 100,000 songs, and is equipped to facilitate several traditional MIR tasks such as music recommendation, genre classification, and mood classification. In this section we proceed to analyse the dataset's contents relevant to our tasks and the methods undertaken to slightly modify M4A and craft our dataset. Readers interested in further details about the dataset can find more information in their paper [54].

4.1.1 Genre Distribution

The dataset is characterized by its diversity in genres, offering as many as eight genre labels per track and featuring more than 600 unique genres. Instances where multiple genre fields contained the same value (e.g., 'genre1: rock' and 'genre2: rock') were streamlined to retain only a single instance of the genre. The distribution of the resulting dataset's 20 most populated genre labels can be seen in Figure 4.1.1. It is important to note the dataset's imbalance, with a predominant representation of Pop and Rock genres. This skew can be attributed to the widespread prevalence of these genres in the music industry, and their broad categorization by non experts. A detailed genre mapping to more generic labels is discussed in a later subsection.

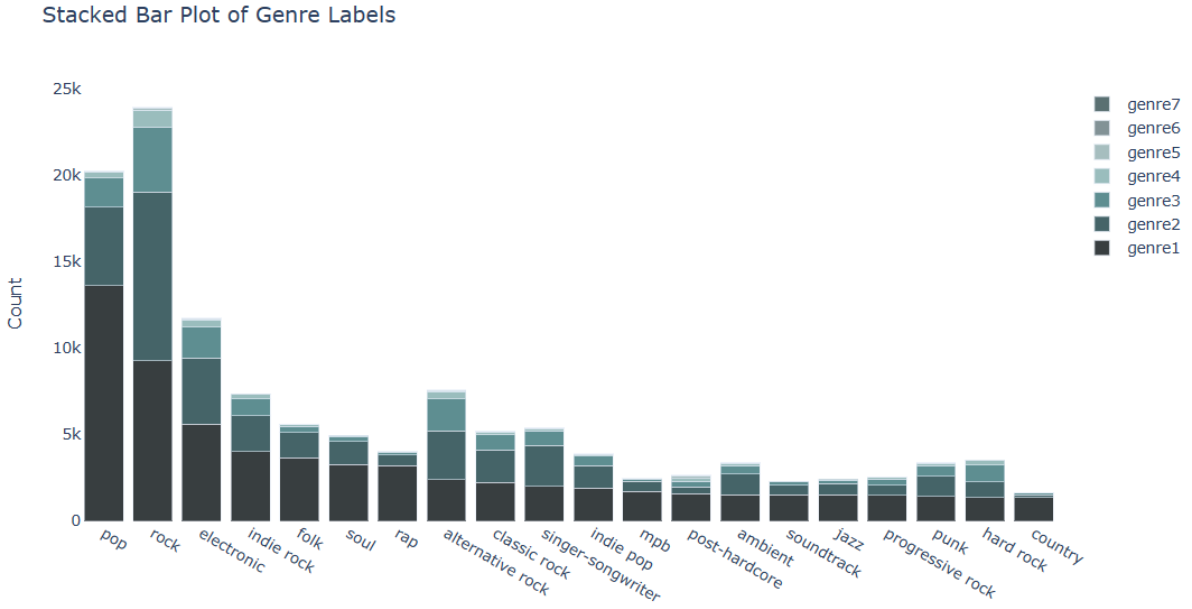
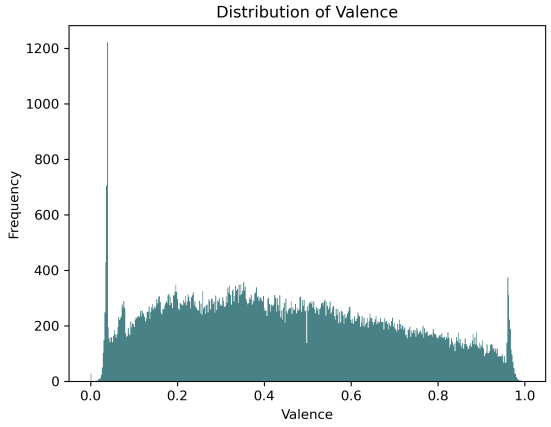


Figure 4.1.1: Figure containing the distribution of genres. Each color represents the distribution of one label for each entry.

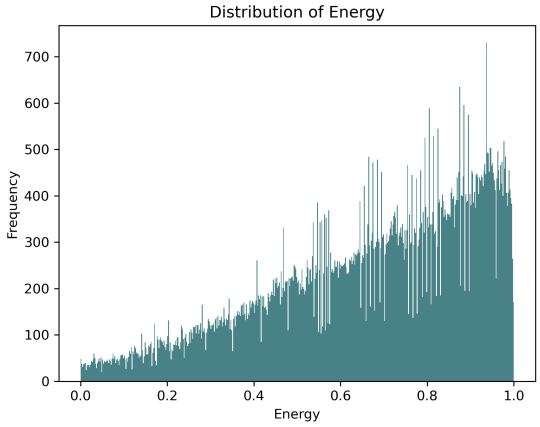
4.1.2 Valence and Energy Characteristics

The valence and energy values contained in the dataset, sourced from Spotify, are continuous, ranging from 0 to 1, and exhibit distinct distribution patterns. The valence scores tend to follow a quasi-normal distribution with a slight inclination towards less happy songs (instances with valence < 0.5), as can be seen in Figure 4.1.2a. In contrast, the energy distribution resembles the shape of a linear function, presented in Figure 4.1.2b, indicating fewer low-energy tracks and an abundance of high-energy ones. A Hexagonal Binning plot is also depicted in Figure 4.1.2c, representing the distribution of valence and energy values in a two-dimensional space. This plot reveals that high valence yet low energy songs are rare, indicated by the bright

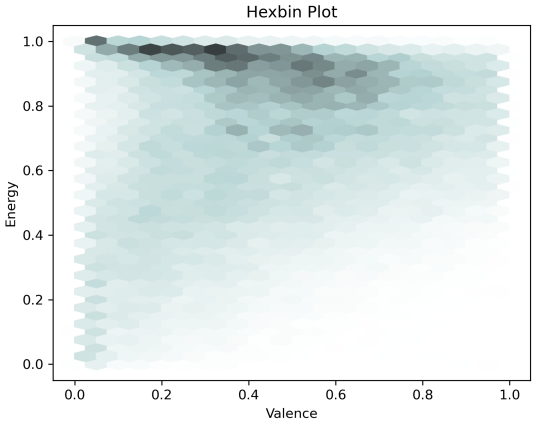
color of those hexes, whereas high energy but low valence songs are more frequent and the corresponding hexagon's color intensity is higher, suggesting a scarcity of relaxing and calm tunes in the dataset. This imbalance can be explained by popular music trends, favoring more upbeat, energetic tracks leading to their over-representation in collections and datasets.



(a) Valence Distribution



(b) Energy Distribution



(c) Hexbin Plot

Figure 4.1.2: Distributions of valence and energy as well as a hexbin plot representing the distribution of valence and energy values in a two-dimensional space, where each hexagon's color intensity corresponds to the concentration of songs with those valence and energy levels. The darker the color the higher the concentration of such songs.

4.1.3 Dataset Customizing, Augmentation, Mapping and Balancing

In our approach to genre classification, we recognized the need to condense the diverse range of music genres into more general categories. This simplification is crucial for effective classification and analysis. To guide our reclassification, we utilized the genre mapping available at Musicmap. This resource provided a plethora of information as well as a visual categorization of genres which proved useful for consolidating various music genres into nine broad classes. These classes were carefully selected to encompass the wide spectrum of musical styles while maintaining distinct and meaningful categories for our classification task. The labels of these classes are:

- Rock, encompassing rock 'n' roll, golden age rock, classic rock and contemprorary rock.
- Pop, a broad category that includes popular music styles prevalent since the 1950s.

- Hip Hop, covering hip hop as well as rap songs.
- Alternative Rock, including indie, alternative and other styles that differ from mainstream rock.
- Heavy Music, this label is used for metal, hardcore, and industrial music with metal elements.
- Punk, consisting of punk rock and new wave songs.
- Electronic, involving electronic dance music, downtempo and industrial music with electronic elements.
- Rhythm Music, consisting of rhythm 'n' blues, blues, gospel, jazz and Jamaican music.
- Folk, also including country music, thus encompassing both traditional and contemporary folk and country songs.
- Other, which contains ambiguous and other genres that do not belong in any of the previous categories.

We encountered some challenges with the genre labels in the M4A dataset that could potentially impact the accuracy and reliability of our genre classification models. In order to get genre labels, the creators of the M4A dataset collect user annotated tags from the last.fm API. Then they filter the tags keeping genre names available at Every Noise at Once. Although last.fm API returns weights alongside the tags, the M4A dataset does not contain such weights and therefore it is hard to determine which genre label is more relevant for each song. Furthermore, the tags can sometimes be incorrect since the annotators are users and not experts in music, resulting in noisy labels. In order to tackle this problem, we also fetch the artist genres available at the SpotifyAPI. We reduce the artist genres in the 10 classes categories mentioned above. Lastly, we consider a genre label in M4A dataset to be correct if it is also included in the artist genres. Considering the last.fm API returns the tags sorted by their importance and therefore genre labels in the M4A dataset that appear first represent a song more accurately, we consider a song's primary genre the first genre label in M4A that is also present in the artist genres by spotifyAPI.

In the task of classifying music based on mood, we transition from continuous valence and energy values to discrete emotional labels. This is achieved through a defined mapping that categorizes each song into 9 emotional states in accordance to Russel's Circumplex Model. Namely entries with high valence (greater than 0.65) are labeled 'Exciting' if their energy is also high, 'Relaxing' if their energy is low (less than 0.35) and 'Happy' otherwise. For low valence (less than 0.35) the tracks are labeled 'Angry', 'Depressing' and 'Sad' if they have high, low or in between energy values respectively. For valence values in the middle range the tracks are deemed 'Tense', 'Calm' or 'Neutral' accordingly.

As was expected, the process of mapping continuous valence and energy values to discrete emotional labels led to an unbalanced dataset, particularly lacking in 'Relaxing' and 'Calm' songs. To address this, we developed scripts to fetch more data from the Spotify API. Since Spotify doesn't offer direct searching by valence and energy, we iterated through a dozen genres for years ranging from 1950 to 2023. The API call returns up to 50 most popular songs per year per genre, with an option to offset for subsequent sets of 50. We filtered these songs based on desired valence and energy values, ensuring the availability of a 30-second audio preview URL. Furthermore, we refined our dataset by removing non-English songs using the langdetect library and then fetched their lyrics using the Genius Lyrics API, thus enhancing the representation of underrepresented moods in our dataset.

To tackle the issues of over-represented classes and genre-ambiguity in the original dataset, we experiment by sub-sampling the original dataset to contain a balanced number of songs for each class, with the requirement that each track must be originally labeled under one specific genre. However training the models with this approach did not have significant impact in the performance metrics. We create our final dataset containing the lyrics alongside a 30s audio clips for each entry as well as mood and genre labels. We remove songs with the genre label "other". Our dataset involves 9 emotion labels for each song and 9 genre labels. We create a train-val-test split of the data, making sure that artists that appear on the train-val split are not also present on the test split. A summary of the process followed to generate our dataset can be seen in Figure 4.1.3. The resulting statistics and distribution of labels can be found in the next chapter.

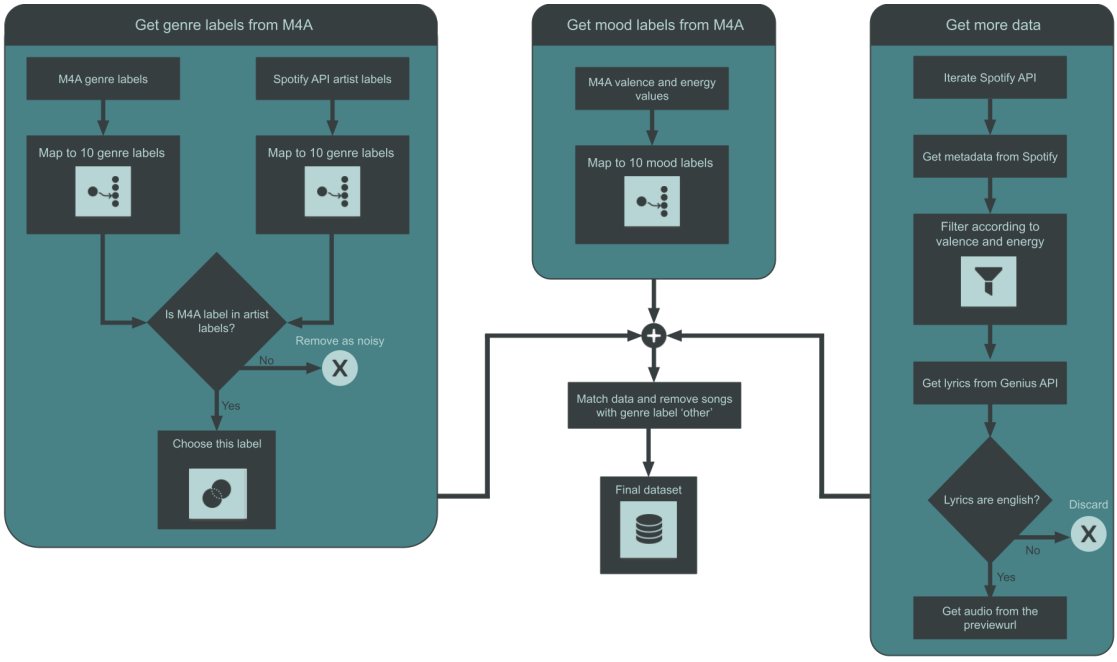


Figure 4.1.3: Summary of the process followed to modify the M4A dataset.

4.2 Feature selection and Model architecture

In recent years, the domain of deep learning (DL) has experienced tremendous growth, which has concurrently driven advances in music information retrieval (MIR) technologies. This section outlines our methodical approach to the selection, training, and integration of models within this vibrant technological landscape. To thoroughly assess the capabilities of various architectures, our strategy entailed a two-stage approach. Initially, in the first stage, we engaged with different regression models that focused on predicting valence and energy levels within music tracks. This phase was crucial for us to understand how different data and features interact within each modality, providing critical insights that guided our subsequent choices. The knowledge gained from this exploration phase was instrumental in pinpointing the most effective models for each modality, fine-tuning them not just for regression tasks but adapting them for classification challenges. Following this preliminary exploration, the second stage concentrated on the meticulous training of three distinct classification models. Each model was designed to harness the unique characteristics of its respective modality (text and audio). Additionally, we developed a multimodal model that combined these modalities, aiming to leverage their collective strengths for enhanced analytical depth and accuracy. We will now delve deeper into the specific methodologies employed at each stage, discussing in detail the processes of data preparation, model training, and rigorous evaluation.

4.2.1 Regression exploration

The aim of the models of this subsection is the prediction of continuous valence and energy values from music. Our initial task is a regression problem with a multimodal input - the lyrics and audio of songs. For a comprehensive understanding of our results, we present a variety of models spanning from rudimentary to more sophisticated designs. Some models utilized only one modality for their predictions, allowing for a better analysis of each modality's contribution. Furthermore, due to computational and time limitations, many of our models were trained on only a portion of the available data. A variety of loss functions were employed during the training process. However to ensure a standardized comparison across all models we chose the Mean Absolute Error as the metric for evaluating performance on the test dataset. This exploration of regression models serves as the backbone for feature extraction methods and model selection in our classification tasks. The results of this exploration are presented in detail in the next chapter.

In order to establish baseline performances for our suite of models, we implemented two fundamental re-

gressors: a dummy regressor and a mean regressor. The dummy regressor generates predictions based on random values within the expected range of our target variables, valence and energy. This model, devoid of any learning or pattern recognition, provides a baseline to ensure that our more complex models are indeed learning and not just randomly guessing. On the other hand, the mean regressor offers a slightly more informed baseline. It predicts the mean value of the training dataset's target variables, regardless of the input. While simplistic, this approach gives us a benchmark of predictability based on the central tendency of our dataset.

The creators of the M4A dataset [54] introduce an architecture for mood classification in Section III.C of their publication, utilizing only lyrics. They chose two Long Short-Term Memory (LSTM) models, one to infer valence values and the other for energy values. Each model contains an embedding layer, with an input sequence size of 500 words and an output dimension set to 200 dimensions, an LSTM layer with 128 units, and a dense layer utilizing the sigmoid activation function. The data was fed to the model with a batch size of 2,048 over 100 epochs.

Agrawa et al. [1] use a XLNet (xlnet-base-cased) to identify emotinal connotations of music based on lyrics. Their approach outperforms the existing methodologies available at the time of their study. Inspired by their work we explored the potential of XLNet in a regression task focused only on predicting valence values. Our implementation sets the max length of the input sequences to 640 and trains the model for 10 epochs on a subset of the available data.

In their study, Delbouys et al. [15] employ mid-level fusion to combine the capabilities of audio and lyric-based features, leveraging the concatenated outputs from distinct unimodal architectures. Their audio-focused model utilizes mel-spectrograms with 40 filterbanks and 1024 time frames, processed through two convolutional layers. On the other hand, their lyric analysis model employs Word2Vec embeddings—trained on 1.6 million lyrics and incorporates a convolutional and LSTM layer to effectively capture the semantic nuances of the lyrics. We adapted this architecture based on an unofficial implementation available at the following GitHub repository [56].

