ARISTOTLE UNIVERSITY OF THESSALONIKI
COMPUTER SCIENCE DEPARTMENT
DATA AND WEB SCIENCE MSc

Master Thesis

Emotion-Aware Music Recommendation Systems

Mitigating the Consequences of Emotional Data Sparsity

Papadopoulos Stefanos Iordanis
Supervisor : Athena Vakali

March 2021

Abstract

Emotion-Aware Music Recommendation Systems

Mitigating the Consequences of Emotional Data Sparsity
Papadopoulos Stefanos-Iordanis

Recommendation Systems (RS) for music have become indispensable for millions of regular music listeners and musicophiles alike, who rely on music streaming platforms such as Spotify, Apple Music and YouTube Music, for receiving personalised musical suggestions. In recent years, research on RS has been mainly focusing on 1) developing more advanced, efficient and accurate recommendation algorithms and on 2) more precisely profiling their users in order to better understand their musical preferences. The former involves the development of **hybrid filtering** models such as factorization machines and two-tower neural networks that are able to utilize collaborative, user-based and content-based information simultaneously. The objective of hybrid models is to improve upon predictive accuracy and overcome technical challenges such as the cold start of new items and data sparsity. On the other hand, **user profiling** techniques rely on user-side information such as demographic characteristics (age, gender) or contextual information (location, weather, temporal tendencies) in order to better understand user preferences, improve the situational personalisation of recommendations and by extension the overall user satisfaction. More recently, a sub-category of context-aware systems that has received notable attention in the research community is **Emotion-Aware Recommendation Systems** (EA-RS).

EA-RS rely on the assumption that music preferences are highly subjective and dependent on the current emotional state of the user. Thus, having access and training recommendation algorithms on **(user, item, emotions) tuples** may improve situational personalisation and user satisfaction. Current literature on EA-RS for music has corroborated this assumption by showing noteworthy improvements in terms of predictive accuracy and user satisfaction, while using emotion-aware collaborative filtering methods. Moreover, numerous recent studies have been experimenting with different types of emotional input sources, namely user reviews, social media posts, face emotion recognition and wearable devices. However, what is missing from the current research studies in EA-RS for music is the experimentation with State-of-the-Art (SOTA) recommendation algorithms such as hybrid filtering. Additionally, the issues of cold start,

popularity bias and data sparsity have not yet been addressed in the context of EA-RS. Thus, the objective of this research study is to address these research gaps.

More specifically, the present diploma thesis addresses the effects of **emotional data sparsity** and how it reinforces the **popularity bias** and **cold start** of new items. As previously noted, EA-RS are trained on (user, item, emotions) tuples which require the existence of an emotional input source in order to work, such as user reviews, face emotion recognition or wearable devices. However, in real-world applications users do not always write reviews for the items they interact with, nor is it realistic to expect them to constantly use face emotion recognition or wear their wearable devices. Users may frequently be disconnected from the emotional input source and this will result in **missing emotional values**. When missing values are aggregated, they will lead to the phenomenon of emotional data sparsity which is hypothesised to reinforce the issues of cold start and popularity bias, since EA-RS will not have enough information about how different users react emotionally to new and less popular items.

The proposed solution for the aforementioned challenges involves the combination of hybrid filtering models with cross-platform data fusion techniques. Firstly, hybrid filtering models utilizing both collaborative and content-based information have been shown to be better equipped at handling the issue of item cold start. Secondly, cross-platform data fusion comprises a specific type of item profiling - that fuses content-based information from multiple sources - can further contribute to the given objective. For the purposes of this study, two SOTA **hybrid filtering** models were selected, factorization machines (FMs) and two-tower neural networks (2TNN). Both models were trained and compared on how they perform in an emotion-aware recommendation task. A significant difference between the two selected models is their objective functions. FMs are trained with a learning-to-rank loss function (the weighted approximate-rank pairwise functions), while 2TNN utilized multi-task optimisation, trained simultaneously for minimizing retrieval and rating prediction errors. Despite their differences, both objective functions rely on **negative sampling**, a method which has been proposed in the context of mitigating the popularity bias.

Additionally, the present thesis suggest two novel proposals of cross platform data fusion, specialised on EA-RS for Music. Firstly, **multi-level profiling** involves the profiling of musical items on multiple levels, namely : ‘Track’ (lyrics, acoustic features etc), ‘Album’ (genre, release date etc) and ‘Artist’ (discography, personality etc) levels. It is hypothesised that profiling items on multiple levels will better express the multi-faceted relationships that listeners have with musical items. Secondly, **cross-platform audience reactions**, extends data fusion approaches from item to user-side profiling by

incorporating the emotional reactions of users from a different platform. It is hypothesised that having access to the emotional reactions of 'proxy' users in relation to new and less popular items will help in mitigating the issues of cold start and popularity bias.

In order to empirically examine the central hypothesis of this study, a custom dataset was collected fusing data from multiple sources (Album of the Year, Spotify, YouTube, Twitter and a custom search engine for song lyrics, which was then separated into two versions. In the first, referred to as **'Ideal dataset'**, there were no missing emotional values and thus the issue of emotional data sparsity was not present. In the second, called **'real-world dataset'**, a high level of emotional sparsity was present. In this way, the **causal effects**, that emotional data sparsity may have on different recommendation models, could be examined. Furthermore, all recommendation models were evaluated on multiple metrics for both **quantitative** (accuracy and root mean squared error) and **qualitative**, also known as 'beyond accuracy' metrics such as Novelty, Item Coverage and Personalisation. In this way, an in-depth comparison of the models was made possible and a more thorough examination of the potential usefulness of EA-RS for music could be assessed. Lastly, in order to establish the effect that each family of features had on the final outcome, all models were trained and evaluated in all possible feature combinations, including a 1) baseline relying only on collaborative information, 2) content-based features, 3) user emotions and 4) hybrid features, which combined collaborative, content-based information and users' emotional reactions.

The empirical results validated the central hypothesis of this thesis by showing that emotion-aware collaborative filtering (CF) models when applied in the 'real-world' dataset, where a high level of emotional data sparsity was present, they had a significant decrease in terms of novelty, item coverage and personalisation, indicating the reinforcement of the popularity bias. Furthermore, both CF and FM models were suffering from the cold start of new items and were unable to recommend new items. On the other hand, the proposed 2TNN, while not surpassing FM's remarkable predictive performance, it was able to maintain noteworthy scores of novelty, item coverage and personalisation even when a high level of emotional sparsity was present in the dataset. Additionally, 2TNN, utilizing hybrid features, was able to mitigate the cold start of new items. This study concludes by 1) proposing how both FMs and 2TNN models can be employed in production for different types of applications, 2) offering technical suggestions for further improving the qualitative performance of 2TNN and 3) discussing potential directions for future research works in the context of Emotion-Aware Recommendation Systems for Music.

Περίληψη

Συστήματα Συστάσεων Μουσικής με Συναισθηματική Επίγνωση
Αντιμετωπίζοντας τις συνέπειες της σποραδικότητας συναισθηματικών δεδομένων
Παπαδόπουλος Στέφανος-Ιορδάνης

Τα μουσικά συστήματα συστάσεων (music recommendation systems) , έχουν γίνει αναπόσπαστο τμήμα της καθημερινότητας ακροατών και μουσικόφιλων οι οποίοι βασίζονται σε πλατφόρμες μουσικής αναπαραγωγής κατ'απαίτηση, όπως το Spotify, το Apple Music ή το YouTube Music, για να λάβουν εξατομικευμένες μουσικές συστάσεις. Τα τελευταία χρόνια, η έρευνα γύρω από τα συστήματα συστάσεων έχει επικεντρωθεί στην 1) ανάπτυξη πιο εξελιγμένων, αποδοτικών και μεγαλύτερης ακρίβειας αλγορίθμων και στην 2) ορθότερη προφιλοποίηση των χρηστών (user profiling) με σκοπό την βαθύτερη κατανόηση των μουσικών τους προτιμήσεων και κατ' επέκταση την βελτίωση της εξατομίκευσης των συστάσεων. Το πρώτο περιλαμβάνει την ανάπτυξη μοντέλων βασισμένων στο **υβριδικό φιλτράρισμα** (hybrid filtering), όπως οι μηχανές παραγοντοποίησης (factorization machines) και τα νευρωνικά δίκτυα 'δύο πύργων' (two tower neural networks), τα οποία μπορούν να χρησιμοποιούν ταυτόχρονα συνεργατικές (collaborative) πληροφορίες, καθώς επίσης μετα-δεδομένα τόσο από την πλευρά των χρηστών όσο και των μουσικών αντικειμένων. Ο στόχος των υβριδικών μοντέλων είναι να βελτιώσουν την ακρίβεια των προβλέψεων, καθώς και να αντιμετωπίσουν τεχνικές προκλήσεις, όπως η 'ψυχρή εκκίνηση' (cold start) των νέων αντικειμένων και η σποραδικότητα δεδομένων (data sparsity). Από την άλλη πλευρά, οι τεχνικές **προφιλοποίησης χρηστών** χρησιμοποιούν πληροφορίες από την πλευρά των χρηστών, όπως δημογραφικά στοιχεία (ηλικία, φύλο) ή πληροφορίες σχετικά με τα συμφραζόμενα και το περιβάλλον τους (όπως τοποθεσία, καιρός, χρονικές τάσεις), προκειμένου να γίνουν καλύτερα κατανοητές οι προτιμήσεις τους, να βελτιωθεί η εξατομίκευση συστάσεων και κατ' επέκταση η συνολική ικανοποίηση τους. Τα τελευταία χρόνια, μια υποκατηγορία των τεχνικών προφιλοποίησης χρηστών - η οποία έχει λάβει ιδιαίτερη προσοχή από την ερευνητική κοινότητα της μηχανικής μάθησης - είναι τα **συστήματα συστάσεων με συναισθηματική επίγνωση** (Emotion-Aware Recommendation Systems) (EA-RS).