The paper [31] we base our next model on presents a novel multimodal approach to music information retrieval. It focuses on the development of a model that integrates high-level audio features such as danceability and acousticness, with lyric content to predict valence and arousal values. In our adaptation, we incorporated metadata values obtained from Spotify to enrich our audio feature set. Regarding lyrical analysis, we also employ Vader to extract sentiment information. Diverging however from the paper's methodology, we replaced the Pricipal Component Analysis (PCA) used in their Term Frequency - Inverse Document Frequency (TF-IDF) process with Singular Vector Decomposition (SVD) and chose to omit the use of Affective Norms for English Words (ANEW) features. These modifications fed into a Multilayer Perceptron (MLP) regressor with three hidden layers of sizes 128, 64 and 32.

MuLaN, as detailed in the paper [28], represents an advanced approach in the field of music emotion recognition. In our adaptation of MuLaN, we attempted to repurpose its methodology by feeding it lyrics instead of natural language music descriptions, training the model over 70,000 steps with a batch size of 10 on half of our available data. Our approach combined the latent features from the audio and textual streams, followed by a regression analysis on these concatenated features.

In our venture to harness state-of-the-art techniques for music emotion recognition, we combined the strengths of two powerful pretrained models: an audio spectrogram transformer based on the Vision Transformer (ViT) architecture, detailed in [22], and the robustly optimized BERT approach, RoBERTa [38], both of which are readily available on the Hugging Face platform. Our custom model integrates these two components: the ASTModel processes audio spectrograms, while the RobertaModel from the roberta-base collection, tokenizes and encodes our lyrics. The concatenated output features – the mean of the last hidden states from the audio model and the first token's hidden state from the RoBERTa model – feed into a neural network with a linear layer, dropout regularization, and sigmoid activation.

4.2.2 Final model architecture and training

After this thorough investigation of feature engineering and model architectures RoBERTa and AST stand out as front runners. To optimize classification performance, we devised three distinct models, each tailored

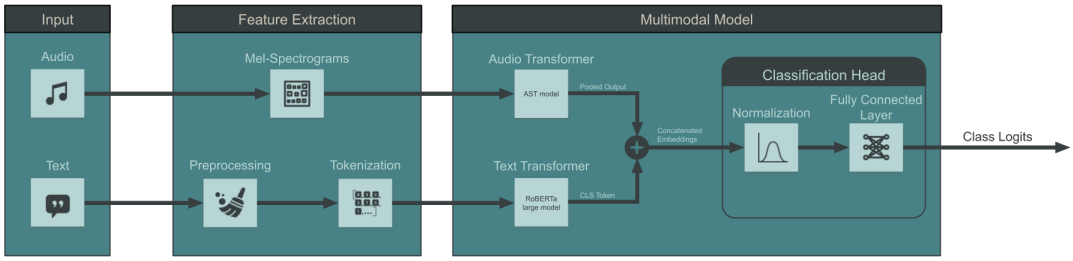


Figure 4.2.1: The pipeline of our multimodal approach. The audio waveform and the lyrics are first converted into mel-spectrograms and tokens respectively. Then each modality input is processed by the respective pretrained transformer model. The pooled output of the audio embeddings produced by the ASTmodel and the CLS token from roberta are concatenated and processed by a classification head, providing the logits for each class.

to handle two specific tasks: one model is dedicated to analyzing lyrical content, the second model focuses on processing audio spectrograms, and the third model integrates both data modalities for a holistic analysis.

The lyrical model leverages the power of the roberta-large architecture for text classification, namely RobertaForSequenceClassification. We first prepare the input text for tokenization by converting all characters to lowercase to maintain consistency. punctuation is removed to reduce complexity and potential noise in the data. Subsequently, newline and tab characters are replaced with spaces, and any superfluous whitespace is condensed into a single space between words. The tokenization function, configured with a maximum sequence length of 256 tokens, encodes the clean text into a format suitable for a RoBERTa model. This function generates numerical ids and an attention mask that indicates which tokens should be attended to by the model. We proceed to train the model by fine-tuning for sequence classification across the distinct label categories present in our dataset. The optimizer of choice is AdamW, set with a learning rate of 9e-7, to guide the learning process. The model is trained over a course of 9 epochs, with a persistent checkpointing mechanism that monitors validation accuracy preserving the model state that achieves the highest validation accuracy, thereby ensuring we retain the most performant model at the end of the training cycle.

The audio model leverages the pretrained Audio Spectrogram Transformer for audio classification loaded from the checkpoint 'ast-finetuned-audioset-10-10-0.4593' provided by MIT. We first use the ASTFeatureExtractor class to convert audio data into spectrograms. It's important to note that only a portion of the original audio is used in this process. With a typical audio sample rate of 44100 Hz and a maximum length of 1024 for the spectrogram as set by this class, only around 10.2 seconds of the original 30-second audio are actually used to create the spectrogram. The optimization of the model's parameters is entrusted to the Adam optimizer, utilizing a fine-tuned learning rate of 6e-6 and training for a total of 5 epochs.

Our final model's architecture is designed to integrate and classify multimodal inputs from both audio and text sources. The model leverages pretrained components: the ASTModel for audio features and the RobertaModel for textual analysis. The audio and text inputs are preprocessed in the same way as in the unimodal cases. The pooled output from the audio model and the CLS token from the text model are chosen as they provide a comprehensive summary of their respective modality's content. These embeddings, with sizes 1024 and 768 respectively, are then combined into a unified feature vector and processed by a classification head. This component initially normalizes the combined features using a layer normalization step to stabilize the learning process. Following normalization, a fully connected layer maps the normalized features to the desired number of output labels (9 labels for both tasks), producing the final classification logits. The pipeline of our approach can be seen in 4.2.1. We train the multimodal models for each task using the Adam optimizer with 6e-6 learning rate over a course of 6 epochs and keep the weights of the model on the epoch that performs best on the validation set.

To train our models effectively, we leveraged the powerful computing resources available on Google Colab

and Kaggle. Specifically, our configurations utilized NVIDIA's V100 GPUs on Google Colab and P100 GPUs on Kaggle. Kaggle provides users with up to 30 hours of free GPU usage each week, making it a cost-effective option for our experiments. On the other hand, Google Colab requires a paid subscription to ensure consistent access to its GPUs, which can be crucial for longer or more resource-intensive training sessions. As part of our methodology, each model was trained with distinct learning rates and batch sizes, tailored to optimize computational efficiency and model performance. The batch sizes varied across models, with each requiring approximately 13 GBs of GPU RAM. Typically, models were trained for around 12 hours. However, the multimodal models, due to their increased complexity and data sizes, demanded up to 24 hours of training time. Looking ahead, future work could explore the potential benefits of reducing learning rates and increasing the number of training epochs. This approach aims to minimize validation loss more effectively and prevent the common pitfalls of overfitting and underfitting, thereby enhancing the overall accuracy and reliability of our models.

4.3 Explainability Methodology

In this section of our report, we delve into the techniques and approaches employed to unravel the decision-making processes of the AI models created. Our focus is on ensuring that these methods not only clarify how decisions are derived but also enhance the transparency and trustworthiness of the models. We outline the practical tools, mainly LIME, we have adopted to explain the lyrical and audio single-modal models. Also, we propose an adaptation of these tools for the task involving both of those modalities. We finally outline the method followed to get global aggregations of local explanations.

4.3.1 Lyrical Explainability Method

While several techniques like TextFooLer[30] and MiCE[53] were considered for explaining text data, we opted to implement LIME explanations. LIME begins by creatinga binary vector with equal length to the original corpus, indicating the presence or absence of words. For example, the phrase "Come as you are as you were" would be converted into a vector of length 7. A vector representing the presence of all the words would be [1, 1, 1, 1, 1, 1, 1] whereas a vector representing the absence of the words "Come" and "were" would be [0, 1, 1, 1, 1, 1, 0]. Lime then perturbs the input data by randomly turning words on or off (present or absent) creating a set number of samples around the original input. In the previous example LIME could create the phrases "as you are as you" and "Come as you are" represented by the two-dimensional vector of vectors [[0, 1, 1, 1, 1, 1, 0], [1, 1, 1, 1, 0, 0, 0]]. After generating these perturbed samples, LIME utilizes the original model to predict outcomes for each, treating these predictions as labels for training a simpler, interpretable model, typically a linear regression. This interpretable model is designed to approximate the decision-making process of the complex original model within the locality of the input example. By emphasizing perturbed samples that are closer to the original text in terms of their transformed feature space, LIME assigns higher weights to them during this training phase. This process ensures that the explanation model focuses on the most relevant variations of the input data, highlighting which words contribute most significantly to the model's prediction. An example can be seen in figure 4.3.1. The example depicts a portion of Nirvana's "Come as You Are" lyrics, with some words that influenced the model to decide the "alternative rock" class, highlighted in blue. The higher the opacity, the higher the weight of the word.

In our study we employ LIME to provide local approximations of RoBERTa's decision-making processes. Given the max_length parameter in RoBERTa, which restricts the number of tokens the model can process in a single input, we implement a truncation function to ensure that the input text to LIME is precisely what RoBERTa evaluates. This truncation function carefully slices the text to fit within RoBERTa's token limit, ensuring that all perturbations generated by LIME are relevant and within the accepted input size of the model. Additionally, since LIME requires not just perturbed texts but also the corresponding output probabilities for each class, we have defined a wrapper function around RoBERTa. This function accepts the perturbed texts as input and returns the class probabilities, facilitating a seamless integration of LIME with RoBERTa to interpret the model's predictions accurately.

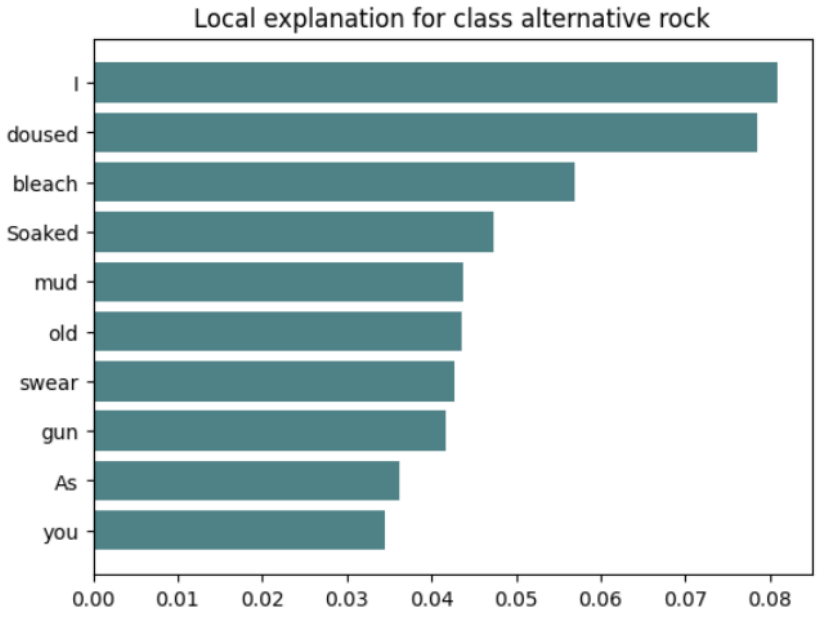
Come as you are, as you were
As I want you to be
As a friend, as a friend
As an known enemy

Take your time, hurry up
The choice is yours, don't be late
Take a rest as a friend
As an old enemy

Memory, memory
Memory, memory

Come doused in mud
Soaked in bleach
As I want you to be
As a trend, as a friend
As an old enemy

(a) Part of the Lyrics of Nirvana's "Come as you are" with important words highlighted.



(b) The weights of the highlighted words according to LIME

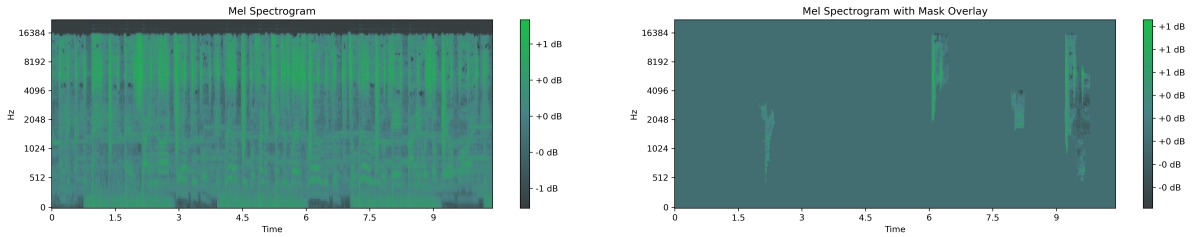
Figure 4.3.1: An example of LIME text explanations given Nirvana's "Come as You Are" as input. Words that influence the model to decide "alternative rock" as the class are highlighted in blue. The higher the opacity, the higher the weight of the word as seen in figure (b).

4.3.2 Audio Explainability Methods

In the realm of audio explainability our journey begins with the insights offered by LIME for interpreting image data. In this case, instead of splitting a text corpus into strings, LIME dissects images into superpixels. It then creates perturbations where each superpixel is toggled on or off, similar to the word perturbations in text. Each superpixel is then weighted based on its influence on the model's output, with the most impactful ones highlighted in the original image to indicate their significance. Adapting this method to our audio domain, we utilize spectrograms, which essentially serve as grayscale images of sound. This allows us to apply LIME directly to our audio data represented in spectrogram format. An example spectrogram and it's masked version as generated by LIME can be seen in figure 4.3.2. To get a listenable explanation from the spectrograms we implement two strategies: The first one involves attempting to recreate the audio from the highlighted parts of the spectrograms outputted by LIME using an inverse short-time fourier transform (STFT). The second strategy entails filtering the original audio to isolate the specific portions of the spectrogram that LIME highlights as the most influential. Although these approaches yielded some listenable explanations, their results, as presented in the next chapter, suggest there is room for further refinement and improvement.

LIME for image explanations did not deliver satisfactory results for our audio analysis needs. That is, despite highlighting the part of the spectrogram most relevant for the model's decision, the listenable explanations generated were not of high quality. Therefore we explored alternative methodologies. Our research led us to audioLIME[25], a LIME variant specifically tailored for music data. Unlike traditional LIME, which operates on spectrograms, audioLIME directly perturbs the audio itself, enabling a more nuanced exploration of sound components. It originally employs source separation technology using spleeter [26] to isolate individual elements of a song such as vocals, drums, bass, and piano. Additionally, audioLIME segments the audio into temporal segments. Finally, similar to how LIME calculates feature importance, audioLIME assesses the impact of each segment on the model's decision-making process. A figure depicting this pipeline can be seen in figure 4.3.4.

However, recognizing limitations in spleeter's ability to accurately decompose complex audio signals, we



(a) Example Mel Spectrogram

(b) The spectrogram with the mask generated by LIME

Figure 4.3.2: Two figures depicting a mel spectrogram and a masked spectrogram. The mask was generated by LIME and presents the areas of the spectrogram that influence a model's decision the most.

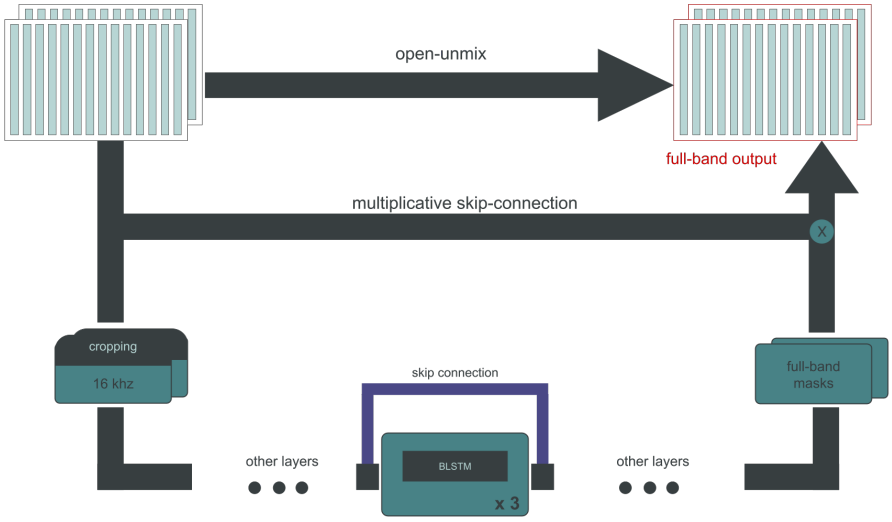


Figure 4.3.3: The openunmix model for one source.

opted to integrate a more successful factorization technique, open-unmix [57] (UMX), which is a deep neural network designed for precise audio source separation. It allows us to separate audio into vocals, drums, bass and other instruments. The system's architecture includes multiple models, each trained specifically for a target source, which allows for customized training data for each source. The core of Open Unmix is a three-layer bidirectional deep LSTM (Long Short-Term Memory) network that predicts the magnitude spectrogram of a target from the mixed input. The model operates in the time-frequency domain, using Short-Time Fourier Transforms (STFTs) to process the input signals. During separation, the network applies a mask to the input spectrogram, and the output is optimized using mean squared error in the magnitude domain. This model for one source can be seen in Figure 4.3.3.