Τα EA-RS στηρίζονται στην υπόθεση ότι οι μουσικές προτιμήσεις είναι κατ' εξοχήν υποκειμενικές και ότι εξαρτώνται από την τρέχουσα συναισθηματική κατάσταση του χρήστη. Επομένως, έχοντας πρόσβαση και εκπαιδεύοντας αλγορίθμους συστάσεων με

πλειάδες της μορφής (**χρήστης, συναισθήματα, μουσικό αντικείμενο**), τα EA-RS μπορούν να βελτιώσουν περαιτέρω την κατά περίπτωση εξατομίκευση (situational personalisation) και άρα την ευρύτερη ικανοποίηση των χρηστών. Η τρέχουσα βιβλιογραφία για τα EA-RS μουσικής επιβεβαιώνει αυτήν την υπόθεση δείχνοντας αξιοσημείωτες βελτιώσεις όσον αφορά την προγνωστική ακρίβεια και την ικανοποίηση των χρηστών κατά τη χρήση μοντέλων συνεργατικού φιλτραρίσματος (collaborative filtering) (CF) με συναισθηματική επίγνωση. Επιπλέον, αρκετές πρόσφατες μελέτες έχουν πειραματιστεί με διαφορετικού τύπου 'συναισθηματικές πηγές', όπως 1) κριτικές χρηστών, 2) δημοσιεύσεις στα κοινωνικά δίκτυα, 3) 'έξυπνες' φορητές συσκευές και 4) την αναγνώριση συναισθημάτων από τις εκφράσεις του προσώπου μέσω τεχνικών μηχανικής όρασης. Ωστόσο, αυτό που απουσιάζει από τις τρέχουσες ερευνητικές μελέτες των EA-RS μουσικής είναι ο πειραματισμός με αλγόριθμους προτάσεων τελευταίας τεχνολογίας, όπως το υβριδικό φιλτράρισμα. Επιπρόσθετα, τα ζητήματα της 'ψυχρής εκκίνησης' (cold start) νέων αντικειμένων, η 'μεροληψία της δημοτικότητας' (popularity bias) και η σποραδικότητα δεδομένων (data sparsity) δεν έχουν ακόμη μελετηθεί και αντιμετωπιστεί στα πλαίσιο των EA-RS. Επομένως, ο στόχος της παρούσας ερευνητικής μελέτης είναι η κάλυψη των προαναφερθέντων ερευνητικών κενών.

Πιο συγκεκριμένα, η παρούσα διπλωματική εργασία μελετά το φαινόμενο και τις συνέπειες της **σποραδικότητας των συναισθηματικών δεδομένων** (emotional data sparsity) και το πώς αυτό ενισχύει περαιτέρω την 'ψυχρή εκκίνηση' των νέων αντικειμένων και την 'μεροληψία της δημοτικότητας'. Όπως προαναφέρθηκε, τα EA-RS εκπαιδεύονται σε πλειάδες της μορφής (χρήστης, συναισθήματα, μουσικό αντικείμενο) τα οποία προϋποθέτουν την ύπαρξη μιας συναισθηματικής πηγής για να λειτουργήσουν. Ωστόσο, σε πραγματικές εφαρμογές οι χρήστες δεν γράφουν πάντα κριτικές για τα αντικείμενα με τα οποία αλληλεπιδρούν, ούτε είναι ρεαλιστική η προσδοκία ότι θα χρησιμοποιούν συνεχώς αναγνώριση συναισθημάτων από τις εκφράσεις του προσώπου ή ότι πάντα θα φοράνε τις 'έξυπνες' φορητές τους συσκευές. Ενδέχεται, οι χρήστες συχνά να αποσυνδέονται από την 'συναισθηματική πηγή' και αυτό θα έχει ως αποτέλεσμα την δημιουργία **ελλιπών συναισθηματικών τιμών**. Μετέπειτα, η συσσώρευση αρκετών 'ελλιπών τιμών' θα οδηγήσουν στο φαινόμενο της σποραδικότητας των συναισθηματικών δεδομένων το οποίο με την σειρά του θα ενισχύει περαιτέρω τα ζητήματα της 'ψυχρής εκκίνησης' και της 'μεροληψίας της δημοτικότητας'. Αυτό οφείλεται στο ότι το EA-RS δεν θα έχει αρκετές πληροφορίες σχετικά με το πώς διαφορετικοί χρήστες αντιδρούν συναισθηματικά σε καινούργια ή λιγότερο δημοφιλή μουσικά αντικείμενα.

Η προτεινόμενη λύση της παρούσας εργασίας για την αντιμετώπιση των προαναφερθέντων προκλήσεων περιλαμβάνει το συνδυασμό μοντέλων υβριδικού φιλτραρίσματος με τεχνικές συγχώνευσης δεδομένων (data fusion). Αρχικά, τα μοντέλα υβριδικού φιλτραρίσματος που χρησιμοποιούν 'συνεργατικές' πληροφορίες ταυτόχρονα με μετα-δεδομένα από την πλευρά του χρήστη και των αντικειμένων' έχουν αποδειχθεί καλύτερα εξοπλισμένα στο χειρισμό της 'ψυχρής εκκίνησης'. Επιπλέον, για τον ίδιο σκοπό έχουν αποδειχθεί χρήσιμες ορισμένες τεχνικές συγχώνευσης δεδομένων - δηλαδή η συλλογή και συνδυασμός δεδομένων από πολλαπλές πηγές - ώστε να αναπαραστήσουν με μεγαλύτερη ακρίβεια τα χαρακτηριστικά των μουσικών αντικειμένων. Στα πλαίσια της παρούσας διπλωματικής μελέτης, επιλέχθηκαν δύο State-of-the-Art μοντέλα **υβριδικού φιλτραρίσματος**, τα Factorization Machines (FM) και Two-Tower Neural Networks (2TNN). Και τα δύο μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν για την απόδοσή τους στα πλαίσια πραγματοποίησης συστάσεων με συναισθηματική επίγνωση. Μια σημαντική διαφορά μεταξύ των δύο επιλεγμένων μοντέλων είναι οι αντικειμενικές τους συναρτήσεις (objective functions). Τα FM εκπαιδεύονται βάσει μιας συνάρτησης 'σταθμισμένου βαθμού ανά ζεύγη' (weighted approximate-rank pairwise) ενώ το 2TNN χρησιμοποιεί τη βελτιστοποίηση πολλαπλών εργασιών (multi-task optimisation), δηλαδή εκπαιδεύεται ταυτόχρονα για την ελαχιστοποίηση των σφαλμάτων ανάκτησης και την πρόβλεψη βαθμολογιών. Όμως, παρά τις διαφορές τους, και οι δύο αντικειμενικές συναρτήσεις κάνουν χρήση της **αρνητικής δειγματοληψίας** (negative sampling), μια μέθοδος η οποία έχει προταθεί στα πλαίσια του περιορισμού της 'μεροληψίας της δημοτικότητας'.

Επιπλέον, η παρούσα διατριβή προτείνει δύο καινοτόμες προτάσεις συγχώνευσης δεδομένων μεταξύ πλατφορμών που είναι εξειδικευμένες στα μουσικά EA-RS. Αρχικά, η **πολυεπίπεδη προφιλοποίηση** (multi-level profiling) προτείνει την δημιουργία προφίλ των μουσικών αντικειμένων σε πολλαπλά επίπεδα, συγκεκριμένα σε επίπεδα Κομματιού (στίχοι, μουσικά χαρακτηριστικά κ.λπ.), Δίσκου (είδος, ημερομηνία κυκλοφορίας κ.λπ.) και Καλλιτέχνη (δισκογραφία, προσωπικότητα κ.λπ.). Έχει θεωρηθεί ότι το προφίλ στοιχείων πολλαπλών επιπέδων θα εκφράσει καλύτερα την περίπλοκη και ποικιλόμορφη σχέση που έχουν οι ακροατές με την μουσική. Δεύτερον, η εισαγωγή **αντιδράσεων κοινού εξωτερικής πλατφόρμας** (cross-platform audience reactions), επεκτείνει τις προσεγγίσεις συγχώνευσης δεδομένων από δημιουργία προφίλ αντικειμένων σε δημιουργία προφίλ χρηστών, ενσωματώνοντας τις συναισθηματικές αντιδράσεις χρηστών 'μεσολάβησης' από μια διαφορετική πλατφόρμα. Η υπόθεση της συγκεκριμένης πρότασης είναι ότι η πρόσβαση στις συναισθηματικές αντιδράσεις ορισμένων χρηστών 'μεσολάβησης' - ειδικά σε σχέση με νέα και λιγότερο δημοφιλή αντικείμενα - θα βοηθήσει στον περιορισμό της 'ψυχρής εκκίνησης' και της 'μεροληψίας της δημοτικότητας'.

Προκειμένου να εξεταστεί εμπειρικά η κεντρική υπόθεση της παρούσας μελέτης, ήταν πρώτα απαραίτητη η συλλογή ενός νέου συνόλου δεδομένων ικανό να καλύψει τις απαιτήσεις του συγκεκριμένου θέματος. Το συλλεχθέν σύνολο δεδομένων συγκέντρωσε και συγχώνευσε πληροφορίες από πολλαπλές πηγές, συγκεκριμένα από τα Album of the Year, Spotify, Youtube, Twitter και μια προσαρμοσμένη μηχανή αναζήτησης για στίχους μουσικών κομματιών. Μετά την συλλογή του, το σύνολο δεδομένων χωρίστηκε σε δύο διαφορετικές εκδοχές. Η πρώτη εκδοχή, η οποία αναφέρεται ως '**εξιδανικευμένο**' σύνολο δεδομένων, αποτελείται μόνο από τις αλληλεπιδράσεις χρηστών που περιλαμβάνουν γραπτή κριτική από την οποία θα μπορούσε να υπολογιστούν οι συναισθηματικές τους αντιδράσεις. Επομένως, από το πρώτο σύνολο δεδομένων δεν απουσιάζουν συναισθηματικές τιμές και, ως εκ τούτου, δεν εμφανίζεται το ζήτημα της σποραδικότητας συναισθηματικών δεδομένων. Στο δεύτερο, το οποίο αποκαλείται σύνολο δεδομένων '**πραγματικού κόσμου**', περιλαμβάνονται όλες οι αλληλεπιδράσεις μεταξύ χρηστών και μουσικών αντικειμένων (κριτικές, βαθμολογίες, ακρόαση) και επομένως υπάρχουν πολλαπλές περιπτώσεις από τις οποίες απουσιάζουν οι συναισθηματικές αντιδράσεις των χρηστών και επομένως, υπάρχει υψηλό ποσοστό συναισθηματικής σποραδικότητας. Με αυτόν τον τρόπο, ήταν δυνατή η εξέταση των **αιτιατών συνεπειών** που μπορεί να έχει η σποραδικότητα συναισθηματικών δεδομένων σε διαφορετικά μοντέλα συστάσεων. Επιπρόσθετα, όλα τα επιλεγμένα μοντέλα συστάσεων αξιολογήθηκαν βάσει πολλαπλών μετρικών τόσο **ποσοτικής** φύσεως, όπως η ακρίβεια προβλέψεων και η ελαχιστοποίηση των λαθών, όσο και **ποιοτικής** φύσεως - γνωστές και ως 'πέραν της ακρίβειας' - όπως η εξατομίκευση (personalisation), η πρωτοτυπία (novelty) των συστάσεων και η κάλυψη στοιχείων (item coverage) από τον συνολικό κατάλογο. Με αυτόν τον τρόπο, έγινε εφικτή η εκτενής σύγκριση των μοντέλων και κατ' επέκταση η αξιολόγηση της πιθανής χρησιμότητας των EA-RS σε πραγματικές εφαρμογές αναπαραγωγής μουσικής. Τέλος, προκειμένου να αποδειχθεί η επίδραση που είχε κάθε οικογένεια χαρακτηριστικών στο τελικό αποτέλεσμα, όλα τα μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν σε όλους τους πιθανούς συνδυασμούς χαρακτηριστικών, συμπεριλαμβανομένου 1) μόνο τις συνεργατικές πληροφορίες, 2) τα μετα-δεδομένα των αντικειμένων, 3) τις συναισθηματικές αντιδράσεις των χρηστών και 4) τα υβριδικά χαρακτηριστικά, που συνδυάζουν όλα τα παραπάνω μαζί.

Τα εμπειρικά αποτελέσματα των διεξαχθέντων πειραμάτων, επιβεβαιώνουν την κεντρική υπόθεση της διπλωματικής. Όταν τα CF μοντέλα χρησιμοποιούν το 'σύνολο δεδομένων πραγματικού κόσμου', παρουσιάζουν μια σημαντικά μειωμένη απόδοση όσον αφορά την πρωτοτυπία, την εξατομίκευση και την κάλυψη των αντικειμένων. Αυτό το φαινόμενο υποδεικνύει την ενίσχυση της 'μεροληψίας της δημοτικότητας'. Επιπλέον, τόσο το CF όσο και το FM έπασχαν από το πρόβλημα της 'ψυχρής εκκίνησης' και αδυνατούσαν πλήρως να προτείνουν καινούργια μουσικά αντικείμενα για τα οποία το σύστημα δεν είχε ήδη γνωστές αλληλεπιδράσεις. Αντιθέτως, το προτεινόμενο μοντέλο 2TNN, ενώ δεν ήταν ικανό να πλησιάσει τα εντυπωσιακά ψηλά ποσοστά ακρίβειας του FM, κατάφερε να διατηρήσει πολύ υψηλά αποτελέσματα πρωτοτυπίας εξατομίκευσης και κάλυψης αντικειμένων, ακόμα και κατά την απουσία υψηλού ποσοστού συναισθηματικών τιμών. Η συγκεκριμένη συμπεριφορά δείχνει την ικανότητα του μοντέλου να μειώσει σημαντικά το πρόβλημα της 'μεροληψίας της δημοτικότητας'. Επιπρόσθετα, το 2TNN κάνοντας χρήση των υβριδικών χαρακτηριστικών ήταν ικανό να προτείνει και καινούργια αντικείμενα δείχνοντας την δυνατότητα του μοντέλου να μειώσει και το πρόβλημα της 'ψυχρής εκκίνησης'.

Τέλος, η παρούσα εργασία καταλήγει με 1) ορισμένα συμπεράσματα για την εν δυνάμει χρήση και των δύο υβριδικών μοντέλων, FM και 2TNN, για διαφορετικές εφαρμογές στην βιομηχανία της μουσικής, 2) προτάσεις για την περαιτέρω βελτίωση της απόδοσης του 2TNN μοντέλου και 3) ορισμένες συστάσεις για μελλοντικές ερευνητικές κατευθύνσεις, που θα μπορούσαν να πραγματοποιηθούν στα πλαίσια έρευνας και ανάπτυξης συστημάτων συστάσεων με συναισθηματική επίγνωση για την μουσική.

Acknowledgements

The procedure of conducting and completing the present diploma thesis was a fairly challenging and laborious objective. Beside my personal efforts and dedication, this study would have not come into fruition without the valuable support from several people to whom I would like to express my deepest gratitude.

First and foremost, i would like to thank my supervisor, Professor Athena Vakali, for offering her guidance, expertise and support throughout this process. Her feedback proven to be vital for directing this study to its fullest potential.
Moreover, I would also like to thank Ph.D. candidate Dimitra Karanatsiou, for helping me stay on course and for offering her time and domain knowledge in psychometric profiling whenever was needed.

I would also like to express my sincerest appreciation to all the people that have been by my side during this time. My partner, Roxanne, who supported and encouraged me during this taxing period.
My parents for unequivocally supporting me all these years, both morally and materially. Thank you for instilling in me a sense of curiosity and wonder from an early age, as well as an appreciation of both arts and science.
Last but not least, i would like to thank my friends and fellow musicians from 'ORIA', for being so understanding even when i was skipping rehearsals and writing sessions during the past few months.

Contents

Chapter 1

Introduction

In this Chapter are discussed the motivation, general outline and contributions of this research study. Firstly, Section 1.1 briefly explores the necessity, evolution and core challenges in the development of Recommendation Systems and introduces the newly birthed variant of Emotion-Aware Recommendation Systems (EA-RS). Thereafter, in Section 1.2 the identified research gaps are addressed and the proposed solutions are presented. The former is concerned with the long term consequences that missing emotional values have on EA-RS, by furthering the Popularity Bias and Cold Start of new items while the latter involves the development of Hybrid emotion-aware Algorithms that utilize both Collaborative Information and Cross-Platform Data Fusion techniques. Afterwards, in Section 1.3, the general contributions of this research work are discussed while Section 1.4 concludes the Introductory Chapter by outlying the structure of this study.

1.1 Background

Recommendation Systems (RS) have been one of the most successful and wide spread applications emerging from the fields of Data Mining and Machine Learning. RS are responsible for filtering and selecting the most relevant items personalised to each individual user's interests, needs and desires. RS have practically become omnipresent and indispensable since contemporary individuals, knowingly or unknowingly, have numerous daily encounters with and have come to rely on them for myriad tasks. The content presented in our Social Media feeds (Facebook), the ads and products in e-Commerce websites (Amazon), the 'similar' articles in online news sites (CNN), movies and TV series suggested in on-demand streaming platforms (Netflix), available job positions (LinkedIn), vacation rentals (Airbnb) and matching dates (Tinder), among numerous others, are the result of recommendation algorithms.

Historically, people have relied on friends, family, colleagues, trusted sources and reviewers to discover products, books, movies, music, job offerings, vacation rentals etc that might be of personal interest, and most still do. When it came to the domain of music, the audience would mostly rely on their local album store, music magazines, radio stations and online blogs for receiving relevant recommendations. Although all these sources are still available, the advent of information technology, the internet, social media and on-demand streaming services - and by extension the immense magnitudes of produced information - have necessitated the development of advanced filtering algorithms, able to select - among thousands or millions of items - the most appropriate and relevant items for each user. Otherwise, navigating through these amounts of data without any filtering mechanism would be practically impossible, labour intensive and extremely time-consuming leading to **Information Overload** and **Choice Paralysis** [1]. In this context, specific to the domain of music, multiple on-demand platforms have been developed, including Spotify, Pandora, Deezer, Youtube Music and Apple Music competing for the retention of user attention and engagement of millions of music lovers.

Each industry may have domain-specific requirements but the logic and architecture behind their recommendation algorithms can be very similar and are generally classified into three broad categories. Collaborative, Content-Based and Hybrid Filtering. **Collaborative Filtering** (CF) models are based on the assumption that users with similar past preferences will also prefer similar items in the future. One important advantage of CF models is their architectural simplicity due to solely relying on past User - Item interactions - either explicit ratings or implicit information - without the need for further user-side or item-side meta data. Nonetheless, they frequently face difficulty recommending items to new users or finding appropriate users for new items, known as the Cold Start problem. Additionally, they tend to disproportionately recommend already popular items and by extension neglecting less popular ones, leading to the problem known as Popularity Bias [2]. On the other hand, **Content-Based Filtering** (CBF) rely on extensive item meta-data in order to identify items with similar characteristics with items that a user has shown interest in the past. CBF models do not suffer from cold start and popularity bias but have a tendency for overspecialisation and a lack in novelty and diversity [3]. Finally, in order to overcome the challenges that CF and CBF models tend to face, **Hybrid Filtering** models are being used that combine different approaches, for example aspects of both CF and CBF, leading to increased performance but usually with the trade-off of increased computational complexity [4].