Similarly to the textual approach, we create a function that receives the perturbed waveforms from audioLIME (e.g. the original waveform without some vocal segments), converts them to spectrograms, feeds them to our model and returns class probabilities. The result of this process are weights for each segment of each source. We decide to split the audio in 10 temporal segments resulting in around 1.2 seconds of listenable audio explanations. The components are labeled with the name of the source followed by a number denoting the temporal sequence ranging from 0 to 9. For example the component label "vocals5" indicate that this component contains the vocals of the audio clip from 4.8s to 6s. This combination of audioLIME and UMX allows us to generate more accurate and insightful explanations of our audio data, providing clearer guidance on the influential elements within each audio segment.

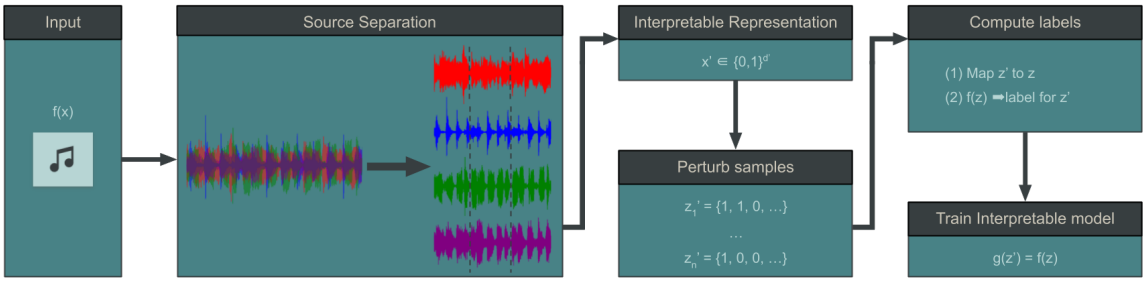


Figure 4.3.4: The aduioLIME pipeline as presented in their paper[25]. It closely follows the general LIME approach with the key difference of using source separation.

4.3.3 Multimodal Explainability Method

We venture into the realm of multimodality and explainability. Influenced by the interesting results by LIME and audioLIME in the unimodal approaches, we propose a way to combine them in order to explain the multimodal model and get multimodal explanations. We begin by creating a binary vector indicating the presence or absence of a feature. The vector has length equal to the sum of the number of text features and the number of audio features. For example, suppose we have the lyric "Come as you are as you were" as text input and a two-second audio clip, set to be split by the explainer into one second segments, factorized into "vocals", "drums" and "other". The total number of features is 13, 7 text features representing each word and 6 audio features representing the vocals, drums and rest of the 1 second audio clips. The process to generate explanations continues similar to LIME, by generating perturbations of the inputs and calculating the weights of every feature. As a result, we can determine which feature are important for the multimodal model's decision, compare the modalities and have listenable explanations as well as textual ones. The respective pipeline is presented in Figure 4.3.5.

We should also mention that we experimented other forms of explanations. One such form was to access the weights of the classification layer in order to determine which modality was more influential for each class. Another approach was to keep one modality fixed and perturb the other modality's input. These methods provided some insights into the model's behavior; however, the multimodal method mentioned above delivered clear, textual, and audible explanations. These not only identify influential features but also allow for further contextual, musical, and cultural analysis.

4.3.4 Global Aggregations of Local Explanations

Local explanations often fail to reflect the model's overall behavior. For a more comprehensive understanding of what features are influential across the model, rather than just specific instances we implement Global Aggregations of Local Explanations as outlined in this paper [35]. The authors first mention three methods to generate aggregates: (a) the Global LIME importance, (b) the Global Average Importance and (c) the Global homogeneity-weighted importance. The Global LIME importance for a class c for a feature j is

$$I_{cj}^{\text{LIME}} = \sqrt{\sum_{i \in S_c} |W_{ij}|}$$

This function assumes that features that occur more often are expected to have a larger effect on model predictions than features that occur less often. In a scenario where the features are words, global LIME importance will be unreasonably biased towards common words that appear more often. To address this assumption the authors propose the global average class importance for feature j and class c defined as:

$$I_{cj}^{\text{AVG}} = \frac{\sum_{i \in S_c} |W_{ij}|}{\sum_{i \in S_c : W_{ij} \neq 0} 1}$$

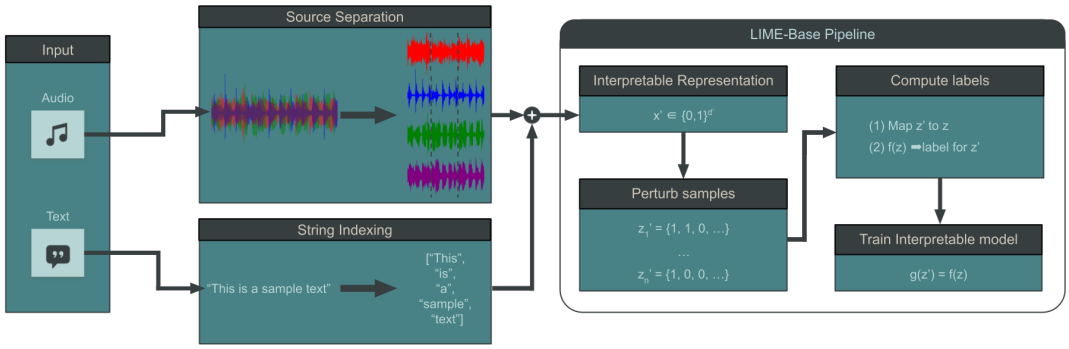


Figure 4.3.5: The pipeline of our approach to generate multimodal explanations. As seen in the figure we first split the audio into it's sources and into temporal segments and the text input into individual components. We create interpretable representation of the combined features and utilize the LIME-base pipeline.

where S_c includes all the instances i classified as class c and W_{ij} is the weight of that feature for that specific instance. Finally, the authors then calculate the vector of normalize LIME importance per class:

$$p_{cj} = \frac{\sqrt{\sum_{i \in S_c} |W_{ij}|}}{\sum_{c \in L} \sqrt{\sum_{i \in S_c} |W_{ij}|}}$$

where L is the set of all labels. The normalized LIME importance p_j represents the distribution of feature j's importance over all classes $c \in L$. The Shannon entropy of this distribution is defined by:

$$H_j = -\sum_{c \in L} p_{cj} \log(p_{cj})$$

and is used to asses the degree of homogeneity with which the feature attributions are distributed over multiple classes. Finally in order make out cases where features appear only once in the test set the authors calculate the homogeneity weighted importance for feature j and class c:

$$I_{cj}^{H} = \left(1 - \frac{H_j - H_{\min}}{H_{\max} - H_{\min}}\right) I_{cj}^{\text{LIME}}$$

where H_{min} and H_{max} are the minimum and maximum entropy measured across all features. In short $I_{cj}H$ addresses the issue of common features that may appear significant due to their presence rather than their informative value and corrects for this by using entropy to penalize features that are uniformly distributed across classes, highlighting those that are truly predictive of specific outcomes. It also addresses the assumption that features uniformly affect model outcomes, by adjusting the importance based on the consistency of their influence across classes and therefore penalizing features that show high variability across classes. This method ensures that the global importance reflects genuine, consistent predictive value, especially in complex multiclass scenarios.

We implement these methods for our tasks to evaluate models' behaviour and trustworthiness. Due to the cost in time and computing power, the global aggregations were generated from a subset of our test dataframe. Once again we utilize Google Colab and Kaggle to get the aggregates utilizing NVIDIA's V100 and P100 respectively. For each model and for each task we first calculate $\sum_{i \in S_c} |W_{ij}|$ which we then utilize to generate the avarage and homogeneity-weighted class importance. For the lyrical models we utilized 608 instances of the test set with 5000 perturbations, for the audio models we utilized 232 instances with 2000 perturbations each and for the multimodal models 63 instances with 5000 perturbations. Each process took approximately 12 hours to run. Although the number of instances was limited, particularly in the case of the

multimodal model, the results of this process as seen in the next chapter were adequate. Future work could not only implement a statistical analysis to determine the number of instances so that the global aggregates are significant but also investigate the number of samples/perturbations necessary for a local explanation to be accurate, given the number of features of the input.

To visualize the results, apart from studying the weights of each feature of the global aggregations, we also plot the most important features for each class for the text features. To achieve that we utilize GloVe embeddings and t-SNE for dimensionality reduction for the lyrical modality. K-means clustering was implemented to identify and analyze the underlying themes that significantly influence the model's decision-making process. We should note that some words like exclamations or vocalizations and names do not have a GloVe embedding and therefore do not appear in the corresponding plots. For the audio modality's global aggregates, we provide heatmaps visualizing the feature weights for every source accross the sequence of temporal segments.

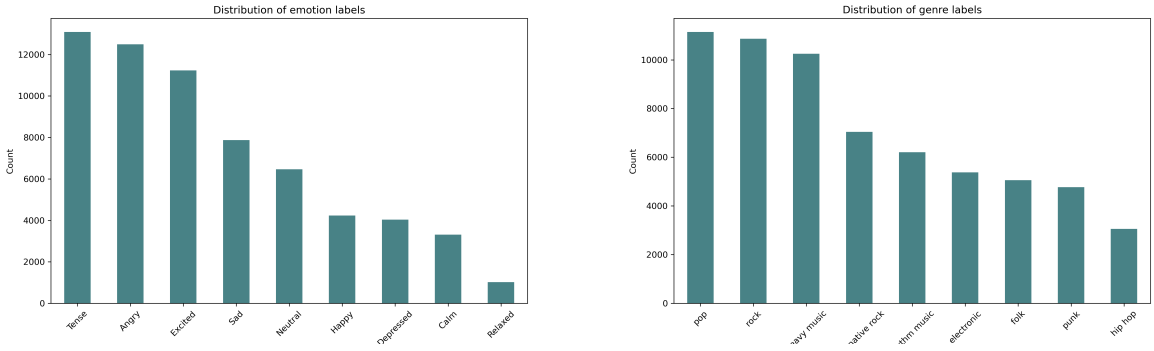
Chapter 5

Results

In this part of the thesis, we present the findings from the methodologies discussed in the previous chapter. Here, readers can expect to find details on the final distributions of our dataset. Furthermore, we showcase the metrics of the models used for each task as well as a comparison and discussion of those results. To further analyze and understand these outcomes, we finally exhibit the in depth examination of these models using the explainable methodologies. This section is particularly important as it highlights the efforts and outcomes of our work across the areas of MIR, multimodality, and explainability. By laying out these results, we aim to give a clear and comprehensive view of our research accomplishments and pave the way for the detailed discussions that will follow.

5.1 The Final Dataset

After the processes of augmenting and refining the M4A dataset described in Chapter 4 we are ready to showcase the resulting dataset's summary. There are 63760 total number of entries, containing the lyrics as well as the RoBERTa tokens, the audio as well as the generated spectrogram and various other metadata including one emotion label and one genre label for each song. The distribution of the emotion labels and the genre labels can be seen in Figure 5.1.1. We note that despite our efforts to collect more 'relaxed' tracks, we found it challenging to enrich our set with any more than the 1020 we included in it (498 of which where in the M4A dataset). This scarcity could be attributed to not only the fact that such songs are not very popular in western music but also might not contain english lyrics or any lyrics at all. We can also see a similar but less intense imbalance in the genre distribution, with a bias towards 'pop' songs while 'hip hop' songs appear less frequently. Out of all the entries, 50660 were utilized as the training set and the reset were split among the validation and the test set. Artists that appear in the train-val split are not included in the test set.



(a) The Emotion labels distribution.

(b) The Genre labels distribution

Figure 5.1.1: Emotion and Genre label distributions of the final dataset.

5.2 Our Models' Performance

In this section we present the performance of the trained models. We begin by analysing the regression model's results which led us to adopt RoBERTa and AST for our classification tasks. We continue with the results of each classification model, providing confusion matrices and classification reports as well as a summary and a discussion about the outcomes. Further investigation on model behaviour is present in the next explainability section.

5.2.1 Regression models' results

By analysing the results of our regression tasks, depicted in Table 5.1, we make the following observations. Firstly, the Mean Absolute Error (MAE) of the mean regressor, around 0.2 for both valence and energy predictions, serves as a baseline for evaluating more complex architectures. Remarkably the simple MLP architecture yielded very successful results. On the other hand, despite the innovative design of MuLaN, our implementation using musiclm-pytorch library lead to a moderate performance. Finally, the superior performance of the RoBERTA and AST implementation, especially for valence predictions, led us to adopting it for our upcoming tasks of mood and genre classifications.

Generally, the models were better at predicting energy values as opposed to valence. This discrepancy may be attributed to the distribution characteristics of the dataset, where valence data exhibits a more uniform distribution, whereas energy values are skewed closer to 1, as also reflected in the predictions of the mean regressors. While further fine-tuning and training with more data could potentially enhance these models' accuracy, such optimization is beyond the scope of our current research task. It's worth noting the rapid advancements in the transformer domain of Deep Learning. New implementation in this area, such as

the Causal Audio Transformer (CAT) [37], have set new benchmarks in audio classification. However, the absence of implementation code for CAT and the difficulty of recreating such a sophisticated model without a comprehensive library presents a significant challenge in applying this technology to our current project.

Model	Mean Absolute Error			Modalities	
	Valence	Energy	Average	Audio	Lyrics
Dummy	0.314	0.342	0.328	-	-
Mean	0.206	0.194	0.200	-	-
Music4All LSTM	0.210	0.197	0.203	-	Embedding
XLNet	0.183	-	-	-	XLNet emb.
Conv. Net.*	0.158	0.107	0.211	Spectograms	Word2Vec emb.
MLP	0.157	**0.085**	0.121	High Level Features	Vader, TF-IDF
MuLaN*	0.195	0.148	0.171	Spectograms	
AST + RoBERTa	**0.140**	0.099	**0.119**	Spectograms	RoBERTa emb.

Table 5.1: Model Performance Comparison and Modality Utilization. *Models marked with an asterisk were trained and tested on a portion of the available data due to computational costs.

5.2.2 Final models' results and discussion

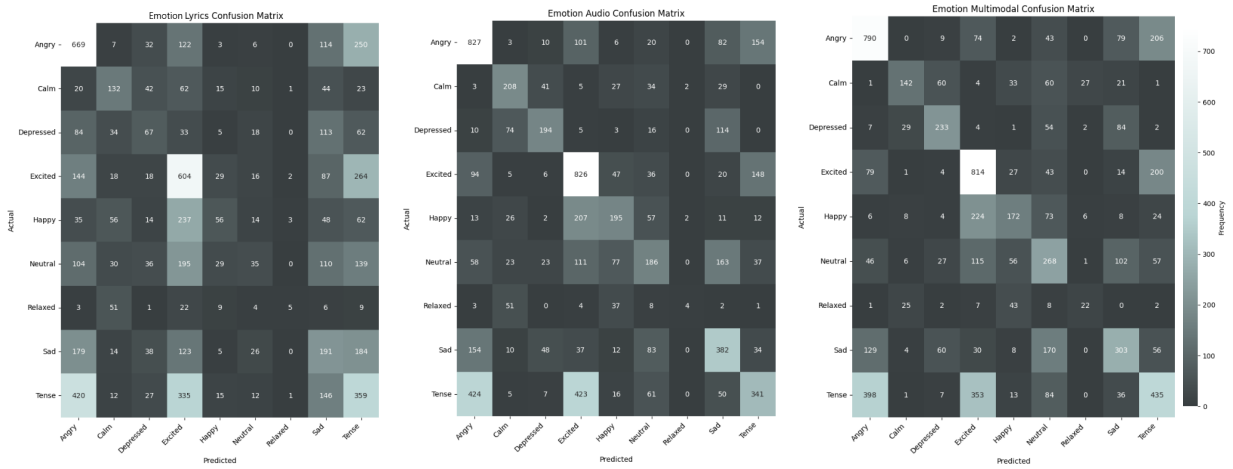
We use the final dataset presented in the previous section to train and evaluate our models. A summary of the performance of our models can be seen in Table 5.2. To gain a deeper insight into the strengths and weaknesses of our models, we provide detailed confusion matrices and classification reports. We proceed by demonstrating and analyzing the resulting models. Further investigation on model behaviour can be found in the next section, where we implement XAI methods to draw accurate conclusions.

Model	Valid Accuracy	Test Accuracy	Epochs
Lyrical Emotion	34.03%	32.33%	9 (5)
Audio Emotion	48.33%	48.29%	5 (3)
Multimodal Emotion	49.05%	48.53%	5 (3)
Lyrical Genre	46.9%	45.14%	9 (7)
Audio Genre	55.63%	53.75%	5 (4)
Multimodal Genre	60.33%	57.34%	5 (2)

Table 5.2: Model Performance Summary

The classification reports and confusion matrices for the emotion models are depicted in Table 5.3 and Figure 5.2.1 respectively. The Lyrical model tasked with predicting emotion labels is the worst performing model of all with validation accuracy of only 34.03%, suggesting that capturing emotion information from lyrical context is not reliable and that such information can be more optimally detected through audio feature analysis. This is also hinted not only by the fact that the audio model greatly outperforms the lyrical one, but also by the observation that the multimodal approach does not yield significantly better results than the audio one.