In recent years, RS have been continuously advancing and becoming increasingly more complicated due to the growing magnitudes of produced information and generated content, the expanding demands of users and the ever increasing competition among different platforms to attract new users and keep them engaged. Researchers in the field of RS, apart from experimenting with models of increased architectural complexity, such as hybrid models and neural networks, are also utilizing advanced data integration and feature engineering techniques in order to increase the performance of their systems and better understand the interests and desires of their users [4]. Established techniques include **Demographic Filtering, Context-Aware, Knowledge-based** and **Cross-Domain** systems [5]. The first two techniques utilize user-based information, such as their demographics (age, gender, location etc) or the contextual environment (e.g time of day, weather etc) while the last two, employ knowledge and rules specific to the target domain or importing them from a separate but relevant domain (e.g transferring knowledge from book-related into movie-related recommenders). All of the above methods, require the collection of additional information that may be difficult to obtain but have proven worth-while endeavors in certain contexts. Indicatively, YouTube will suggest the most popular and trending items based on a new user's location and specific contextual information and demographics - mitigating the problem of user cold start (Demographic Filtering) [6], while Spotify will recommend different playlists for different time windows of the day, depending on each user's listening habits (Context-Aware) [7].

Along the same lines, more recently, with the advancement of psychometric tools, **Personality Detection** and **Emotion Analysis** techniques, researchers have been experimenting with extracting the personality traits or the emotional states of users and integrating them in the pipelines of recommendations algorithms. The central idea is that different personality traits may be correlated with different item characteristics (e.g acoustic features or genres in music) and that users under different emotional states and moods will prefer different types of items. Consequently, having access to users' psychological background could lead to increased predictive accuracy and personalisation. In this vein, a recent survey on music recommendation systems (MRS) [5] concluded that three of the most promising visions for further improving MRS are:

1. Psychologically inspired music recommendation (Personality and Emotions)
2. Situation-aware music recommendation
3. Culture-aware music recommendation

So far, in the domain of music, **Personality-Based** recommendation systems have shown limited performance and mixed results, probably due to small participant sizes and difficulty in accurately assessing personality traits [8]. On the other hand, **Emotion-Aware**

Recommendation Systems (EA-RS) have had very promising results, showing that recommendation algorithms utilizing the emotional states and responses of users can enhance their predictive accuracy and refine personalisation [9] [10] [11]. In this direction, numerous research works have been testing different possible sources of extracting user emotions, including written comments or reviews [12], social media posts [9], face emotion recognition [13] and wearable devices [14].

The current state of research on EA-RS focuses on:

1. Validating the **usefulness** of EA-RS meaning what they have to offer in terms of improved Accuracy, Personalisation [9] [11] and User Satisfaction[10] when compared to 'traditional' approaches.
2. Testing different types of '**emotion input sources**' for extracting users' emotions such as Social Media [9], Wearable Devices [14] and Face Emotion Recognition [15].

These studies are very important for advancing the research of EA-RS. However, what is lacking in the recent literature, is the examination of potential challenges specific to EA-RS, especially when implemented in a real-world situation, and subsequently provide preemptive solutions. Investigating technical issues such as Cold Start, Data Sparsity, Scalability or issues of Fairness such as Popularity Bias and Gray Sheep users are missing from current research works on EA-RS. Thus, the present diploma thesis is attempting to start filling these research gaps by examining the issue of **Data Sparsity** of emotional information and its effects in propagating the **Item Cold Start** and **Popularity Bias** specifically in the context of Emotion-Aware Recommendation Systems.

1.2 Addressed Problem

EA-RS systems rely on having an input source for acquiring the current emotional state of its users or the emotional responses that certain items elicit on them. As was previously mentioned, the **emotion input source** may be user comments or reviews, social media posts - if users choose to connect their personal social media pages with the music platform - Face Emotion Recognition (FER) or wearable devices such as smart watches. However, in a real-world implementation, users may frequently be disconnected from the platform's emotion source. They may be not wearing their wearable device, or choose not to use FER, or have not updated in their social media for a while. In these cases, the system will not have access to the emotional states and reactions of its users and thus can not perform emotionally-aware recommendations. This is due to the fact that EA-RS are trained on (user, item, emotions) tuples and the emotion information will be missing in those situations, as can be seen in figure 3.1. This issue is often raised in the literature from the user's point of view meaning that if an EA-RS platform does not have access to a user's recent emotional state it can not perform emotion-appropriate suggestions. It is often acknowledged that in these cases an EA-RS can not function properly and thus a traditional recommendation approach will be utilized instead, relying on user's past interactions, preferences etc. Nonetheless, the same phenomena is not discussed or studied from the item's point of view nor what the potential long-term consequences that **Missing Emotional Values** (MEV) and **Emotional Data Sparsity** (EDS) may have on EA-RS.

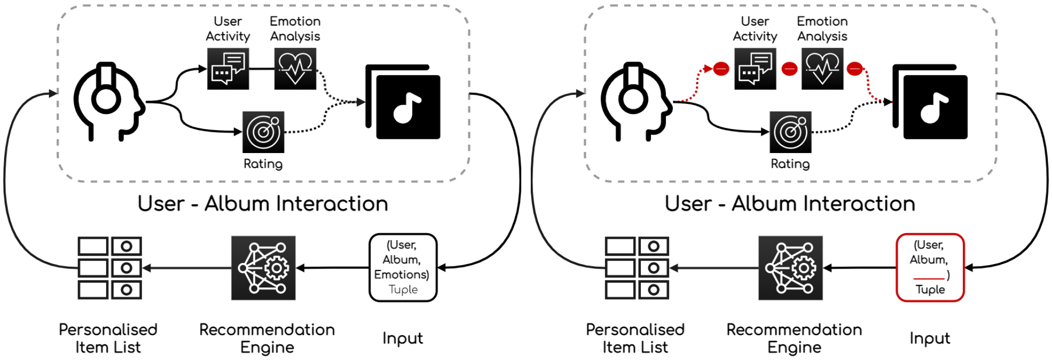


FIGURE 1.1: Indicative workflow of an Emotion-Aware System (left) and a case of a missing emotional input (right) that leads into a Missing Emotional Value.

The most central premise of this diploma is that multiple cases of MEV and by extension the resulting Sparsity of Emotional Interactions will negatively contribute to the phenomena of **Item Cold Start** and **Popularity Bias** since an EA-RS will not have enough information about what new and less popular items elicit to its users and will not be able to recommend them. Thus, the main objective is to develop an EA-RS based on real-world User-Item interactions and propose and evaluate solutions for the issues resulting from MEV, problems very similar to 'Item Cold Start' and 'Popularity Bias' of traditional collaborative filtering approaches. The proposed solution requires a hybrid algorithm that can utilize both collaborative information, content meta-data and dynamically changing user information, namely emotional responses. **Hybrid Filtering models** have been shown to be better equipped at mitigating the challenges of Popularity Bias and Item Cold Start [4] since they can identify patterns and significant relationships between users and item features instead of solely relying on user-item interactions. This enables the recommendation of new and less popular items based on their meta-data and characteristics.

Furthermore, the proposed solution relies on two novel proposals, Multi-Level Profiling and Cross Platform Audience Reaction. Fundamentally, both can be considered as extensions of Cross-Platform **Data Fusion** techniques for mitigating the issue of Item

Cold Start [16] but are further expanded and specialised on EA-RS for Music. Firstly, **Multi-Level Profiling** is the idea that music audiences have a multi-faceted relationship with musical items, at the level of each individual track, its acoustic characteristics, musical mood, lyrical emotions, that exist with other musical tracks forming the totality of a musical album as well as a relationship with the artist's biography and projected persona. It is hypothesised that having information representing to these multi-layer relationships can help a hybrid MRS to discover significant latent patterns between users and musical items and by extension improve its performance including difficult cases of new and less popular items. Secondly, **Cross Platform Audience Reaction** (CPAR) involves the collection of an average emotional reaction for a given musical item from a different music-related platform that can later be used as a proxy for the target platform. In theory, having access to the CPAR of a new or unpopular item may be adequate to inform the RS to which users it should recommend said item even if no previous in-platform interactions were present. Hence, the novelty of CPAR lies in extending cross-platform data fusion from a Item Profiling technique to a User Profiling one by using users' emotional reactions to specific items as proxy for our target platform.

Since these is no available existing dataset meeting all the requirements of this study, a data fusion pipeline was developed web scrapping the review aggregator website Album of the Year collecting User-Album interactions and users' reviews, as well as being connected to Twitter, YouTube and Spotify APIs and a custom Google search engine for collecting lyrics. All in all, the proposed solution can be thought of as a combination of Hybrid Filtering recommendation algorithms with Data Fusion techniques which both have been shown able to mitigate the problems of Cold Start and Popularity Bias but here are specialised for the application on EA-RS.

1.3 Novelty and Contribution

After discussing the theoretical context of this study, the identified research gaps in EA-RS, the addressed challenges as well as the proposed solutions, it can be extrapolated that the central objectives and contributions of this study are the following:

1. Exploration of Emotional Data Sparsity in EA-RS.

It is hypothesised that when missing emotional values are aggregated and the sparsity of emotional data is increased, it will negatively affect the performance of Emotion Aware systems. More specifically, it is hypothesised that high levels of emotional sparsity will contribute to the problem known as **Popularity Bias** and further affect the **Cold Start** of new items. Both hypothesis are examined empirically in relation to multiple models.

While both Cold Start and Popularity Bias are very important challenges and concern numerous research efforts [3] they have been neglected in the context of Emotion-Aware systems.

2. Mitigation of the Popularity Bias and Item Cold Start in EA-RS.

The proposed solution for overcoming the Popularity Bias and Item Cold Start of EA-RS consists of three interconnected parts.

- Hybrid Filtering algorithms able to utilize Collaborative, User-Based and Content-Based information simultaneously,
- Specialised Objective Functions relying on Negative Sampling (Learning-to-Rank and Multi-Task learning),
- Cross-Platform Data Fusion for both User and Item profiling.

All three solutions have been previously considered for mitigating the challenges of Popularity Bias and Item Cold Start [17] [16] [18]. They have not, however, been applied in the context of Emotion-Aware systems nor have been applied cooperatively, meaning for all three to work in tangent. Furthermore, the proposed profiling involves two novel methods, **Multi-Level profiling** and **Cross Platform Audience Reactions** specialised for the domain of music and for EA-RS that have not be previously examined and thus can be considered the most Novel of the proposed three-part solution.