The classification reports in table 5.4 and the confusion matrices in Figure 5.2.2 depict the results of the genre classification models. The lyrical model although has relatively low metrics, with a validation accuracy of 46.9%, is particularly successful in predicting certain classes, namely 'hip hop' and 'heavy music'. This



(a) Lyrical Model (b) Audio Model (c) Multimodal Model

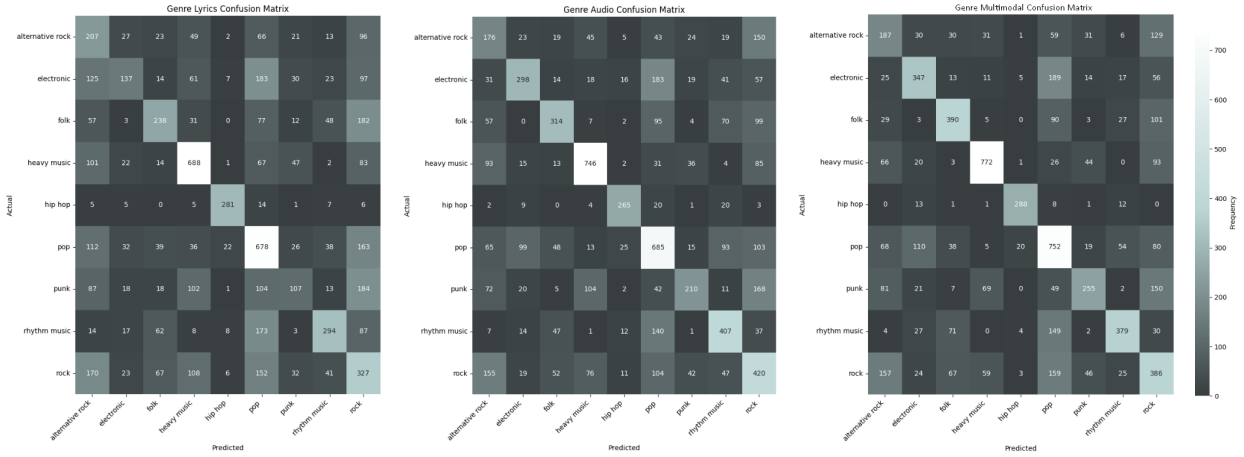
Figure 5.2.1: The confusion matrices of the emotion classification lyrics, audio and multimodal models. Brighter cells indicate higher concentration of model predictions. Ideally the diagonal of the matrix should contain all the values and the rest of the matrix should be null.

Table 5.3: Classification Reports for Three Models on Emotional Classes

Emotion	Lyrical Model Precision, Recall, F1	Audio Model Precision, Recall, F1	Multimodal Model Precision, Recall, F1	Support
Angry	0.40, 0.56, 0.47	0.52, 0.69, 0.59	0.54, 0.66, 0.59	1203
Calm	0.37, 0.38, 0.38	0.51, 0.60, 0.55	0.66, 0.41, 0.50	349
Depressed	0.24, 0.16, 0.19	0.59, 0.47, 0.52	0.57, 0.56, 0.57	416
Excited	0.35, 0.51, 0.41	0.48, 0.70, 0.57	0.50, 0.69, 0.58	1182
Happy	0.34, 0.11, 0.16	0.46, 0.37, 0.41	0.48, 0.33, 0.39	525
Neutral	0.25, 0.05, 0.09	0.37, 0.27, 0.32	0.33, 0.40, 0.36	678
Relaxed	0.42, 0.05, 0.08	0.50, 0.04, 0.07	0.38, 0.20, 0.26	110
Sad	0.22, 0.25, 0.24	0.45, 0.50, 0.47	0.47, 0.40, 0.43	760
Tense	0.27, 0.27, 0.27	0.47, 0.26, 0.33	0.44, 0.33, 0.38	1327

could be attributed to the recurrent themes present in this kind of music. Although the audio model is not as precise in predicting 'hip hop' instances, it outperforms the lyrical model in all other class metrics with a validation accuracy of 55.3%. Finally, the multimodal model surpasses both of these models with validation accuracy of 60.33%, proving that it is able to detect information that is not discernible when each modality is used independently.

Some interesting patterns emerge from the metrics of our models not only across the different modalities but also across the different tasks. To begin with, the models tasked with predicting genre labels greatly outperform the models responsible for emotion prediction. This pattern confirms our intuition that capturing the emotional complexity of musical tracks can be challenging due to the subjective nature of human emotions and the difficulty to provide as well as detect a generalized, objective and therefore accurate estimates of the emotional characteristics listeners might perceive in a track [21]. On the other hand, detecting themes prevalent in certain genres and as a result correctly learning to predict genre labels can be more successful. We should also note that the emotion models lagged behind perhaps also due to the more intense imbalance of the train data. Furthermore, our endeavours show that for both tasks the lyrical models had the worse performance and the multimodal models achieved the best results. This behaviour leads us to infer that information relevant to music classification tasks, especially for emotion classification, is present mainly in



(a) Lyrical Model (b) Audio Model (c) Multimodal Model

Figure 5.2.2: The confusion matrices of the genre classification lyrics, audio and multimodal models. Brighter cells indicate higher concentration of model predictions. Ideally the diagonal of the matrix should contain all the values and the rest of the matrix should be null.

Table 5.4: Classification Reports for Three Models

Genre	Lyrical Model Precision, Recall, F1	Audio Model Precision, Recall, F1	Multimodal Model Precision, Recall, F1	Support
Alternative Rock	0.24, 0.41, 0.30	0.27, 0.35, 0.30	0.30, 0.37, 0.33	504
Electronic	0.48, 0.20, 0.29	0.60, 0.44, 0.51	0.58, 0.51, 0.55	677
Folk	0.50, 0.37, 0.42	0.61, 0.48, 0.54	0.63, 0.60, 0.62	648
Heavy Music	0.63, 0.67, 0.65	0.74, 0.73, 0.73	0.81, 0.75, 0.78	1025
Hip Hop	0.86, 0.87, 0.86	0.78, 0.82, 0.80	0.89, 0.89, 0.89	324
Pop	0.45, 0.59, 0.51	0.51, 0.60, 0.55	0.51, 0.66, 0.57	1146
Punk	0.38, 0.17, 0.23	0.60, 0.33, 0.43	0.61, 0.40, 0.49	634
Rhythm Music	0.61, 0.44, 0.51	0.57, 0.61, 0.59	0.73, 0.57, 0.64	666
Rock	0.27, 0.35, 0.30	0.37, 0.45, 0.41	0.38, 0.42, 0.40	926

the audio domain. The performance of combining the text and audio modalities was optimal proving the potential of multimodality in enhancing music classification tasks and MIR tasks in general. We aim to analyse the models' behaviour even more in depth in the next section.

5.3 Explaining the Models

In this section our main objective is to shed light to the models' decision making process. As mentioned in the previous chapter we will implement LIME, audioLIME and the combinational MusicLIME in order to analyze our results. We also calculate the global aggregates of such local explanations using two aggregation methods. We analyse in detail the genre task's results, providing figures presenting the local explanations of some cherry picked instances or prototypes as well as the global aggregates for each class.

5.3.1 Lyrical Genre Model Explanations

We begin our analysis by presenting the results of the lyrical genre model explainability. For each class, we first present some themes that characterize them. We then choose three instances: a True Positive, a False

Positive and a False Negative and examine their local explanations. We combine the observations made for those instances with the those made from the global aggregates of local explanations to find what themes prevail for each class along with an evaluation of the model's behaviour.

Hip Hop

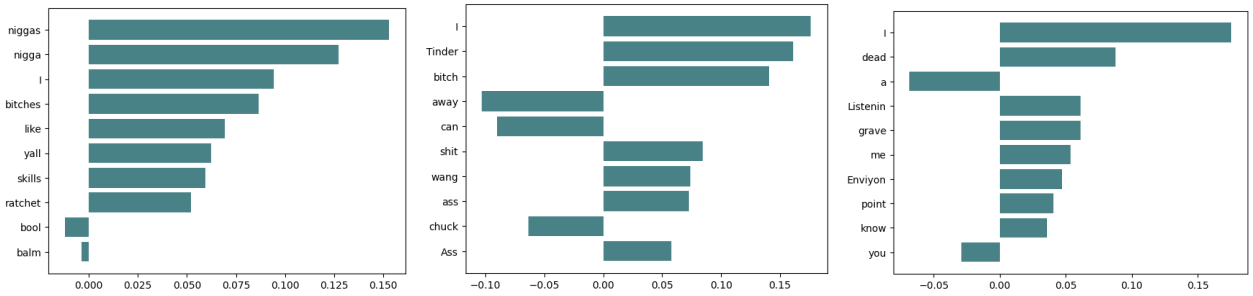
In this lyrical analysis, no other class stands out as prominently as Hip Hop does. This genre, originating as an anti-drug and anti-violence genre, includes some common themes such as social issues (poverty, racism and police brutality), personal struggles and Lifestyle or Culture. These recurring themes could explain why the model achieves almost 90% precision and recall on the test set.

To investigate this assumption we choose the following three songs and generate local explanations: Bool, Balm & Bollective by YG, Poontang Boomerang by Steel Panther and Lucid Dreams by Juice WRLD. The first song has a probability of 0.97 of being "hip hop" and serves as a true positive. The second one is mislabeled by the model as "hip hop" when in fact it belongs to the "heavy music" class. The third song is mislabeled "pop" by the model while it is an instance of "hip hop".

The feature (word) contributions for the "hip hop" class for each of these songs are depicted in 5.3.1. The first-person singular subject pronoun "I" seems to prevail across all three examples and contribute to the model's "hip hop" prediction probability. At first glance, from a data analysis perspective, it may appear worthy of being overlooked as a stopword. However this trend can be attributed to the genre's emphasis on personal experience, identity, self-expression and rivalry [2]. Moreover the presence of words associated with violence and incarceration, such as "prison", "dead" and "grave", influences the model's tendency to categorize content as "hip hop" underscoring the genre's connection to the systemic challenges that pervade the lived experiences of many black artists. Further insights can be found in Michelle Alexander's book [5]. Finally the model's decision is influenced by the presence of African American Vernacular English (AAVE) in lyrics like "yall" and "outta".

While the above themes are undoubtedly influential in the model's decision, our analysis reveals a curious trend. The model appears to prioritize the presence of profanity, racial slurs and misogynistic language in this classification task. This is evident in the first example, with the prevalence of the racial slur with origins in the African slave trade, followed by a word often used to disparage, women which appears on the first two examples. Additionally, other curse words such as "shit" and "Ass" influence the model's classification towards deciding this category. These findings hint at the fact that hip hop music has a history of objectifying women [3] and almost always contains explicit content.

This is also confirmed in our global aggregations analysis. The 4 most influential words for predicting "hip hop" can be seen in Figure 5.3.6 and are all considered offensive language. The rest of the words do not seem to sway the model's decision as much as the first 4. Although not as impactful these words other than profanity included cultural references, like "Brooklyn", hip hop's influential artist Jay-Z's nickname, "mic", "flow" referring to the rhythm and rhyme style of an artist and "bars". Other themes prevailing in our results are Street Life and Social status, with words like "hood", "police", "block", "45" invoking the .45 caliber firearm, "baller" and "pimp".



(a) Bool, Balm & Bollective by YG, correctly predicted as "hip hop" with a lot of certainty.

(b) Poontang Boomerang by Steel Panther, a song with label "heavy music" but predicted as "hip hop".

(c) Lucid Dreams by Juice WRLD, a song predicted as "pop" but belongs to class "hip hop".

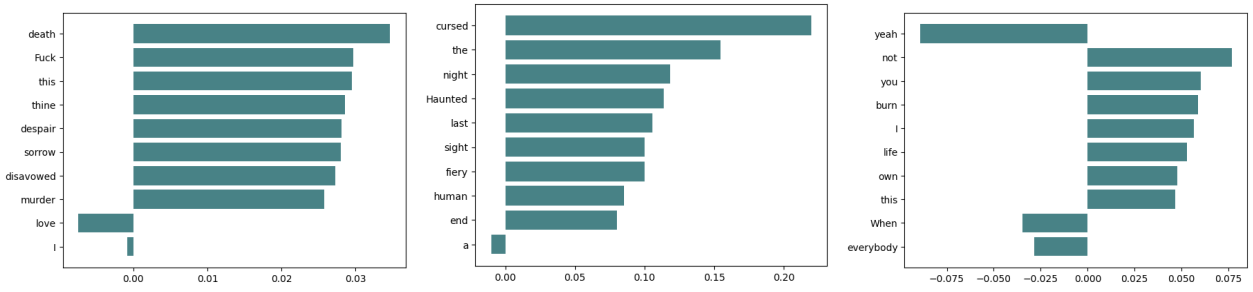
Figure 5.3.1: Local explanation for class "hip hop" for three instances of the test set: (a) is a true positive (b) is a false positive and (c) is a false negative. These graphs depict which features (words) contribute most to the class "hip hop".

Heavy Music

Following the model's success in "hip hop", the next best performer is "heavy music" class. This class includes hardcore, heavy metal and industrial metal music with their distinct identity and themes. Such songs give away their genre through their lyrics, that delve into darkness, rebellion, social and political critique and dystopian visions.

To investigate model behaviour we present again one true positive, one false positive and one false negative instance for this class. The respective songs and their artists are Beautiful Mourning by Machine Head, Sole Survivor by Blue Öyster Cult and Fiction by Avenged Sevenfold. The model is fairly certain, with a probability greater than 0.9, that the first two instances belong to "heavy music" while it misclassifies the last song as "rock".

The results of the local explanations, depicting the feature weights for class "heavy music" can be seen in Figure 5.3.2. Words talking about death, emotional pain and darkness cause the model to skew its predictions toward "heavy music", as can be seen on the first two examples. It is noticeable that yet again profanity plays important role for this class. The last instance does not contain any of those themes and therefore is misclassified by the model as rock. Positive emotions like "love" signal to the model that the instances likely do not belong to "heavy music".



(a) Beautiful Mourning by Machine Head, correctly predicted as "heavy music" with a lot of certainty.

(b) Sole Survivor by Blue Öyster Cult, a song with label "rock" but predicted as "heavy music".

(c) Fiction by Avenged Sevenfold, a song predicted as "rock" but belongs to class "heavy music".

Figure 5.3.2: Local explanation for class "heavy music" for three instances of the test set: (a) is a true positive (b) is a false positive and (c) is a false negative. These graphs depict which features (words) contribute most to the class "heavy music".

In our attempt to aggregate global features from local explanations, we recognize some patterns for the genre

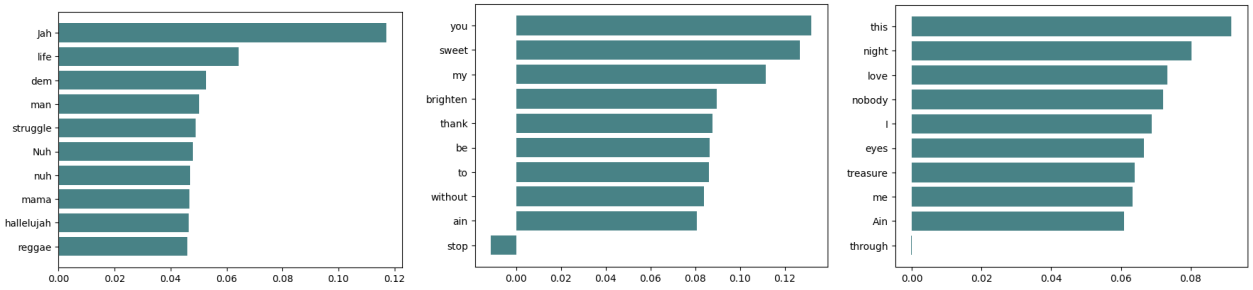
lyrical model. By investigating the most relevant features for the model's decision, some of which can be seen in Figure 5.3.6, several themes arise that align with the heavy music genre lyrics' thematology. To begin with words like "blood", "wrath", "destruction, "killer" and "torture" suggest that themes including violence, aggression and conflict that are part of heavy music[10] influence the model to decide that a song belongs in this class. This violent pattern extends into the emotional realm, where words such as "fear," "misery," "grief," "despair," "sorrow,", "hatred" and "emptiness" indicate emotional pain, distress and suffering. Furthermore, not only death and the macabre are detected by the model, e.g. "dead," "grave," "death" "rotting" and "cremation" but also occult and religious themes as demonstrated by the features "evil," "cursed," "soul," "hell," and "prophecy". Finally, heavy music is characterised by rebellious, political and social themes which are highlighted in our analysis with words like "refuse," "rise," and "unrelenting". Finally, in our observations, apart from "darkness" and "nightmare" which yielded a high weight from the global analysis, most other features were similarly influential. In summary, heavy music's distinct dark themes are detected by our lyrical model and boost it's performance in comparison with other classes.

Rhythm Music

The "rhythm music" class as mentioned in previous chapters, encapsulates not only blue note music (Gospel, Jazz and Blues) but also Jamaican/Reggae and R&B. These genres can cover a wide range of subjects ranging from slavery and oppression to religious hymns. Our model, although not ass successfully as the previous classes, seems to distinguish some of the patterns included in the lyrics of such music.

For this class we present the local explanations of Skankin' Sweet by Chronixx, How Sweet It Is (To Be Loved by You) by James Taylor and Ain't Nobody by Rufus. The first instance is the true positive, the second one is "rock" but the model thinks it is "rhythm music" and the final song was misclassified as "pop" although it belongs to the "rhythm music" class. The most import features that influence the model towards deciding "rhythm music" are depicted in Figure 5.3.3.

From this local analysis we can derive the following: The word "Jah" which is the Rastafarian name of God leads the model with certainty to guess "rhythm music". This word is present in many Jamaican and specifically Reggae lyrics. We also see that "dem" which is a variant of the English word "them", frequently used in Jamaican Patois also similarly influences the model. Apart from the word to praise God "hallelujah", "struggle" which might refer to the struggle of being a person-of-color and "reggae" which gives away the song genre, not many other words seem to fall under similar thematic categories.



(a) Skankin' Sweet by Chronixx, correctly predicted as "rhythm music" with a lot of certainty.

(b) How Sweet It Is (To Be Loved by You) by James Taylor, a song with label "rock" but predicted as "rhythm music".

(c) Ain't Nobody by Rufus, a song predicted as "pop" but belongs to class "rhythm music".

Figure 5.3.3: Local explanation for class "rhythm music" for three instances of the test set: (a) is a true positive (b) is a false positive and (c) is a false negative. These graphs depict which features (words) contribute most to the class "rhythm music".

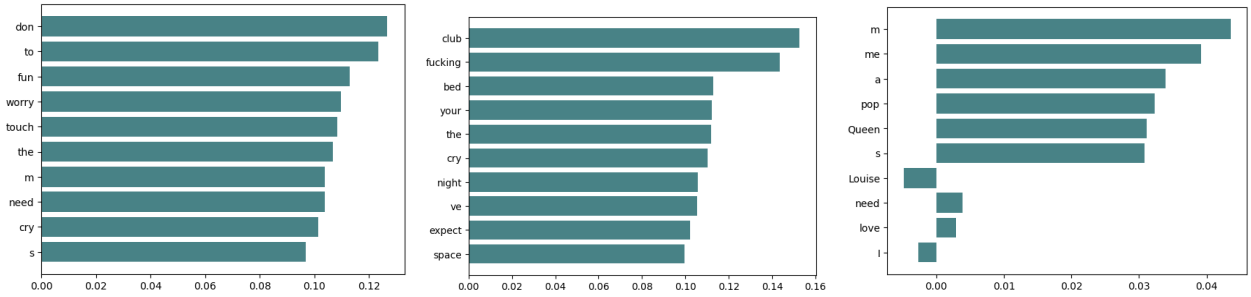
To further investigate how the model behaves around this class we also investigate the global aggregations from local explanations. We find that the word "Jah" has significantly more weight than other features. However among these features we can detect some thematic categories that also match with the "rhythm music" class. To begin with we find words with religious context such as "gospel", "pray", "saviour", "Galilee" and "lord"

that most likely appear in gospel music. Furthermore we note the presence of Jamaican Patois or Caribbean Dialect through words like "dem", "jammin", references of "jamaica", "bout" and "ya". The feature "Jah" belongs in both of the previous categories, which might explain it's high weight. Other categories include geographic references such as "Mississippi" and "Tennessee" with roots in blues music and social issues such as "ghetto" and "poor". In short, the model can easily recognize Jamaican and Gospel music but struggles to pinpoint other genres that belong to the broad label "rhythm music".

Pop

The model also performs decently when trying to find instances that belong to the class "pop", with F1-score of 0.51 for the test set. Pop music often explores themes of romance and relationships, celebrating love, heartbreak, and emotional connections. It also focuses on dancing and partying, capturing the energetic atmosphere of nightlife. Emotional experiences, including personal empowerment and resilience, are frequently highlighted. Additionally, pop music often emphasizes fashion, glamour, and the modern lifestyle, reflecting contemporary culture and trends.

To investigate if the model can capture those themes we present in Figure 5.3.4 Just a Touch by AlunaGeorge which was correctly labeled as pop, Expectations by Lauren Jauregui which was misclassified as pop and Bubble Pop Electric Gwen Stefani which was mislabeled as "rhythm music". Based on the local explanations, the model appears to capture several key themes of pop music, such as emotional experiences with words like "cry", "love" and "worry", nightlife and fun with words like "club", "fun" and "night" and relationships and intimacy with words like "touch" and "bed".



(a) Just a Touch by AlunaGeorge, correctly predicted as "pop" with a lot of certainty.

(b) Expectations by Lauren Jauregui, a song with label "rhythm music" but falsely predicted as "pop".

(c) Bubble Pop Electric Gwen Stefani, a song falsely predicted as "rhythm music" but belongs to class "pop".

Figure 5.3.4: Local explanation for class "pop" for three instances of the test set: (a) is a true positive (b) is a false positive and (c) is a false negative. These graphs depict which features (words) contribute most to the class "pop".

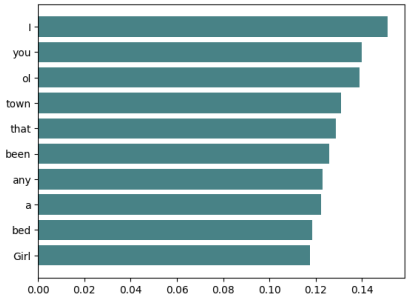
The global aggregations also bring forward some features that could be semantically grouped. To begin with, the most influential group would be romance and relationships with words like "kiss", "girlfriend", "sexy", "boy", "heart", "someone" and "beauty" showing attraction, love and emotional connections. Another important group for the model's decision is dancing and partying with words like "club", "dance", "dancin", "disco", "bars" and "vegas" having relatively high weights. We finally discern words hinting emotional experiences lie "forever", "heart", "hoping" and "cry" as well as themes of fashion and style like "dress", "chic", "glitter" and "glow". The model's focus in these themes could explain not only why the model can adequately capture pop songs but also why these songs are sometimes confused with rock and alternative rock songs which might have common themes.

Folk

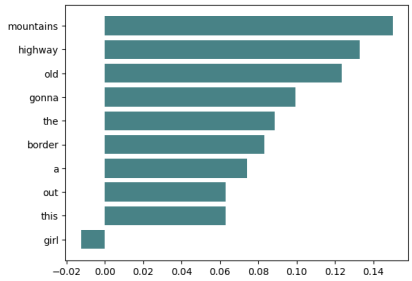
The next class worth analyzing, in order to understand the model's behavior is "folk". This class contains the folk and country genres under a common label. Folk and country music encompass a wide variety of themes, reflecting diverse aspects of human experience. Folk music often delves into social and political

issues, preserving cultural traditions and heritage, and recounting stories of work and labor. It also explores personal themes of love and relationships, as well as a deep connection to nature and the environment. Country music, on the other hand, frequently focuses on everyday life, romantic love and heartache, and expressions of patriotism and national pride. It also addresses themes of resilience and perseverance, social gatherings and escapism, and spirituality and faith.

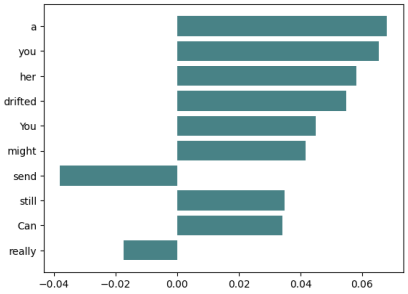
The local explanations of 3 instances are shown in Figure 5.3.5. Any Ol' Barstool by Jason Aldean was correctly labeled as "folk", Key To The Highway by Eric Clapton was misclassified as "folk" and Twinkle, Twinkle Lucky Star by Merle Haggard was misclassified as "rhythm music" while it is "folk". From the first two instances we see that themes regarding small town life, personal relationships and nature might be detected by the lyrical model and classified as folk music. On the other hand we see that the last instance does not contain such themes and therefore is misclassified.



(a) Any Ol' Barstool by Jason Aldean, correctly predicted as "folk" with a lot of certainty.



(b) Key To The Highway by Eric Clapton, a song with label "rhythm music" but falsely predicted as "folk".



(c) Twinkle, Twinkle Lucky Star by Merle Haggard, a song falsely predicted as "rhythm music" but belongs to class "folk".

Figure 5.3.5: Local explanation for class "folk" for three instances of the test set: (a) is a true positive (b) is a false positive and (c) is a false negative. These graphs depict which features (words) contribute most to the class "folk".

Our global aggregation analysis revealed that no single feature can definitively lead the model to predict this class with certainty, since all feature weights are similarly low in value. However it is clear that to identify this class the model recognizes themes realted to nature and environment with words like "river", "pastures", "mountain", "woods", "trees", "leaves" and others, celebrating landscapes, flora and fauna. Moreover the model recognizes themes regarding personal relationships with words such as "hurt", "Mary", "fell" and "pretty" and small town life with words like "county", "town", "barley", "bakers" and others. Finally, words like "Bethlehem", "Belfast" and "Government" suggest a connection to historical events. Those themes are not exclusive to this category which could explain why the model does not decide "folk" with certainty when coming across these features.

Alternative Rock, Rock, Electronic and Punk

Although the model performed relatively well in predicting the previous classes and found themes present in the lyrics of each class, it struggles to find instances that belong in "alternative rock", "rock", "electronic" and "punk" with each class's F1-score being lower than 0.3. The model confuses "alternative rock" music with "rock", "punk" music with "heavy music", "pop", "rock" and "alternative rock", "electronic" with "pop" and "alternative rock" and finally "rock" with "alternative rock", "heavy music" and "pop. In Figure 5.3.7 we present two examples. In the first example we present the t-SNE scatter plot of the word embeddings for the 30 most influential features for "hip hop" and "heavy music". As can be seen inf Figure 5.3.7a the themes that the model consider relevant for each class are clearly separable with a bit of confusion when it comes to profanity. This fact justifies the model's ability to distinguish between those classes. On the other hand, the second example includes a similar plot for features that influence "alternative rock" and "rock". As can be seen in Figure 5.3.7b the themes for each of these classes are not easily distinguishable from each other and the model confuses "alternative rock" with "rock" instances and vice versa.

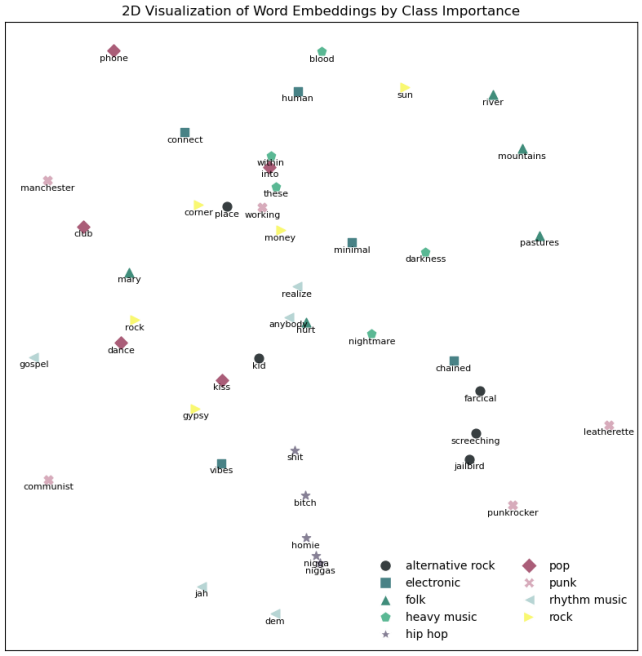
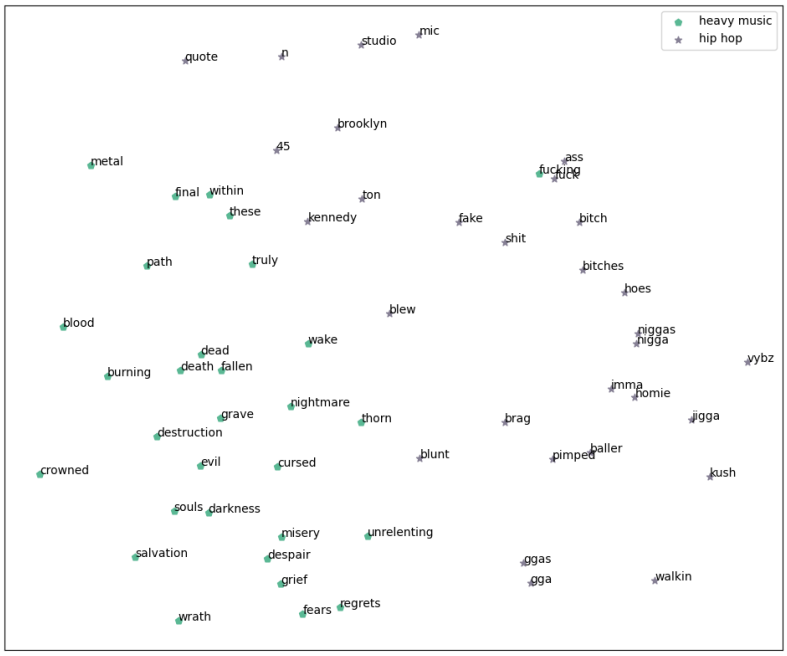
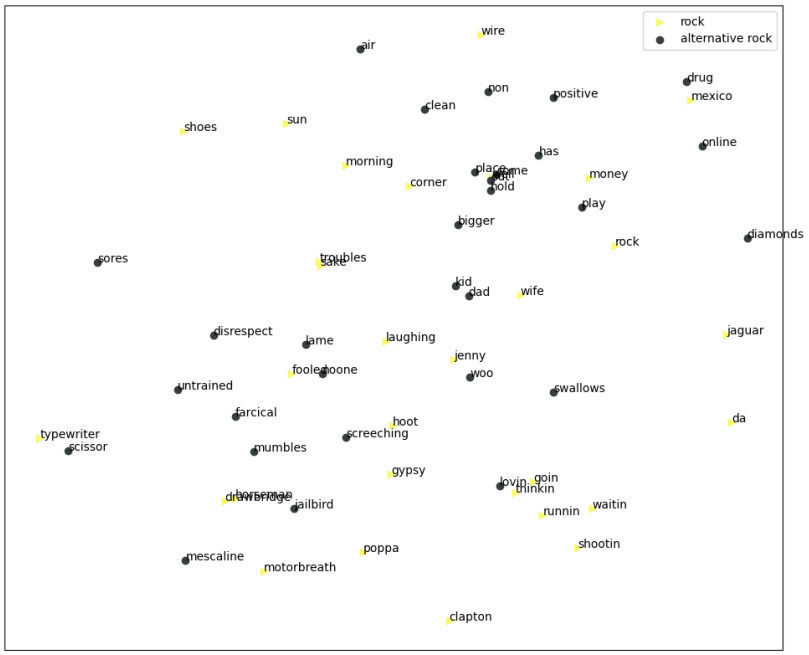


Figure 5.3.6: Top 5 features of global aggregates of local explanations for each genre class.



(a) The 30 most influential features for "hip hop" and "heavy music" according to the model. The themes are clearly separable.



(b) The 30 most influential features for "alternative rock" and "rock" according to the model. The themes are not clearly separable.

Figure 5.3.7: The t-SNE plot of glove embeddings for the 30 most influential features for each class. In fig. (a) we present the features of "hip hop" and "heavy music" and in fig. (b) the features of "alternative rock" and "rock".

Some amount of confusion by the model is to be expected. This can partially be explained by the fact that a single song can blend elements from various genres, making it challenging to classify it into just one category. This can lead to a dataset that contains noisy instances and, consequently, a model that learns from this noisy data. Moreover, while certain themes are unique to specific genres, the majority of semantic groupings are shared across multiple genres. For example, songs about love can have similar lyrics but differ in rhythm, timbre or even musical scale and therefore genre. Our global aggregate analysis for these genres shows that, although some themes are present among the important features of each class (eg. "punkrocker", "communist", "worker" and "manchester" for punk), the presence of generic and ambiguous words, combined with a lack of strong imagery and themes, indicates that the model's ability to identify those kinds of music instances is limited. We conclude that lyrical thematology is only one part of a musical creation. Classifying music into genres solely on lyrical content can be accurate for certain genres but in general it does not, and should not, yield highly accurate results.

5.3.2 Lyrical Emotion Model Explanations

Although the genre explanations provide interesting results, as far as the model behaviour is concerned, the analysis of the emotion lyrical model didn't yield similar results. In an analogous manner as before we generate local explanations and global aggregates in order to get a grasp of the model's decision making process and shed light to the reasons of it's poor performance. This leads us to the observation that the model recognizes dark and macabre themes, with words like "dead", "destroy" and "bleed", and attributes them to the "Angry" class. This behaviour justifies the relatively high F1-score of the model for this class. Similarly but less accurately the model recognizes party and fun themes and attributes them to the "Excited" class and religious themes, attributing them to the "Calm" class. However the model performs poorly in recognizing themes related to the other classes. The weights of the global aggregate of features influencing the model to predict those classes were low, indicating the model's undecidability. This difficulty in making decisive classifications for this task is also hinted by the fact that even the highest probability scores during the evaluation on the test set for these classes are low. In Figure 5.3.8 we present the 5 most influential features for each class. We should note that some features that are expressive sounds like "Mmm" or names like "Bethlehem" do not have a GloVe embedding and therefore are not depicted in the Figure.

The model's inability to perform well on this task can be attributed to several factors. Firstly, the emotion recognition is inherently complex and subjective, heavily depending on human experiences and interpretations, which vary widely among individuals. Even though stronger words might occasionally provide clear indicators of the highly energetic classes "Angry" and "Excited", music is a multifaceted art form. Lyrics, while important, represent only one dimension of music creation. Critical information for accurately classifying music often lies beyond the lyrical content, encompassing elements like melody, harmony, rhythm, and tempo, which are not captured in this domain. Since the training and test data were annotated considering the whole music creations and not just the lyrical corpora of songs, it is to be expected that instances that convey a certain emotion, but their audio expresses a different emotion, add significant confusion to the model. Consequently, the model struggles to make decisive and accurate predictions which is evident by the inability to find common themes for each class.