3. Re-Examining the usefulness of Emotion-Aware systems.

This is achieved by comparing all emotion-aware algorithms with fine-tuned baselines that do not utilize users' emotional reactions and moreover, evaluate these models not only be in terms of their quantitative **Predictive Accuracy** but also on a variety of qualitative **'Beyond Accuracy'** metrics such as Personalisation, Novelty, Item Coverage and Scalability, in order to more thoroughly access their potential advantages and disadvantages. This is in contrast to the majority of the current literature that solely relies on accuracy and error minimization metrics [11] [13] [19] for validating the usefulness of EA-RS.

4. Comparative Analysis between Hybrid Emotion-Aware Algorithms.

In the current literature, most studies in EA-RS for music rely on adjusting traditional collaborative filtering models [9] [10] which are not in accordance with the State-of-the-Art [4]. In this study more advanced and complicated recommendation models, namely

Factorization Machines and **Two-Tower Neural Networks**, are also trained and evaluated in the context of EA-RS.

The first issue, the effects that emotional sparsity has on real-world EA-RS, can be considered as the central question of this study while the second contribution can be considered as the 'answer' and solution to 1). Subsequently, Contributions 3) and 4) can be considered as positive research corollaries deriving from the examination of 1) and 2) that while not being the focal point of this study, can still offer valuable insights in relation to the current research of EA-RS.

1.4 Research Structure

Section 1 concludes the general introduction of the main research objectives and contributions of this study. This work proceeds with an orderly structure of four more sections. In **Section 2** are discussed the fundamental theoretical underpinnings of recommendation systems and the most current literature in relation to context-aware and psychometric-based recommendation systems for music. This is done in order to identify potential gaps in the related research and contextualise this study. In **Section 3** is described the methodology for developing two hybrid emotion-aware recommendation systems for music, including the data collection process and the development of emotion and mood detection models for user reviews and track lyrics respectively. Subsequently, in **Section 4** are presented the results and the experimental design, for validating the central hypothesis of this study and evaluating the recommendation models. Finally, **Section 5**, concludes this diploma thesis by offering some final remarks and directions for future studies in relation to emotion-aware recommendations for music.

Chapter 2

Literature Review

This chapter provides an overview of the theoretical background and the most current research relevant to this diploma thesis. Initially, Section 2.1, offers a general overview of the most fundamental methodologies, evaluation strategies and challenges present in the field of Recommendation Systems. Thereafter, Section 2.2, describes the procedure for the conducted systematic review that facilitated the selection of the present study's central subject. The collected papers from the aforementioned procedure, related with Contextual, Emotion-Aware, Personality- and Data Fusion-based Recommendation systems, are presented and analysed in Section 2.3. Finally, Section 2.4 contextualises the current literature and discusses the identified research gaps in the sub-field of Emotion-Aware Recommendation Systems that motivated the present diploma thesis.

2.1 Theoretical Background - Fundamentals

At their core, Recommendation Systems are information filtering systems that try to suggest the most relevant items for each user based on their interests, needs and desires in a given domain. Such a system attempts to understand the complicated and multifaceted relationships between multiple users and items by analysing their past interactions in order to be able to accurately predict future interactions and perform appropriate and relevant suggestions. RS are one of the most successful and prevalent developments to arise from the fields of machine learning and data mining. This phenomenon is due to the, by now, indispensable role they play in various industries (e-commerce, social networks, streaming platforms and digital media, among many other) in combination to the ever-expanding generation of new data and the need for filtering and serving those data to the most interested users. Consequently, there exists a continuous stream of new research works on RS in order to improve upon existing algorithms or to solve specific, and ever more complicated, challenges. Among many, ACM's RecSys is arguably the most important and influential conferences, wholly dedicated on Recommendation Systems, advancing innovative research works on RS since 2007. But before presenting and analysing the most recent literature on Context-Aware, Psychometric-Based and Data Fusion-based Recommendation Systems, it is deemed important to concisely discuss the most fundamental concepts, methods, main challenges and evaluation methods in the research field of RS.

2.1.1 Model Families in Recommendation Systems

Since their inception, there has been an extensive amount of proposed models, algorithms and approaches that can be broadly categorised into three, Collaborative Filtering, Content-Based Filtering and Hybrid Filtering. In this section each family of methods is discussed as well as some of their advantages and disadvantages.

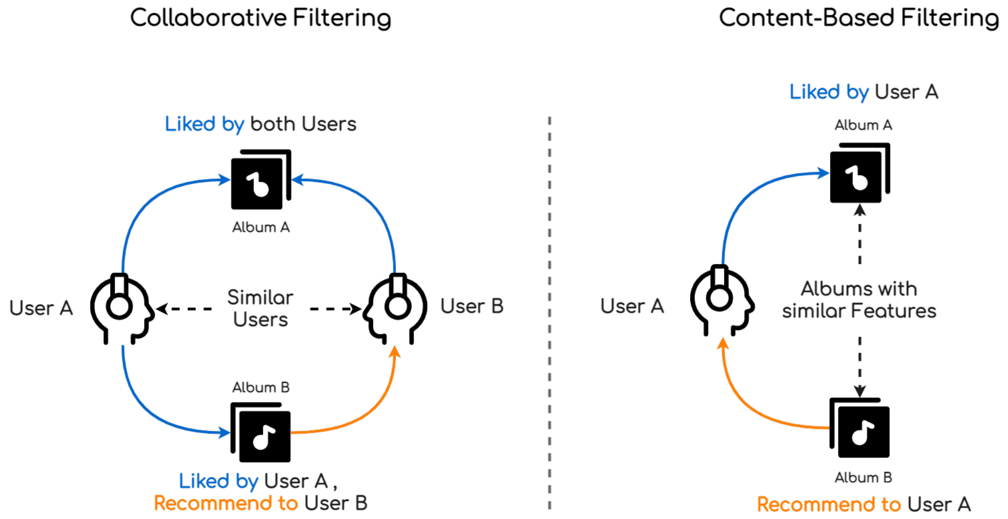


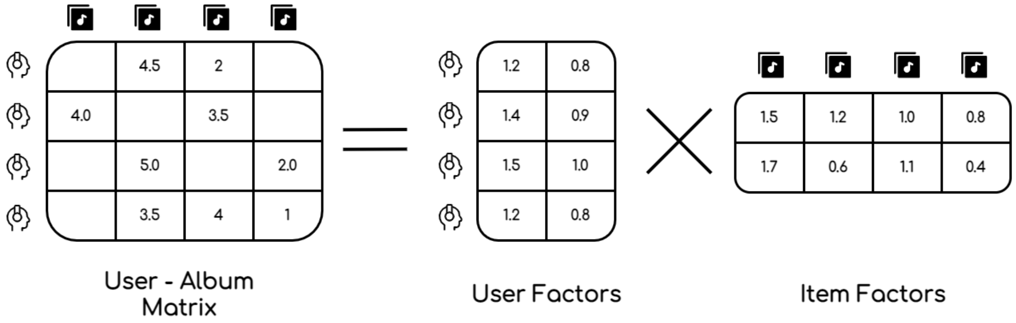
Figure 2.1: An example of Memory-Based Collaborative Filtering (left) and Content-Based Filtering (right)

Firstly, **Collaborative Filtering** is probably the most widely known and used family of techniques for recommendation systems. Their fundamental assumption is that users who have exhibited similar liking patterns and behaviors in the past will also have a tendency to like similar items in the future. CF systems require a significant volume of User (U) - Item(I) interactions - which can be either explicit (scaled ratings) or implicit (clicks, listening time, purchase) - and a measure of similarity in order to calculate 'neighborhoods' of similar Users that have liked the same Items and perform recommendations between them. Consequently, if Users U1 and U2 are considered similar users and U1 highly rated Item I1 that B has not interacted with, then I1 can be recommended to

U2. One significant advantage of CF systems is their simplicity of Implementation since they rely on a User * Item matrix and do not require additional Features like Meta-Data or User's demographic information to function properly. Additionally their suggestions are generally more Novel and Diverse when compared with Content-based Filtering. An important distinction must be made between **Memory-based CF** and **Model-based CF** approaches.

Memory-based CF more closely matches the previous description since it relies on a similarity measure (Cosine Similarity or Pearson Correlation Coefficient) and the calculation between All-to-All user ratings. This approach can be further divided into User-based (UBCF) and Item-based (IBCF) models, the first suggesting Items liked by similar users while the second suggests Items that are similarly rated. However, a common disadvantage of Memory-Based CF is that their Predictive Accuracy tends to decrease with the degree of Data Sparsity while the increasing scale of vectors decrease their Scalability [3].

Secondly, **Model-Based CF** approaches employ data mining and machine learning algorithms in order to predict missing values of unrated items. This category includes Cluster Analysis, Association Rules, Link Analysis but **Matrix Factorization** (MF) is the most commonly used family of models. Dimensionality reduction algorithms, like SVD or PCA, are utilized in order to discover significant latent factors and the initial User-Item matrix to be compressed into a low-dimensional representation without missing values. This approach is better suited to handle Data Sparsity and is significantly more Scalable since it does not require the whole High-Dimensional dataset to perform rating predictions [20].



	4.5	2	
4.0		3.5	
	5.0		2.0
	3.5	4	1

1.2	0.8
1.4	0.9
1.5	1.0
1.2	0.8

1.5	1.2	1.0	0.8
1.7	0.6	1.1	0.4

Figure 2.2: Matrix Factorization : low dimensional latent factors extracted from the sparse User-Item interaction matrix

Furthermore, in recent years, many researchers have been experimenting with Neural Network based CF (NN-CF), claiming them to be the new State of the Art [21]. However, a recent 'Performance Comparison Study' showed that many NN-CF research works were not replicable and those that were, when fairly compared with simpler 'traditional' CF approaches, their performance was found limited [22]. The researchers stressed the importance of reproducibility and the importance of correctly-tuned baseline models especially when proposing novel models with complicated architectures, such as **Neural Collaborative Filtering** [23].