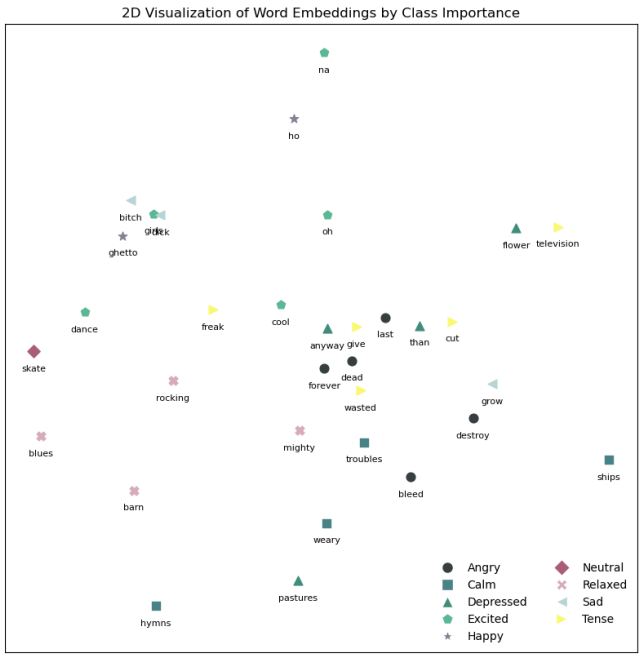


Figure 5.3.8: Top 5 features of global aggregates of local explanations for each genre class. Words representing town names might not be included.

5.3.3 Audio Genre Model Explanations

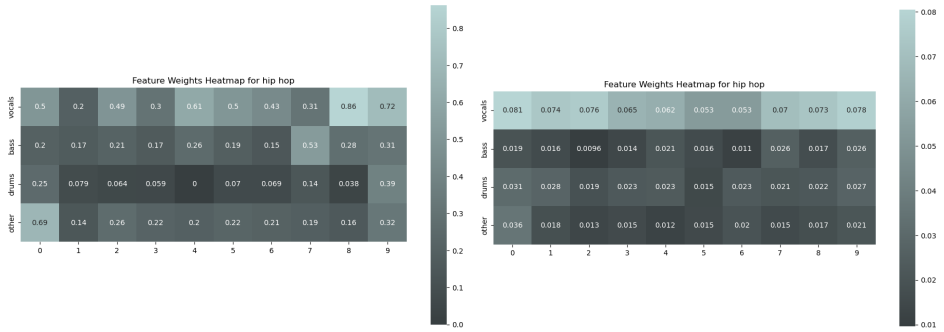
The main objective of this subsection is to determine what audio features influence the audio model to make it's decisions and evaluate and discuss if those features are indeed relevant for the specific genre. As mentioned in the previous chapter, the local explanations in the audio domain are listenable. The XAI technique implemented divides the audio into 10 temporal segments and separates it into vocals, drums, bass, and other components. For example, a listenable explanation labeled as 'vocals6' represents the 7th (indexed from 0 to 9) temporal segment of the vocals of the song. Some explanations can be listened to at our GitHub repository. We also provide two methods of aggregating local explanations, I_{cj}^{H} and I_{cj}^{AVG} as defined in the previous chapter.

Hip Hop

Surprisingly, the only class that performs worse in the audio domain is "Hip Hop." Aside from the prevalent lyrical themes described previously, "Hip Hop" is also distinct in its audio characteristics. These include rhythmic beats often syncopated to create a head-nodding groove, strong and driving drums with a kick laying down the main pulse, and deep bass lines providing a low-frequency thump. The vocals in "Hip Hop" are dominated by rhythmic and rhyming speech (rapping), with styles ranging from fast-paced and aggressive to slow and poetic, depending on the artist and the mood of the song. For instance, the legendary rapper Eminem is known for his rapid-fire delivery and intricate wordplay, while artists like Kendrick Lamar often utilize a smoother flow with a focus on conscious lyrics. Other elements of "Hip Hop" might include sampling, beatboxing, and turntablism.

In order to understand what is important for the model's decision for this class, we identify a prototype from the test set. The prototype instance is the song Statue of Limitations by 2 Chainz. By generating local explanations and analysing the results, we find that the most important features for this prototype are the

vocals dominating the first 10 spots in the feature importance leader board, with the first two most important vocal features singing the lyrics "people's girl" and "ex-athlete". The global explanations for this class show a similar pattern. As can be seen in Figure 5.3.9 vocal features seem to consistently influence the model when predicting "hip hop", while drum and bass features are not as important. This fact indicates not only the continuous presence of vocals throughout the song but also that the model successfully detects rapping in the those segments and therefore classifies the instance as "hip hop".



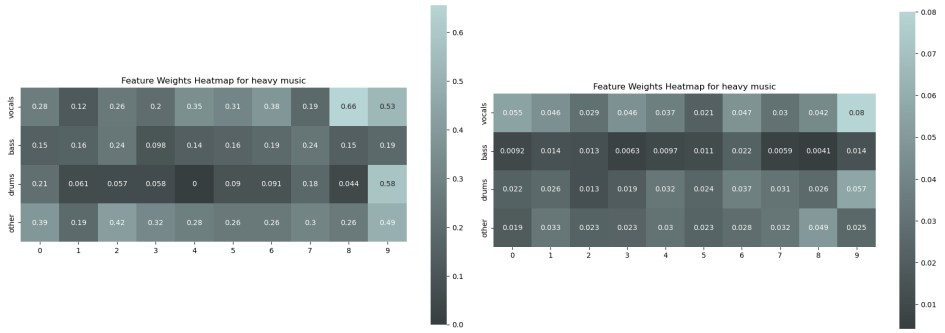
(a) Homogeneity importance weights (b) Average weights

Figure 5.3.9: Global aggregates heatmap for class "hip hop". The first heatmap refers to the homogeneity importance weights whereas the second heatmap refers to the average lime weights per feature.

Heavy Music

The model performs very accurately when predicting metal, hardcore and industrial music under the umbrella class "Heavy Music". Characterized by its aggressive and intense audio features this class, includes heavily distorted guitars often featuring heavy riffs and extended solos, rapid and complex drumming with a focus on double bass and blast beats, and thick and distorted bass lines that add depth. The vocals range from growling and screaming in metal and hardcore to more robotic and processed sounds in industrial music. The use of unconventional instruments, synthesized effects, and dark atmospheric elements also distinguishes industrial from other genres.

The prototype for this class is the song Tread Lightly by Mastodon. Yet again the local explanations attribute high weights to the vocal segments, with 6 of them being in the top 7 most impactful features. The feature with the highest weight is a shout of the lyrics "when there is", that is in higher pitch compared to the rest of the vocals. We can also find a drumming pattern with high weight, as well as segments labeled as other which mainly contain guitar power chords. The rest of the vocal segments that were not attributed a high weight are mostly silent. Again we can generalize these observations by analysing the global aggregates presented in Figure 5.3.10. We find that vocal and other features influence the model the most, as indicated by the relatively high homogeneity importance weights. Drum and bass features seem to be more influential in this class compared to the previous one, however the model's focus on vocals and other features suggest the ability to detect shouts, screams and distorted guitars which are important elements of this class.



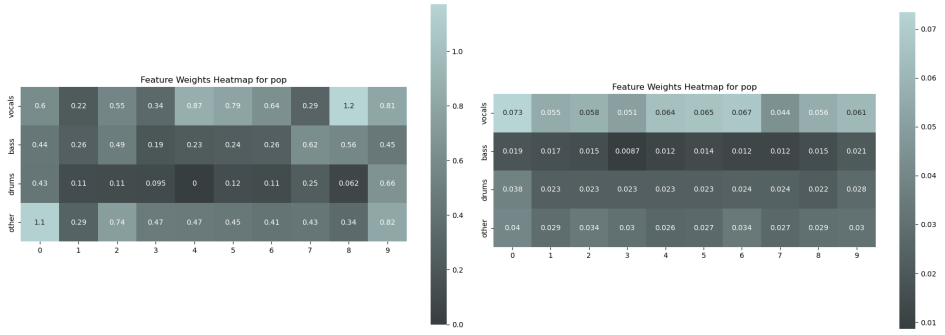
(a) Homogeneity importance weights (b) Average weights

Figure 5.3.10: Global aggregates heatmap for class "heavy music". The first heatmap refers to the homogeneity importance weights whereas the second heatmap refers to the average lime weights per feature.

Pop

The model also generates better results when predicting "pop" instances. Pop music is characterized by its catchy melodies, simple chord progressions, and polished production. It often features a verse-chorus structure, with hooks that are easy to remember and sing along to. The beats are typically straightforward, with a focus on danceable rhythms and a strong emphasis on the vocals. Pop songs frequently incorporate electronic elements, such as synthesizers and drum machines. The genre is known for its broad appeal and often blends elements from various other musical styles.

The prototype for this class is the song I Love You Always Forever by Betty Who. Once more, vocal features are important for the model's decision for this instance. The feature with the highest weight is the vocals singing the lyric "you've" from the third verse of the song and seems to be very impactful compared to the rest of the features. Some drum features also seem to play an important role, while bass features are the least impactful. From the global explanations we observe the importance of vocal features regardless of their temporal order. In comparison to the prototype instance we find that drums have a minimal impact on the model's predictions. Other features sometimes play a significant role in detecting instances that belong to the pop class.



(a) Homogeneity importance weights (b) Average weights

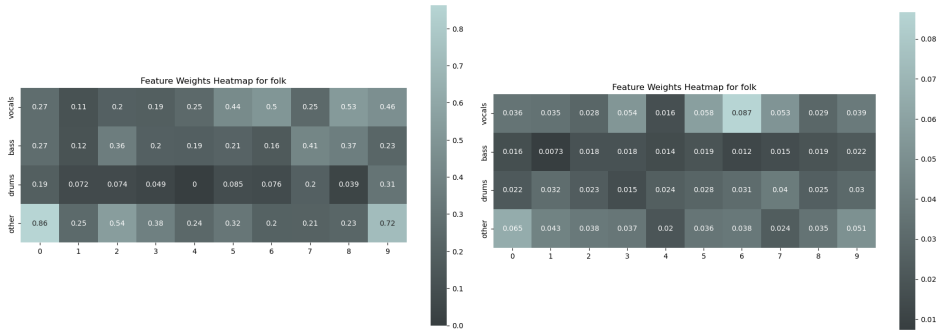
Figure 5.3.11: Global aggregates heatmap for class "pop". The first heatmap refers to the homogeneity importance weights whereas the second heatmap refers to the average lime weights per feature.

Folk

The model F1-score for this class is 0.54, increased by 0.12 compared to the lyrical model. Folk music, including country, is characterized by its acoustic instrumentation and storytelling aspect. Traditional folk often features acoustic guitars, banjos, fiddles, and harmonicas, creating a rustic and organic sound. Country music builds on this foundation with additional elements like pedal steel guitars and prominent, twangy

vocals. The melodies are typically simple and memorable, with a strong emphasis on lyrical content and the vocals are typically natural and unprocessed. Drums can range from simple and brushed to more prominent backbeats depending on the subgenre.

The prototype for this class is the song Throw It All Away by Brandi Carlile. Guitar elements heard in "other" features and vocals are the most important influences for this class, with drums following suit while bass features do not rank highly for this example. The global aggregates showcase vocal and other features as impactful, suggesting the ability to detect banjos, harmonicas and of course guitar sounds prevailing in folk music. The different methods to generate global aggregations disagree as to whether bass or drums influence the model the most.



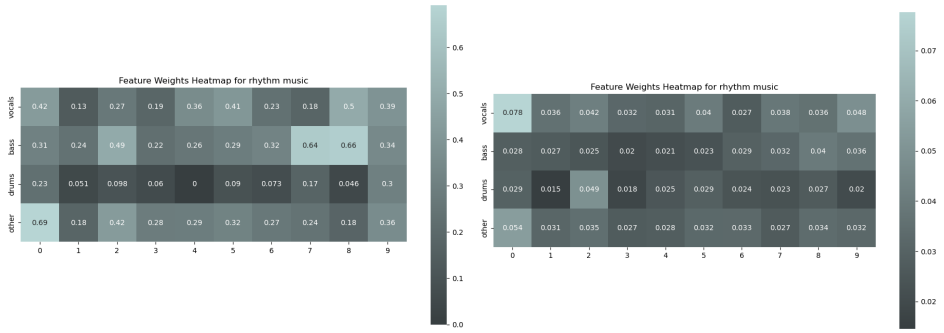
(a) Homogeneity importance weights (b) Average weights

Figure 5.3.12: Global aggregates heatmap for class "folk". The first heatmap refers to the homogeneity importance weights whereas the second heatmap refers to the average lime weights per feature.

Rhythm Music

"Rhythm music" includesR&B, jazz, blues, gospel, and Jamaican genres. R&B often features smooth, melodic vocals over groovy bass lines and syncopated drum patterns. Jazz is marked by improvisation, complex chord progressions, and instrumental solos, typically featuring saxophones, trumpets, and pianos. Blues music highlights emotive guitar riffs, often using the pentatonic scale, with a strong backbeat. Gospel is distinguished by powerful vocal harmonies and organ or piano accompaniment, while Jamaican music, such as reggae, emphasizes offbeat rhythms and deep, resonant bass lines.

As a prototype for this class Power To The People - Demo Version by Curtis Mayfield was selected. The most important feature according to the local explanations appear to be vocal, however the rest of the most impactful features appears to be a mixture of vocals, bass and other features with emphasis on bass features. The global aggregations show that drums are not influential for this class but do not highlight other features as more important. In short the model seems to detect the presence of basslines in R&B instances, such as our prototype instance, as well as the intricate vocals, saxophones and pianos of other subgenres in this class. The global aggregates are seen in Figure 5.3.13.



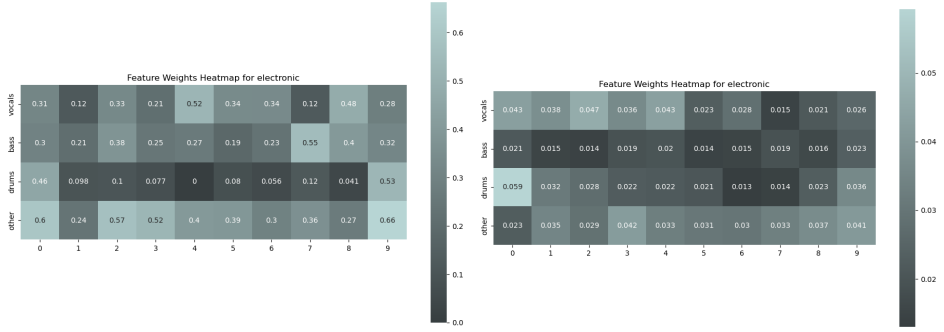
(a) Homogeneity importance weights (b) Average weights

Figure 5.3.13: Global aggregates heatmap for class "folk". The first heatmap refers to the homogeneity importance weights whereas the second heatmap refers to the average lime weights per feature.

Electronic

The audio model performs significantly better at recognizing "electronic" instances compared to the lyrical model, with a 0.22 increase in the F1-score. Electronic music is defined by its use of synthetic sounds, created and manipulated through digital and analog electronic instruments. It includes sub-genres like techno, house and trance, each with its own distinct characteristics. The music is often built around repetitive rhythmic patterns, deep bass lines, and layered synthesizer textures. Electronic tracks frequently feature effects such as reverb, delay, and modulation to create immersive soundscapes. The genre is known for its danceability, with beats ranging from the steady four-on-the-floor patterns of house music to the more complex and faster rhythms of techno and drum and bass. Some electronic subgenres, like ambient or techno, may forgo vocals entirely and when vocals are present they can be processed and manipulated with effects like autotune, vocoders, or pitch-shifting to create unique timbres and textures.

Our prototype from the test set is the song Colourless Colour by La Roux. Although this instance's original label is 'electronic', one could also argue that it belongs to the 'pop' category, highlighting the correlation between these two genres. According to the local explanations, the most important features for this instance are "other" features. Those include the distinct synthesizer melodies present in this song. Once again, the rest of the features seem to be of equal importance with an exception of drum features that have a lower weight. This behaviour is present in the weights of global aggregates for this class presented in Figure 5.3.14



(a) Homogeneity importance weights (b) Average weights

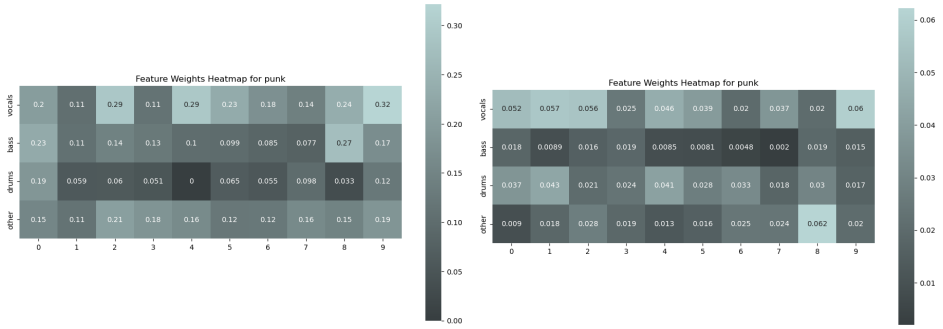
Figure 5.3.14: Global aggregates heatmap for class "electronic". The first heatmap refers to the homogeneity importance weights whereas the second heatmap refers to the average lime weights per feature.

Punk

This class shows a great 0.2 increase in the F1-score compared to the lyrical model. Although the audio model can distinguish better between "punk" and "pop" instances it still missclasifies "punk" songs as

"rock", "alternative rock" and "heavy music". Punk music is known for its raw, energetic, and rebellious sound. It typically features fast tempos, short song durations, and simple, power-chord-based guitar riffs. The drumming is aggressive and straightforward, often featuring rapid-fire snare hits and cymbal crashes. Bass lines are typically simple and follow the root notes of the chords. Vocals in punk are often shouted or delivered with a snarly, confrontational tone.

The prototype song for this class is It'll Be A Long Time by The Offspring. All features in the explanations appear to be relatively important, with 3 vocal features having the heavies weights. Those features present drum, bass and guitar patterns unique for this class. The global aggregates show an emphasis on vocal segments, with the rest of the features having similar importance weights.



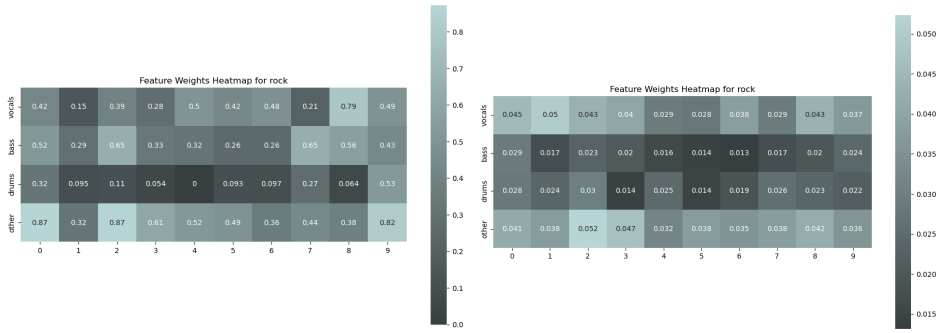
(a) Homogeneity importance weights (b) Average weights

Figure 5.3.15: Global aggregates heatmap for class "punk". The first heatmap refers to the homogeneity importance weights whereas the second heatmap refers to the average lime weights per feature.

Rock

Rock music, encompassing rock 'n' roll and golden age rock along with their various subgenres also has better metrics compared to the lyrical model. This class is characterized by the power of the electric guitar, often featuring distinctive guitar riffs that lay the foundation for the song. For example Led Zeppelin's "Immigrant Song" or Guns N' Roses' "Sweet Child o' Mine" wouldn't be the same without their iconic opening riffs. Soaring guitar solos are another element, adding moments of virtuosity and excitement. While punk's raw energy might influence some rock subgenres, classic rock generally leans towards a more polished and layered sound with the rhythm section providing a solid foundation for the guitars to take center stage. Furthermore, a strong rhythm section of bass and drums create a driving beat and vocals can range from smooth to more powerful singing.

Broadway by Goo Goo Dolls serves as the prototype of the test set for this class. As expected, other features that showcase guitar chords mostly influence the model and vocals appear as the second most influential. The heatmaps if Figure 5.3.16 also present other and vocal features as more important for this class, highlighting the fact that guitars take center stage in this kind of music.



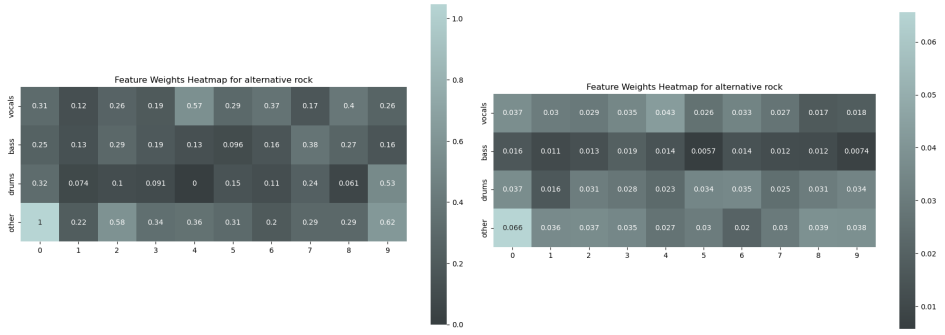
(a) Homogeneity importance weights (b) Average weights

Figure 5.3.16: Global aggregates heatmap for class "rock". The first heatmap refers to the homogeneity importance weights whereas the second heatmap refers to the average lime weights per feature.

Alternative Rock

The audio model performed just as badly at predicting "alternative rock" songs as the lyrical model. Although more precise, the model misclassifies more "alternative rock" songs as "rock". This genre is diverse and blends elements from punk, post-punk, and mainstream rock, often characterized by its experimental approach to sound and structure. It features the electric guitar as a core element, but it's artists tend to be more adventurous than their classic rock counterparts, experimenting with a wider range of influences. This music might feature distorted guitars reminiscent of punk or metal, but with a distinct alternative rock flavor. The rhythm section can be quite diverse as well, with some bands maintaining a strong rock foundation with prominent drums and bass lines and others exploring more nuanced and textured rhythms, creating a less predictable and more atmospheric soundscape. Vocals in alternative rock are just as varied ranging from more powerful and emotional vocals to use a more dissonant or spoken-word approach. What mainly distinguishes this class from "rock" is its experimental approach to sound and production, often featuring a raw, unpolished aesthetic.

Stephens's Malkumus Jo Jo's Jacket is chose as the prototype for this class. Some drumming patterns seem to influence the model the most for this instance but in general drum, vocal and other features (which once again contain guitar sounds) seem to equally contribute to the model's "alternative rock" decision. This pattern is also present in the global explanations. While the model seems to decently capture some elements of alternative rock, it cannot distinguish them from rock elements. This leads us to believe that the audio model lacks the capability to assess production quality, which is a key difference between the two genres. As can be expected, the model cannot identify the indie origins of a song, and the overlap in musical elements between rock and alternative rock further complicates accurate classification.



(a) Homogeneity importance weights (b) Average weights

Figure 5.3.17: Global aggregates heatmap for class "alternative rock". The first heatmap refers to the homogeneity importance weights whereas the second heatmap refers to the average lime weights per feature.

To conclude, the audio model performs better in every class except "hip hop". By observing the explanations we find that the model indeed detects elements characteristic for each class whether it be the unique rhyming pattern present in "hip hop" vocals or the distorted guitar sounds in "heavy music". The class "alternative rock" does not perform better in the audio domain and it still gets mixed up with "rock" instances. The authors of this paper [43] find that "art styles that are confused the most are styles that share common characteristics" and that "Art movements inherited features from their predecessors and influenced their successors". Similarly in music, rock and alternative rock not only share common characteristics but also alternative rock is an evolution of classic rock. It is also important to mention that music categories are often mixed and the boundaries between them are relatively fluid as mentioned in this work [60].

5.3.4 Audio Emotion Model Explanations

The emotion audio model is significantly superior compared to its lyrical counterpart with certain classes having F1-score improvement of more than 0.3 units. In the global aggregates of our audio feature analysis, "happy" and "excited" classes both heavily rely on drum features, with "happy" also emphasizing vocals, suggesting a strong association between rhythm and positive, high-energy emotions. For "neutral" and "relaxed" classes, features named "other" dominate, with bass features notably less important for "neutral" emotions. We should note that since the model does not make many "relaxed" predictions on the test set, we do not have enough data to conclusively determine feature importance for Relaxed. In "angry" and "tense" emotions, other features are most significant, but "tense" also highly values vocals, indicating a complex blend needed to convey negative and high-strung feelings. The "sad" and 'depressed" classes show diverging patterns; drums are key for "sad" but negligible for "depressed", highlighting how energy levels influence feature relevance. Lastly, "calm" class uniquely values vocals and other features, with occasional importance of bass and drums, illustrating a varied audio component mix that aids in producing a soothing emotional effect. We finally observe that adjacent classes on the emotion map often consider similar features as important, particularly when they share similar levels of either valence or energy.

5.3.5 Multimodal Genre Model Explanations

In previous subsections we presented some elements that characterize each genre in the lyrical and in the audio/music domain. We found that each unimodal model can detect some of these and make accurate predictions. The lyrical model was less successful in general but could predict certain classes, with distinct thematology very accurately. The audio model on the other hand was better overall and was able to identify concepts that distinguish genres from one another. Here, we analyze the results of the explainability process so as to determine what features influence the multimodal model the most and if it is able to combine both modalities to be more accurate.

Generating global aggregates using the homogeneity-weighted importance does not accurately capture the influence of features class. This is because we have two types of features: words and audio features. While the audio features are the same 40 each time, word features can differ from instance to instance. This means that vocal, drum, bass and other audio features will similarly impact each class due to their different content. For example, an "other" feature could contain guitar power chords and contribute to the model deciding "heavy music" or it could contain saxophone and therefore influence the model to decide "rhythm music". This leads to audio features having high entropy meaning that audio feature attributions point to many classes. On the other hand, word attributions have lower entropy since they are less homogeneous. This leads to mistakenly consider audio features as less important. Therefore the global average importance is more suitable for the multimodal case I_{cj}^{AVG}.

The local explanations as well as the global aggregates present some promising outcomes. In classes where the lyrical model was more accurate than the audio model, the most impactful features for the multimodal model's decision appear to be the lyrical features, whereas in classes that the lyrical model underperformed, the model pays attention to the audio features. For example for the "hip hop" class the global aggregates, as presented in Figure 5.3.18, show clearly that lyrical features are more impactful. On the other hand, the model focuses more on audio features for "punk" instances, whose global aggregates are presented in Figure 5.3.19. In other cases like the "pop" class whose global aggregates are available in Figure 5.3.20 the audio and lyrical features equally influence the model. In general the multimodal approach manages to integrate and combine information from the unimodal models for superior results.

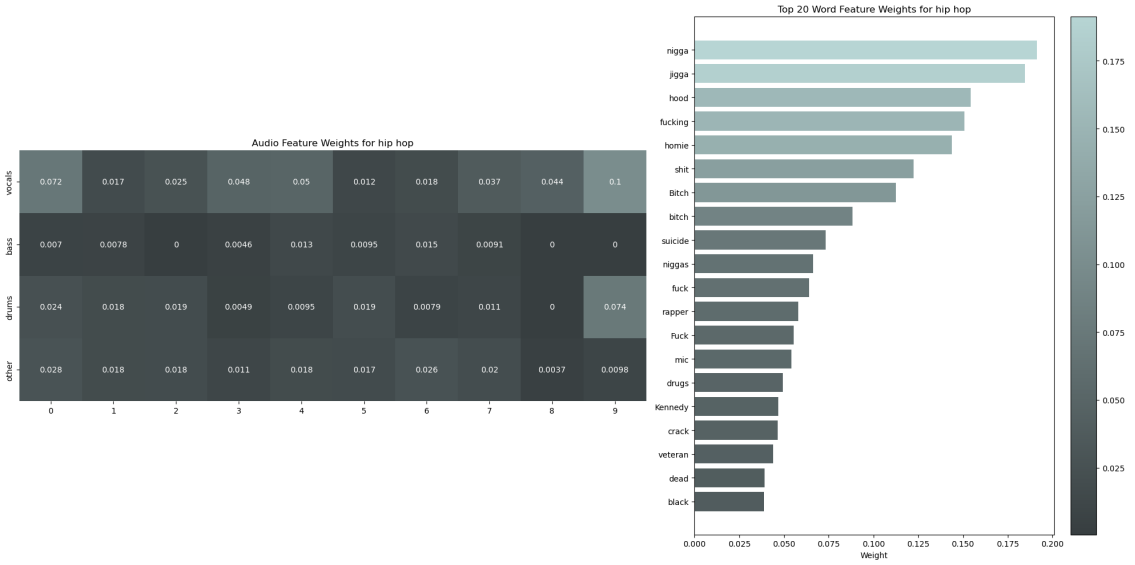


Figure 5.3.18: The global aggregates of local explanations for instances classified as "hip hop" for the multimodal model. The first heatmap depicts the weights of the audio features while the barplot shows the weights of the 20 most impactful word features.

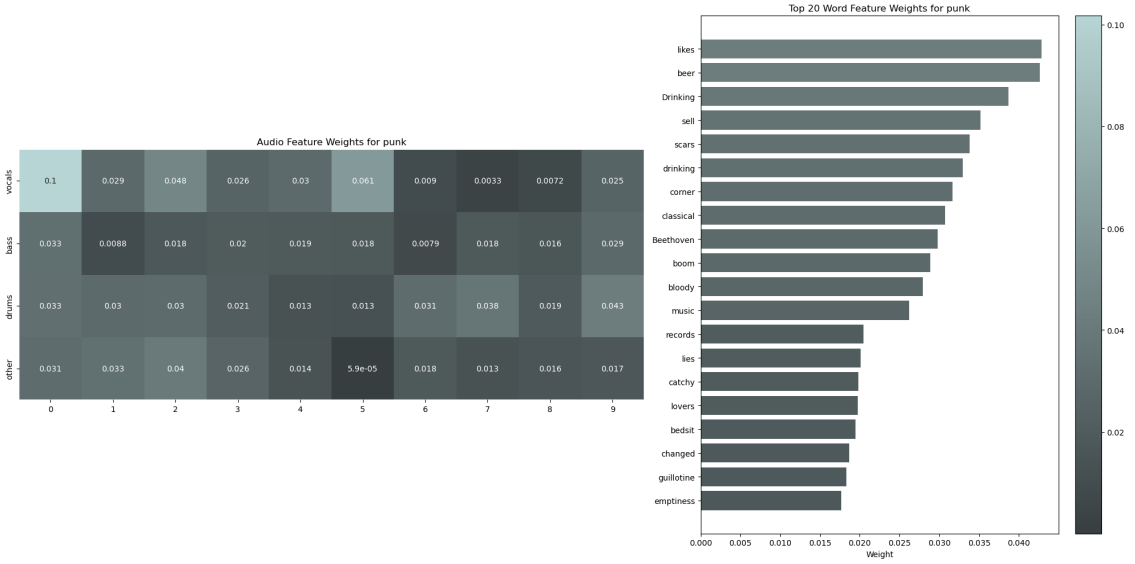


Figure 5.3.19: The global aggregates of local explanations for instances classified as "punk" for the multimodal model. The first heatmap depicts the weights of the audio features while the barplot shows the weights of the 20 most impactful word features.

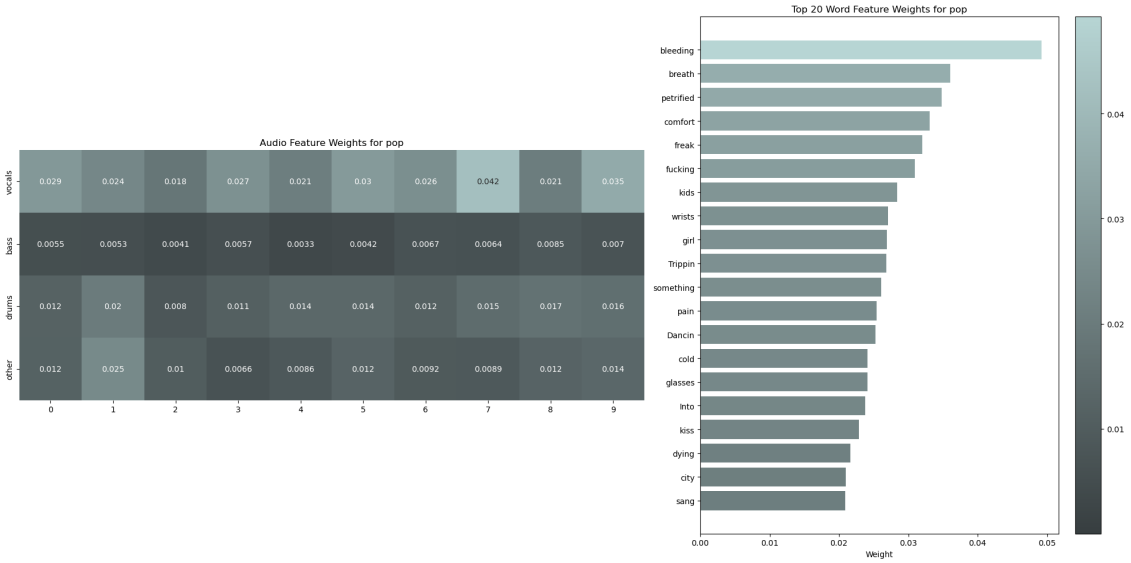


Figure 5.3.20: The global aggregates of local explanations for instances classified as "pop" for the multimodal model. The first heatmap depicts the weights of the audio features while the barplot shows the weights of the 20 most impactful word features.