Content-based Filtering (CBF), is a user-specific approach recommending new items with a high degree of similarity based on Item Features (content description or meta-data) with other items or groups of items the user likes. If for example, a user highly rated Song S1 tagged as T1 type of music then if S2 is also tagged as T1, A2 will be recommended to the user. This approach requires a method of Feature Representation (like tf-idf) for each Item, an individual profile for each user in relation to the Item Features and a filtering stage, selecting the most relevant unseen items for each user, usually some type of machine learning algorithm [5]. CBF is differentiated with IBCF is that the later solely relies on ratings while the former requires additional Item-related information like the Artist's name, Music Genre, Tags, Mood or others. This approach offers user independence and - unlike CF - does not suffer from the Cold-Start problem. Nevertheless, CBF has a tendency to overspecialize, suffering from a lack of Novelty and Diversity [3].

Hybrid Filtering (HF) is the approach of combining the functionality of two or more recommendation methods. R. Burke has offered a helpful framework for classifying Hybrid Filtering techniques into seven different categories [24]. In Mixed Hybrid, the results of two different models are calculated separately and then combined in a single list while in Weighted Hybrid the combination is based on weighted linear function based on different types of users. In Switching Hybrid different models are selected depending on specific criteria. For example, a Emotion-Aware RS relying on Social Media posts to extract the user's emotional state will not be used if the user has not made any recent posts and instead a different RS will be selected. In Feature Combination Hybrid and Meta-level Hybrid the output of the first model is being used as input in a second model with the only difference being in the first it is used as an additional feature among others while in the second it is the only feature. Feature-augmenting methods use a Data Mining model that its output is fed into a RS. For example, a Classifier may calculate 'Similar Artists' or 'Songs with similar mood' and these results will be used as input for a RS. Finally, according to Burke's taxonomy, Cascade Hybrid a RS is given a hierarchy of priorities and judges the results of the second RS. HF approaches are used in order to improve the overall accuracy or mitigate some specific problem of CF or CBF when used individually like Cold-start or Overspecialization respectively. However, in HF systems, there usually exists a tradeoff between increased performance and increased computational complexity [4].

More recently, there has also been an increased interest in algorithms with **Hybrid**

Architecture, such as Factorization Machines (FM) [25] and Two Tower Neural Networks (2T-NN) [26] that can simultaneously utilize collaborative information, item-side and user-side meta-data. Thus, maintaining the advantages of Hybrid Filtering without the need to train separate models. Variations of both FMs and 2T-NN model families have been used extensively in real-world applications such as Lyst's fashion e-commerce recommendation [17] and AirBnB's recent use of 2T-NNs [6] [27]. These models will be discussed in more detail in chapter 3.

2.1.2 Main Challenges in Recommendation Systems

As is was passingly mentioned in the previous section, each family of models has certain advantages and limitations. In this section, these limitations are put into context and the most fundamental challenges facing Recommendation Systems are discussed as well as presenting some recent attempts to mitigate these issues.

First of all, improving the **Accuracy** of recommendation systems is a fundamental issue, intensively preoccupying researchers in the field, since it translates into more precise and reliable recommendations and more broadly contributes to increased user satisfaction. The measures of Accuracy will be dependent on the type of each given Recommendation Task and generally categorised into Rating, Retrieval and Ranking prediction. Each approach is used for different purposes and are followed by different evaluation metrics. The issue of evaluating the accuracy of recommendation systems is will be discussed in more detail in the next section.

Secondly, since recommendation systems are intended to be used in a wide variety of real-life application of various domains (Movies and Music streaming, e-Commerce etc) and these domains tend to have enormous amounts of users, items and their interactions, renders the issue of **Scalability** to be of extreme importance. Furthermore, another issue related to the volume of data is that of **Data Sparsity**. Provided that most users have not interacted with most items the U - I matrix contains numerous missing values and is rather sparse. This phenomenon affects the effectiveness of recommendation systems, especially CF approaches. Common solutions include Dimensionality Reduction and Hybrid Filtering. One recently proposed solution, instead of only relying on User-Item interaction, incorporated Tag-User interactions with a low-rank matrix factorization model in order to extract latent factors related to Semantic Information about that Music Tracks that help mitigate the issue of Sparse Data [28].

More recently, researchers have stressed the importance of 'going beyond accuracy', showing an increased awareness towards the issue of lacking **Diversity** regarding attempts to increase the range and variety of recommended items each user receives so as to avoid overspecialization and keep them interested and engaged in the long run. Diverse recommendations need to balance between similarity and novelty. A recommendation should be similar enough to the user's existing preferences but dissimilar enough as to be novel. A recently proposed solution was based on a Graph-based approach that calculates user-user similarity based on degrees of Diversity or Similarity and creates a 'Virtual-Friend' profile based on the PageRank algorithm for improved personalized recommendations [28].

The **Gray Sheep User** problem involves users with idiosyncratic preferences that tend to favor lesser popular and 'underground' items and thus - especially - CF methods find it difficult to classify them into appropriate neighborhoods and suggest relevant items. Hybrid methods are usually selected to overcome this problem. A more recent attempt computed user-specific coefficients in relation to the Popularity of their preferred Artists, without the need for hybrid filtering or additional features (demographic, meta-data) and was able to improve the accuracy of both rating prediction and item recommendation when compared to traditional CF [29]. Finally, a related challenge is known as the **Popularity Bias** and has been observed in many recommendation systems, especially collaborative filtering methods, who show a tendency (bias) towards disproportionately recommending popular items - by definition items that have already received numerous ratings - and contrariwise neglect unpopular items [30] [31].

2.1.3 Evaluation of Recommendation Systems

Prior to deploying a Recommendation System into production it is very important to thoroughly evaluate its performance and effectiveness on various aspects. An offline evaluation can accessed either with a **user study** or with **task-specific metrics**. User studies require the participation of numerous users, who are presented with lists of recommended items, created by different methods and the users will explicitly express their personal preferences and rate each list. The participant's feedback best approximates the reactions of users in online evaluations, meaning when the model is deployed. However, user studies are costly and time consuming and require a significant scale of participants as well as very careful extensive experimental design. Conversely, the second offline evaluation approach involves splitting the dataset into training and test sets, training the recommendation algorithm on the test set, performing predictions on the previously unseen test set and thereafter calculating various task-specific metrics in comparison to the ground truth.

First and foremost is the evaluation of a model's predictive accuracy and how consistently it can suggest items that the users will find relevant. To this end, various ways have been proposed and utilized for accessing a model's predictive performance. Traditionally, one commonly used approach was **Rating Prediction**, the process of predicting the known explicit ratings that each user gives to different items with the minimum possible error. For that purpose, the most commonly used metrics are error-based metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) that calculate the difference between predicted ratings and user's known explicit ratings. The latter metric was used for the famous Netflix Prize competition [32]. But since then, it has been shown that solely optimising for minimizing explicit rating errors on already seen items, does not necessarily lead to useful recommendations [33].

H. Steck (2010) empirically showed that models optimized for explicit rating prediction, might be performing excellent in terms of RMSE but they would perform worse in terms of their Ranking ability when compared with models trained on implicit information. In the same work, that has since been reproduced repeatedly, models utilizing both explicit ratings and implicit interactions had shown better performances. The main reason is that ratings are not missing at random since the enjoyment of an item influences the likelihood that the user will rate said item. Additionally, the likelihood of selecting an item by a user is correlated with how favorably she will rate the item or what the expected enjoyment will be. This means that in most cases users do not select items at random and they base their selection on some factor related to the item and that they expect to enjoy. Furthermore, solely relying on explicit ratings and favouring positively predicted ratings can miss the preferences of more idiosyncratic users for example, users frequently select 'guilty pleasures'. In summation, analyzing implicit interactions can reveal important hidden patterns.

Consequently, most current research work, even when training on explicit ratings, tend to evaluate their model's accuracy either with **Retrieval** or **Ranking** metrics on all possible user to item interactions. Retrieval metrics include Precision, Recall and ROC Curve metrics that are useful for indicating the overall performance of the model. Ranking metrics including nDCG, Mean Reciprocal Rank and Average Precision are more appropriate for tasks where the order in which the items are placed is important.

Recently, more and more researchers are stressing the importance of going beyond solely evaluating RS in terms of their retrieval or ranking accuracy and evaluating the Diversity of recommended items. As back as 2010, Zhou, T et al., introduced the metrics of Novelty and Personalisation [34]. **Novelty** indicates the ability of the RS to recommend novel and less known items that the user was unlikely to have previously encountered. **Personalisation** calculates the rate of dissimilarity between the lists suggested to different users. It requires a measure of distance such as cosine similarity and the highest scores, meaning the highest dissimilarity between recommended lists, translate into greater levels of personalisation. Furthermore, Mouzhi Ge, et al. in a paper aptly called 'Beyond Accuracy' proposed various metrics of **Coverage** indicating the percentage of items that the RS is able to recommends from the totality of known items meaning that if a RS has a Popularity Bias it will not be able to score high in terms of Coverage [35]. It is important to note, that there is a possible trade-off between predictive accuracy and metrics of diversity but a working RS should ideally have high values in terms of Diversity while at the same time maintaining high levels of Accuracy.

2.2 Systematic Review Procedure

An extensive survey was published in 2018, discussing the 'current challenges and visions in music recommender systems research' [36]. The paper presented the trending research topics in MRS, the main challenges they face and various proposed solutions to said challenges. Among them, Markus Scheld et al., discussed the 1) Cold Start Problem, 2) Automatic Playlist continuation and the 3) consistent evaluation of MRS. A substantial body of research was discussed for each topic and the proposed solutions, while naturally having limitations, were workable. What the researchers considered under-researched while being fruitful topics for future research was threefold including 1) 'Psychologically inspired music recommendation', 2) 'Situation-aware music recommendation' and 3) 'Culture-aware music recommendation'. Therefore, for the purposes of the present diploma, it was deemed appropriate to focus on the aforementioned 'open subtopics' and 'future directions' and to examine how their research has progressed since. After selecting the general topics of interest, namely contextual and psychology-inspired recommendations for music, the procedure of systematic reviews, specifically Kitchenham's methodology [37], was partly employed in order to structure and organise the process of finding, filtering and categorising relevant papers.

For retrieving the most relevant research papers, the search engines of *Scopus*, *Elsevier*, *ACM* and *IEEE Xplore* were used as sources under the most relevant categories of 'Computer Science', 'Data Science' or 'Artificial Intelligence'. Google Scholar was also used in case some relevant research work was missed. Moreover, the selection criteria were that the study should be relatively recent, between 2015 and 2020. That is because previous works on the selected topics were, to some extent, already analysed by the aforementioned review. Furthermore, the full text should be accessible and written

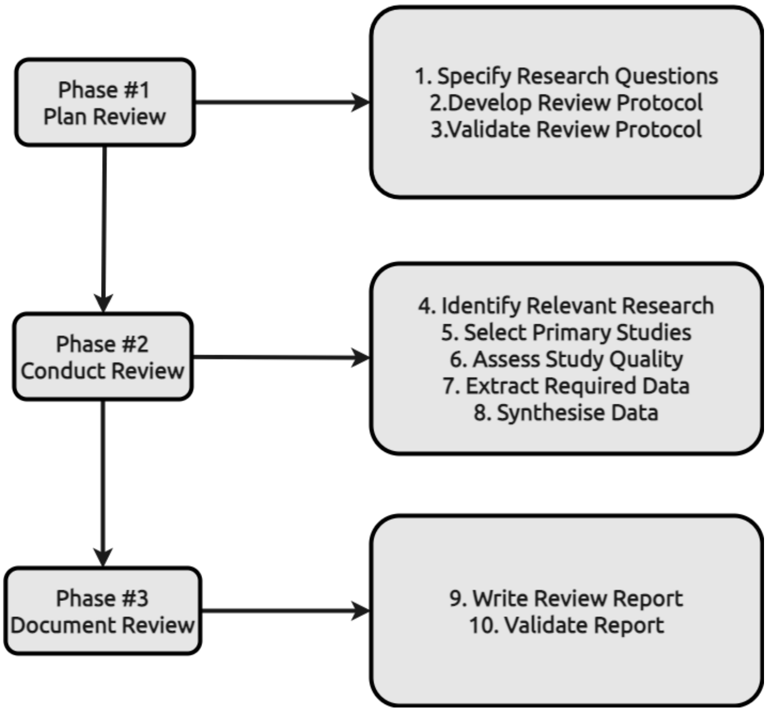


Figure 2.3: Kitchenham's Systematic Review procedure

in English. Papers were first filtered by reading the title and the abstract. The criteria was that the study should be related with the field of Recommendation Systems for Music and also address at least one of the following issues : Personality Modelling, Mood Detection, Emotion Analysis, Contextual- or Situational- or Cultural-Awareness, Cross Platform, Cross Domain or Data Fusion methodology. Only the papers addressing at least one of these sub-topics were selected for a full reading. While collecting and reading the most relevant papers for each subtopic that met the selection criteria, the papers were categorised and if important papers were found in the references, they were also collected. In the process, the categorisation of papers was evolving and changing as new information and types of papers were analysed.

By attempting to categorise the whole body of collected research works it could be said that all, with the exception of Data Fusion-related studies, were concerned with a different types of contextual information. The source of contextual information is either **Extrinsic** or **Intrinsic**, meaning they are part of the user's external environment or they are a part of their internal psychology. In the literature, Intrinsic Context is frequently referred to as Psychometric- or Psychology-Inspired RS. On the other hand, the temporal nature of contextual information can be categorised either as **Static** or as **Dynamic**. The user's personality or cultural environment could be considered relatively static and stable categories while emotions, moods or time of day and weather are dynamic phenomena. This categorisation could be summarized in the form of a 2 by 2 matrix.

	Static	Dynamic
Internal	Personality-Based	Emotion-Aware
External	Culture-Aware	Environment-Aware

Table 2.1: Mapping Contextual Information.

This framework could be helpful for understanding and contextualising these seeming unrelated subtopics under the same family of context-aware recommenders. However, in current literature, it is more common to refer to 'Context-Aware' RS only to those related to dynamic external context, while RS relying on internal context is usually referred to as Psychology- or Psychometric-Based RS. Therefore, for the rest of the literature review, they will be referred with the most commonly accepted categorisation framework.

2.3 Related Work

Following the mapping ensuing from the Systematic Review procedure, discussed in Section 2.2, the collected research works that are recent and relevant to the present work, were categorised into:

1. Context-Aware Music Recommendation Systems (CA-MRS)
2. Psychometric-Based Music Recommendation Systems (PB-MRS)
    - Emotion-Aware
    - Personality-Based
3. Data Fusion-based Music Recommendation Systems (DF-MRS).
    - Cross Platform
    - Cross Domain

CA-MRS are recommendation systems that take into consideration various contextual factors and information of their users with the goal of increasing personalisation, prediction accuracy and user satisfaction. The rationale of contextual systems is that a user's musical preferences may vary depending on the situation they are in, for example while working, studying, doing a physical activity or going to sleep. Without explicit input from the user, a system may not be able to know the exact situation of the user, but contextual-data such as approximate location (extracted from IPs), weather reports and log files (timestamps) can work as reliable proxies and identify useful patterns in a user's behavior. Contextual factors can be further divided depending on their temporal nature into Static or Dynamic, meaning that they are relatively stable and unchanging (such as Culture) or Dynamic (Time and Location) respectively.

PB-MRS attempt to utilize psychometric analysis tools (Emotion Analysis, Mood Detection or Personality Detection) to extract information related to a users' psychological makeup, in an a attempt to improve the accuracy and personalisation of RS. PB-MRS can be further categorised into Emotion-Aware and Personality-Based MRS. The rationale of developing Emotion-Aware systems is that selecting musical items is significantly linked with and dependent on a user's current mood and emotional state. Thus accessing the emotional state and reactions would improve the predictive accuracy in real-time recommendations. On the other hand, Personality-Based MRS rely on the idea that music categories and characteristics are correlated with different Personality Traits that can then be used as a static user-side features, similarly to demographic information. Due to the fact that both Emotion-Aware and Personality-Based MRS have been extensively researched in recent years, they will be discussed separately, in two different sections.

Finally, DF-MRS are utilized in order to address difficult challenges and solve specific issues resulting from algorithmic biases or lacking information in the target platform. DF-MRS can be further categorised into **Cross-Platform** or **Cross-Domain** depending on whether the data are fused from different platforms of the same domain (different music related platform in our case) or from different domains altogether (e.g from Movies or Books datasets). In both cases, data are collected from a different source and are fused into the target platform's existing data in a way that can improve its performance or ameliorate some RS-specific challenge such as Data Sparsity and Cold Start.

2.3.1 Context-Aware Music Recommendation Systems

Applied in the domain of music [38] examined the correlations between musical preferences in relation to the locations they were being selected in order to identify culture specific listening patterns. Culture-specific features were extracted from socioeconomic data sources like the World happiness report (which included GDP, freedom, healthy life expectancy, generosity, social support, corruption, happiness). The music preferences were collected from Twitter's "Currently listening on spotify" feature and were represented by their acoustic features. When the acoustic features and the socioeconomic data were correlated with the users' locations some significant correlations were identified giving credence to the idea of context-aware MRS.

To that end, [39], developed a contextual MRS that tracks each user's musical preferences in relation to time of day, location, weather and season that a musical item is selected. This model utilizes a heterogeneous information network (a graph-based with different types of edges and links) that consists of a Topic Extraction stage and a Recommendation stage. First the current contextual information is collected and the most appropriate type of music given these parameters is selected. Secondly a Personalized PageRank algorithm identifies suitable songs the selected 'Topic'. When tested against Item-based CF, Matrix Factorization, Bayesian personalized ranking and Latent Semantic Indexing the proposed method outperformed them scoring a MAP of 0.34 and NDCG of 0.23 with Item-based CF coming second with 0.33 and 0.17 respectively. The researchers also claim that when applied in a Mobile Network this approach can mitigate the New User Cold-start based on 'current mobile environment.

Dirapisut et al., based their model on the idea of Micro-profiling, creating separate recommendation profiles for each user for different time windows. Their rationale is that while user's preferences may vary there still exist some stable time-related patterns and that micro-profiles can identify them [40]. The history of each User-Item interactions are segmented at the level of daytime (morning, afternoon or evening), week (weekend or weekday) and year (cold or hot season) with a total of 6 micro-profiles for each user. Furthermore, in order to overcome the usual cold-start problem present in CF models, the researchers proposed an alternative known as Tendency-based CF. Instead of calculating the similarity between all to all users, this approach calculates the tendency of each user in relation to their rating habits (generally rating positively or negatively) and similarly how each item tends to be rated. Trained on a dataset - relatively small dataset - of 357 last.fm users - the proposed model required a lower computation time from both training and prediction speed. When compared to CF and MF, Tendency-based CF was on average better performing and had lower complexity for both training and prediction time.

2.3.2 Emotion-Aware Music Recommendation Systems

For emotion-aware systems to work, there must be an emotion-related source that can derive and later analyse the user's current emotional state. There is a wide range of potential sources since emotions can be detected from Wearable Physiological Sensors [14], Face Emotion Recognition [15] [13]. Social Media posts and Mood-tags which can in turn be analysed by lexicon-based approaches or machine learning approaches trained on labeled corpuses [9]. Additionally, there is the research field of Music Emotion Recognition (MER) that analyses music and its acoustic features like pitch, intensity, timbre, tempo, rhythm etc [41] or by analysing music lyrics.