5.3.6 Multimodal Emotion Model Explanations

For the reasons described in the previous subsection using the homogeneity-weighted importance does not accurately capture the influence of features for each class. It is clear that the multimodal model considers the audio features more important. Although the model has better performance than the unimodal approaches, the difference is marginal. We observe from the global aggregates that the multimodal approach considers audio features as more impactful across all classes. For the classes "happy", "depressed", "excited", "tense", "calm" and "relaxed" the top 10 features with the highest weights are all audio features. For instances labeled as "sad" and "neutral" only a few word features have high weights. For "angry" songs we find a combination of text and audio features as important, with the audio ones being once again more relevant, and yet the multimodal model's score does not exceed that of the audio approach.

Chapter 6

Conclusion

6.1 Discussion

In this work we explored genre and emotion, multimodal music classification. Our research begins by investigating various existing datasets that offer both lyrical and audio content, as well as genre and emotion labels. Since the Music4All dataset satisfies multimodality, enables the exploration of multiple tasks and is one of the biggest datasets of its kind, we proceed by analysing and evaluating its contents. We propose methods to group the dataset's instances into 9 emotion categories and 9 music genres and also augment the dataset due to it's emotion imbalances and clean some noisy instances as far as the genre labels are concerned. Although we went to great lengths in order to curate the M4A dataset, our final dataset is imbalanced especially for emotion labels and contains some ambiguous genre instances. This imbalance can be attributed to the fact that the distribution of music tracks in the respective industry favors certain categories, like "pop" or "angry" songs, and makes it more difficult to find "relaxing" and "calm" songs using SpotifyAPI. The genre ambiguity lies in the fact that many songs fall into multiple genres [60] and selecting one label to represent a track introduces confusion.

We continue our endeavours by studying related work on MIR tasks, namely music emotion regression. Although during this procedure we find recent SOTA technologies we choose to follow a transformer finetuning approach by creating 3 models for each classification task, a text model utilizing the pretarined roberta-large, an audio model with spectrograms as input based on AudioSpectrogramTrsnformer and a multimodal model combining the previous two. According to our research it is the first time those models are combined for music classification. This approach outperforms both of the unimodal ones, while the lyrical model was the worst performer for both tasks. In general, the classification of instances according to their genre was more successful than the classification according to their emotion label.

In order to understand the behaviour of our models we look for explainability methods in the text and audio/image domain. We find that Local Interpretable Model-Agnostic Explanations neatly fit our needs. Therefore we implement LIME explanations on the lyrical model, to capture themes relevant for each class. In the audio domain, instead of treating spectrogram as images and generate explanations for those, we implement audioLIME, an approach based on LIME that separates the audio into temporal segments and each segment into vocal, drums, bass and other components and provides us with listenable explanations. We introduce MusicLime, a way to combine the two methods and produce multimodal local explanations with text and audio as input. Finally we generate global aggregates of local explanations to have a more complete view for each task.

Our analysis yielded interesting results. Firstly, the lyrical explainability methods manage to capture certain themes for each class, particularly ones with strong words, profanity and death. However, classifying music only taking into account lyrical content is not particularly successful since music is multifaceted. Furthermore, the audio explanations, show that the audio model can capture music elements distinct for each class, such as shouts and screams contained in the vocals of "heavy music". The genre model that utilizes both lyrics and audio managed to combine both modalities and achieve superior results. For the emotional domain since the

lyrical model was not very accurate, the multimodal model closely follows the audio approach which is also evident during the explanation phase. The main hindrance to the models performing better appears to be the label ambiguity present in both tasks. As mentioned in this paper [21] "tagging a musical excerpt with an emotion label can be a vague and ambivalent exercise due to its subjective nature". Although genre tags might seem as more discrete than emotion labels, many songs incorporate elements from multiple genres and different experts might use different criteria for genre classification [60].

6.2 Future Work

The results also uncover several avenues for further investigation outlining potential directions for future research that can improve upon the foundational work presented here. To begin with, in order to address the label ambiguity impediment, future research could focus on building a dataset with genre annotations from experts. Although they could also have disagreements as to the genre of many instances, such a dataset would improve upon amateur labeled pieces of work. Also, since many songs might fall under multiple genre labels, for example being both "pop" and "electronic", a logical next step could be to implement the models of this study in a multi-label setting. Moreover, in our work we concatenate the embeddings outputed from each unimodal model and then use a classification head to make predictions. In order to improve performance future endeavours could study different ways to train the multimodal models which might include different combinations of the modalities (e.g. multiplying or adding the embeddings together). Finally, the models utilized are pretrained on data outside the music domain (e.g. Wikipedia corpora, or YouTube video spectrograms). A multilingual BERT like model pretrained on lyrics and perhaps poetry and a spectrogram transformer pretrained on pieces of music could yield interesting results.

Although the explanation methods and their global aggregates in this study gave adequate results, additional studies could improve upon them. Generating local explanations with LIME requires setting the number of samples variable. This variable controls the number of samples close to the original instance that are going to be created. As of now an exhaustive number of samples is prohibitive for large models with many input features. Subsequent studies could investigate what is sufficient number of samples, or specific ways to perturb the input so that the local explanations are accurate. Also, such studies could also include a statistical analysis to determine the appropriate number of instances to include in the global aggregates from the test set, so that an accurate representation of the model's behaviour is ensured. Furthermore explainability endeavours might find it fruitful to generate explanations based on lyric lines instead of individual words. Furthermore, exploring counterfactual explainability, that is altering certain features and observing how these changes impact the model's predictions, can offer a different perspective for emotion and genre recognition tasks. Finally, continuation of this work could include human evaluations of the explanations through user surveys providing insights into the perveived accuracy of the predictions from a listener's/reader's perspective.

Bibliography

[1] Agrawal, Y., Shanker, R. G. R., and Alluri, V. "Transformer-based approach towards music emotion recognition from lyrics". In: *CoRR* abs/2101.02051 (2021). arXiv: 2101.02051. URL:

[2] Ajayi, T. M. and Filani, I. "Pragmatic Function(s) of Pronouns in Nigerian Hip Hop Music". In: *UJAH: Unizik Journal of Arts and Humanities* (2016).

[3] Akalp, H. et al. "Language Representation Models for Music Genre Classification Using Lyrics". In: *2021 International Symposium on Electrical, Electronics and Information Engineering*. ISEEIE 2021. Seoul, Republic of Korea: Association for Computing Machinery, 2021, pp. 408–414. ISBN: 9781450389839. DOI: 10.1145/3459104.3459171. URL:

[4] Akman, A. and Schuller, B. W. "Audio Explainable Artificial Intelligence: A Review". In: *Intelligent Computing* 3 (2024), p. 0074. DOI: 10.34133/icomputing.0074. eprint: URL:

[5] Alexander, M. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press, 2010.

[6] Aljanaki, A., Yang, h., and Soleymani, M. "Developing a benchmark for emotional analysis of music". In: *PLOS ONE* 12 (Mar. 2017), e0173392. DOI: 10.1371/journal.pone.0173392.

[7] Baltrušaitis, T., Ahuja, C., and Morency, L.-P. "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019), pp. 423–443. DOI: 10.1109/TPAMI.2018.2798607.

[8] Benzi, K. et al. "FMA: A Dataset For Music Analysis". In: *CoRR* abs/1612.01840 (2016). arXiv: 1612.01840. URL:

[9] Bertin-Mahieux, T. et al. "The Million Song Dataset". In: *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*. 2011.

[10] Bogue, R. "Chapter 5 Violence in Three Shades of Metal: Death, Doom and Black". In: *Deleuze and Music*. Edinburgh: Edinburgh University Press, 2004, pp. 95–117. ISBN: 9781474465489. DOI: doi:10.1515/9781474465489-006. URL:

[11] Brown, T. B. et al. "Language Models are Few-Shot Learners". In: *CoRR* abs/2005.14165 (2020). arXiv: 2005.14165. URL:

[12] Çano, E. and Morisio, M. "MoodyLyrics: A Sentiment Annotated Lyrics Dataset". In: *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*. ISMSI '17. Hong Kong, Hong Kong: Association for Computing Machinery, 2017, pp. 118–124. ISBN: 9781450347983. DOI: 10.1145/3059336.3059340. URL:

[13] Chen, K. et al. "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection". In: *CoRR* abs/2202.00874 (2022). arXiv: 2202.00874. URL:

[14] Chowdhury, S. et al. "Towards Explainable Music Emotion Recognition: The Route via Mid-level Features". In: *CoRR* abs/1907.03572 (2019). arXiv: 1907.03572. URL:

[15] Delbouys, R. et al. "Music Mood Detection Based On Audio And Lyrics With Deep Neural Net". In: *CoRR* abs/1809.07276 (2018). arXiv: 1809.07276. URL:

[16] Dervakos, E., Kotsani, N., and Stamou, G. "Genre Recognition from Symbolic Music with CNNs: Performance and Explainability". In: *SN Computer Science* 4 (Dec. 2022). DOI: 10.1007/s42979-022-01490-6.

[17] Devlin, J. et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL:

[18] Dosovitskiy, A. et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *CoRR* abs/2010.11929 (2020). arXiv: 2010.11929. URL:

[19] Elizalde, B. et al. *CLAP: Learning Audio Concepts From Natural Language Supervision*. 2022. arXiv: 2206.04769.

[20] Filandrianos, G. et al. "Counterfactuals of Counterfactuals: a back-translation-inspired approach to analyse counterfactual editors". In: *arXiv preprint arXiv:2305.17055* (2023).

[21] Gómez-Cañón, J. et al. "Joyful for you and tender for us: the influence of individual characteristics and language on emotion labeling and classification". In: *Proceedings of the 21st International Society for Music Information Retrieval Conference*. Ed. by J. Cumming et al. ISMIR. Montréal, Canada: ISMIR, Oct. 2020, pp. 853–860.

[22] Gong, Y., Chung, Y., and Glass, J. R. "AST: Audio Spectrogram Transformer". In: *CoRR* abs/2104.01778 (2021). arXiv: 2104.01778. URL:

[23] Grekow, J. "Music emotion recognition using recurrent neural networks and pretrained models". In: *Journal of Intelligent Information Systems* 57.3 (2021), pp. 531–546. DOI: 10.1007/s10844-021-00658-5. URL:

[24] Han, D. et al. "A survey of music emotion recognition". In: *Frontiers of Computer Science* 16.6 (Jan. 22, 2022), p. 166335. DOI: 10.1007/s11704-021-0569-4. URL:

[25] Haunschmid, V., Manilow, E., and Widmer, G. "audioLIME: Listenable Explanations Using Source Separation". In: *CoRR* abs/2008.00582 (2020). arXiv: 2008.00582. URL:

[26] Hennequin, R. et al. "Spleeter: a fast and efficient music source separation tool with pre-trained models". In: *Journal of Open Source Software* 5.50 (2020). Deezer Research, p. 2154. DOI: 10.21105/joss.02154. URL:

[27] Hevner, K. "Experimental Studies of the Elements of Expression in Music". In: *The American Journal of Psychology* 48.2 (1936), pp. 246–268. ISSN: 00029556. URL: (visited on 07/03/2024).

[28] Huang, Q. et al. *MuLan: A Joint Embedding of Music Audio and Natural Language*. 2022. arXiv: 2208.12415 [eess.AS].

[29] International Society for Music Information Retrieval. *ISMIR*. Retrieved April 10, 2024. URL:

[30] Jin, D. et al. "Is BERT Really Robust? Natural Language Attack on Text Classification and Entailment". In: *CoRR* abs/1907.11932 (2019). arXiv: 1907.11932. URL:

[31] Krols, T., Nikolova, Y., and Oldenburg, N. *Multi-Modality in Music: Predicting Emotion in Music from High-Level Audio Features and Lyrics*. 2023. arXiv: 2302.13321 [cs.SD].

[32] Kwinten Crauwels. *MusicMap*. Accessed: 2024. 2023.

[33] Liartis, J. et al. "Semantic Queries Explaining Opaque Machine Learning Classifiers." In: *DAO-XAI*. 2021.

[34] Liartis, J. et al. "Searching for explanations of black-box classifiers in the space of semantic queries". In: *Semantic Web* Preprint (2023), pp. 1–42.

[35] Linden, I. van der, Haned, H., and Kanoulas, E. "Global Aggregations of Local Explanations for Black Box models". In: *CoRR* abs/1907.03039 (2019). arXiv: 1907.03039. URL:

[36] Lipton, Z. C. "The mythos of model interpretability (2016)". In: *arXiv preprint arXiv:1606.03490* (2016), pp. 14–18.

[37] Liu, X. et al. *CAT: Causal Audio Transformer for Audio Classification*. 2023. arXiv: 2303.07626 [cs.SD].

[38] Liu, Y. et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL:

[39] Lyberatos, V. et al. *Perceptual Musical Features for Interpretable Audio Tagging*. 2024. arXiv: 2312.11234 [cs.SD].

[40] Mastromichalakis, O. M., Liartis, J., and Stamou, G. "Beyond One-Size-Fits-All: Adapting Counterfactual Explanations to User Objectives". In: *arXiv preprint arXiv:2404.08721* (2024).

[41] Mastromichalakis, O. M. et al. "Rule-Based Explanations of Machine Learning Classifiers Using Knowledge Graphs". In: *Proceedings of the AAAI Symposium Series*. Vol. 3. 1. 2024, pp. 193–202.

[42] MENIS-MASTROMICHALAKIS, O. "Explainable Artificial Intelligence: An STS perspective". In: ().

[43] Menis-Mastromichalakis, O., Sofou, N., and Stamou, G. "Deep Ensemble Art Style Recognition". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9207645.

[44] Menis-Mastromichalakis, O. et al. "Semantic Prototypes: Enhancing Transparency Without Black Boxes". In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2024.

[45] Miller, T. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38.

[46] Neumayer, R. and Rauber, A. "Integration of Text and Audio Features for Genre Classification in Music Information Retrieval". In: *Advances in Information Retrieval*. Ed. by G. Amati, C. Carpineto, and G. Romano. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 724–727. ISBN: 978-3-540-71496-5.

[47] Panda, R. et al. "Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis". In: Oct. 2013.

[48] Pandeya, Y. R. et al. "Multi-modal, Multi-task and Multi-label for Music Genre Classification and Emotion Regression". In: *2021 International Conference on Information and Communication Technology Convergence (ICTC)*. 2021, pp. 1042–1045. DOI: 10.1109/ICTC52510.2021.9620826.

[49] Pyrovolakis, K., Tzouveli, P., and Stamou, G. "Multi-Modal Song Mood Detection with Deep Learning". In: *Sensors* 22.3 (2022). ISSN: 1424-8220. DOI: 10.3390/s22031065. URL:

[50] Raffel, C. et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *CoRR* abs/1910.10683 (2019). arXiv: 1910.10683. URL:

[51] Ribeiro, M. T., Singh, S., and Guestrin, C. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[52] Rodis, N. et al. *Multimodal Explainable Artificial Intelligence: A Comprehensive Review of Methodological Advances and Future Research Directions*. 2023. arXiv: 2306.05731 [cs.AI].

[53] Ross, A., Marasovic, A., and Peters, M. E. "Explaining NLP Models via Minimal Contrastive Editing (MiCE)". In: *CoRR* abs/2012.13985 (2020). arXiv: 2012.13985. URL:

[54] Santana, I. A. P. et al. "Music4All: A New Music Database and its Applications". In: *27th International Conference on Systems, Signals and Image Processing (IWSSIP 2020)*. Niterói, Brazil, 2020, pp. 1–6.

[55] Seo, Y.-S. and Huh, J.-H. "Automatic Emotion-Based Music Classification for Supporting Intelligent IoT Applications". In: *Electronics* 8 (Feb. 2019), p. 164. DOI: 10.3390/electronics8020164.

[56] seungheondoh. *audio-lyrics-emotion-recognition*. Accessed: 16 Oct 2023. 2019. URL:

[57] Stöter, F.-R. et al. "Open-Unmix - A Reference Implementation for Music Source Separation". In: *Journal of Open Source Software* (2019). DOI: 10.21105/joss.01667. URL:

[58] Strauss, H. et al. "The Emotion-to-Music Mapping Atlas (EMMA): A systematically organized online database of emotionally evocative music excerpts". In: *Behavior Research Methods* 56.4 (June 1, 2024), pp. 3560–3577. ISSN: 1554-3528. DOI: 10.3758/s13428-024-02336-0. URL:

[59] Sturm, B. L. "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use". In: *CoRR* abs/1306.1461 (2013). arXiv: 1306.1461. URL:

[60] van Venrooij, A. and Schmutz, V. "Categorical ambiguity in cultural fields: The effects of genre fuzziness in popular music". In: *Poetics* 66 (2018), pp. 1–18. ISSN: 0304-422X. DOI: https://doi.org/10.1016/j.poetic.2018.02.001. URL:

[61] Vaswani, A. et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL:

[62] Wickstrøm, K. K. et al. *RELAX: Representation Learning Explainability*. 2022. arXiv: 2112.10161 [stat.ML].

[63] Won, M., Chun, S., and Serra, X. "Toward Interpretable Music Tagging with Self-Attention". In: *CoRR* abs/1906.04972 (2019). arXiv: 1906.04972. URL:

[64] Wullenweber, A., Akman, A., and Schuller, B. W. "CoughLIME: Sonified Explanations for the Predictions of COVID-19 Cough Classifiers". In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2022, pp. 1342–1345. DOI: 10.1109/EMBC48229.2022.9871291.

[65] Zhang, Y. et al. *Interpreting Song Lyrics with an Audio-Informed Pre-trained Language Model*. 2022. arXiv: 2208.11671 [cs.SD].