Depending on the selected source of emotional input, a different processing technique must be developed and different issues must be addressed. For example, compared to musical tracks, music lyrics are an easily accessible and easier to process feature. Lyrics is a unique form of written expressions that differs significantly from other forms such as newspaper articles, research papers or social media posts. Lyrics are more similar to poems, containing colloquial and format-specific expressions. Lyrics can be abstract, cryptic or metaphorical and may contain wordplay or rhymes. As a result, typical lexicon-based sentiment analysis approaches find it difficult to analyse and correctly detect their emotional state [42]. Contrarily, creating manually labelled lyric corpuses is a costly, time consuming and labor intensive process and as a result lead to corpuses with limited size. At the same time, mood-tag-based can be used as a stand-in for manual labeling however this approach may suffer from the issues of Synonymy and Polysemy where a general agreement on the meanings of tags is lacking [43]. Researchers have to navigate these issues and balance between competing trade-offs of different emotion sources and emotion analysis methods.

Apart from requiring an input source, each emotion analysis method requires a theoretical framework for modelling emotions. Frameworks for classifying mood and emotional states can generally be divided into two types: Categorical and Dimensional. Categorical consists of separately defined emotional states such as sadness, joy and anger while dimensional which maps an emotion into a two-dimensional space. Such a model is James Russell's Circumplex which consists of two axes : Valence (negative-positive) and Arousal or Excitement (Low - High). Similarly, Thayer's model comprises a Stress axis (Cheerful / Restless) and Energy axis (Smooth / Vitality) [43]. The researchers of *MoodPlay* have criticised the over-use of Russell's model for MER [10]. They claim music is too multifaceted and emotionally complex that can not be captured by Russell's model. They present the example of Fear and Anger, both emotions with Negative Valence but High Arousal and differentiating between them in the two dimensional circumplex is difficult. Instead, they propose the use of GEMS (Geneva Emotional Music Scale) model, a more music-oriented emotional framework.

Finally, it is important to define the distinction between Emotion and Mood. Most psychologists agree that there are significant differences between the two terms concerning their duration, intentionality, cause and consequences. More specifically, emotions are considered short-lived mental states that are caused by some internal or external trigger. On the other hand, mood is considered a more 'general and background' affective state of unclear cause that tends to have longer duration [44]. Furthermore emotions can be detectable by facial expressions but moods can not [45]. Despite that, many researchers in recommendation systems use these terms interchangeably.

Applied in the field of recommendation systems, Deng et al., combined both extrinsic and intrinsic context by correlating Emotional states with music preferences in different time windows [11]. The central idea of this approach is to identify musical items that similar users prefer when under similar emotional state. User's emotions were extracted from microblogs and represented into granular form. More specifically, Ekkman's granular emotion framework was used, where the emotions can be expressed in 2, 7 and 21 dimensions. Subsequently the association between user, emotion vector and music type was represented by a three-tuple relation. The researchers compared an Emotion-Aware user-based CF with six other models, including Item-Based CF, User-Based CF, two Graph-Based filtering models, Bayesian Personalized Ranking and a Hybrid model combining UBCF and IBCF. The UBCF utilizing user emotional information showed a significant improvement of both HitRate and Precision compared with all other models. This strongly indicates the importance of emotion features for MRS. A second empirical evaluation examined the effect of different time windows in the model's performance. Different emotion-granularities had a relatively small effect on the outcome with a medium time window (3 to 5 hours) performing slightly better than smaller (1-2 hours) or larger (9 - 24hours) time-windows due to sparsity of posts in the first case and the introduction of noise in the second case.

Rosa et al., developed a smart media player based on the intensity of current emotional states in relation to demographic information in order to improve user satisfaction [9]. The model calculated the intensity of the expressed sentiments that were extracted from the user's social media profile with the use of a hybrid sentiment analysis method that utilized Russell's circumplex as the emotional framework. Furthermore, a correction factor function was developed that took into account a user's demographic information: age, educational level and gender. These data were obtained by a user study and by analysing posts from their social media profile and were used to inform a 'correction factor' function. Depending on the current emotional state and intensity of a user's and factoring in their demographic information the recommendation engine was able to improve User Satisfaction as reported by 78% of participants.

Andjelkovic et al., developed an Artist RS with an interactive affective virtual interface relying on the emotional relationship and similarity between the artist and the listener's current state [10]. The proposed model utilized a two-layer cascade hybrid filtering approach. The first level required the user to declare her current emotional state and select a number of her preferred artists. Then, the User-Artist mood similarity was calculated and the artists with dissimilar Mood-Tags to the user's current mood are filtered out. Based on the filtered list the second level calculated the Artist to Artist similarity based on their Acoustic and Musical Features (timbre, tempo, loudness etc) with the use of kNN algorithm. This model was evaluated by an online user study. Regarding 'predictive accuracy' it yielded a relatively low average of 36.2 - 49.8. As noted before, Content-based Filtering generally scores lower than Collaborative models and since the proposed model was a hybrid of two content-based filtering models the restrained predictive power may be expected. However, an interesting finding was that the same MRS was rated higher on 'User Satisfaction' and 'User Control' when an Interactive Virtual Space was added.

While the three previously discussed emotion-aware systems were focused on improving the accuracy and by extension the user satisfaction of 'traditional' MRS, the study of De Assuncao et al., differs in that it tries to alter the user's current mood towards a desired state [19]. Current user emotion was averaged between self-report and a filled questionnaire. Based on that, the system generated a music playlist informed by a distance function calculating the most effective track-list. The evaluation was based on a user study where the participant's emotions were assessed on face emotion recognition, a questionnaire and self-report. On average, among 20 participants, there was a 53.7% improvement, towards the desired emotion based on self-assessment and 68.5% based on questionnaires. Nevertheless, there was no 'control test' to sufficiently assess the efficacy of the proposed method. Most participants initially were in a negative emotional state (in terms of both Valence and Arousal) and were able to improve towards a higher Valence and Arousal, a result that could easily be interpreted as that continuous listening to music may in general improve the participants mood.

2.3.3 Personality-Based Music Recommendation Systems

In psychological research, the personality of a person is considered the relatively stable and coherent set of psychological characteristics that can describe her behavioral, emotional and cognitive patterns. Traits-based theories of personality do not attempt to describe a person as a 'type' but rather make use of personality traits that can predict the subject's behavior in various aspects and degrees. One such model is known as the Big Five Personality Traits consisting of: Openness to experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism. Generally, Personality Traits can be acquired either by Explicit Means (the subject answering to a specialised questionnaire) or Implicit Means (by analysing user-generated content like social media posts and microblogs). Since Personality can describe and predict certain aspects of an individual's behavioral and decision making patterns, there have been attempts to utilize Traits in Recommender Systems in order to enhance personalization and predictive accuracy [46].

More specifically on Music Recommendation Systems, Paudel et al., studied the impact of user's personality traits on Collaborative Filtering methods [47]. In order to extract user personality features, a Naive Bayes classifier was trained on the 'myPersonality' dataset and then used to analyse the user's textual posts from Facebook on the Big Five model. Thereafter, a user to user CF model calculated the cosine similarity between users by taking into account the rating matrix and personality vector. Six different feature combinations were given input to a UBCF model and then compared against Global Baseline Estimate and Matrix Factorization. Ratings-based CF yielded a RMSE score of 3.89 while personality-based CF outperformed it with a RMSE of 3.20, a noticeable improvement signifying the potential usefulness of personality in MRS. However, more advanced methods like Global Baseline Estimate (RMSE : 2.86) (a method that takes the total average score and calculates each user's deviation, as well as the deviation of each specific item from the total average) and Matrix Factorization (RMSE : 0.88) clearly exceeded Personality-based CF.

Similarly, Onori et al., developed a user to user memory-based Collaborative Filtering model based on personality but experimented with both explicit and implicit Big Five personality traits [8]. A user study was carried out and participants' explicit traits were acquired by the '44-item Big Five' questionnaire while the Implicit traits were extracted by analysing their Facebook textual posts with the use of Cambridge Psychometric 'Apply Magic Sauce' application. Four different models were developed: 1. Association rules between Personality traits and Music Genres, 2. Collaborative Filtering based on explicit personality similarity, 3. Collaborative Filtering based on implicit personality similarity and 4. Baseline Collaborative Filtering. The participants were asked to evaluate the results of each model on the basis of Novelty, Serendipity, Diversity, Interest and Future Use. The mean scores among personality based and baseline CF regarding Novelty, Interest, Future Use and Serendipity were very close but the one significant difference was found in Diversity. The Personality-based Models had an average of 3 while the baseline CF scored 1.7 out of 5 indicating the possible usefulness of personality traits in mitigating the lack of Diversity in MRS. Nevertheless, a notable limitation of the study was its small number of participants, 65 in total of which 22 did not have enough Facebook posts so as to calculate their implicit personality.

Lu, Fend et al., study more thoroughly the degree of musical Diversity that people with different personality traits prefer and how this statistical relationship may affect MRS [48]. The researchers carried out a pilot user study with 25 participants where their personality traits were studied in relation to their musical preferences. The explicit personality traits were acquired from the 'Ten Item Personality Inventory' which is also based on the Big Five model while their musical preferences were represented in terms of their acoustic features (key, tempo etc) and their meta-data (musical genre, release date etc). The findings were that Emotional Stability shows a positive correlation with 'Artist', 'Tempo', 'Genre' diversity as well as the overall level of diversity. Extraversion is positively correlated with Key changes and Agreeableness with the number of different Artists. These observations, about the relationship between personality and diversity, were used as a parameter in an objective function that informs a diversifying greedy heuristic that tries to minimize the balance between similarity and diversity. In this study, Diversity was defined as the dissimilarity between musical items that belong in the recommendation list and not the whole dataset. In simpler terms, these items are similar to the user's general preferences but are different enough between them to create an engaging list that will maintain the user's interest. Afterwards, a recommendation engine, based on Factorization Machines, produced a list and the diversifying heuristic adjusted this initial list and created a re-ranked recommended list. In a second evaluation-stage user study, the subjects rated higher the diversity-based reranking lists on their Quality, Perceived Diversity and User Satisfaction compared to the initial list.

Cheng, Rui et al., also studied the relationship between personality traits and musical preferences and their effects on MRS. In their online user study, the assessed personality traits included the Big Five with the addition of Social Dominance, Depression, Self-esteem, and Intelligence [49]. The participant's musical preferences were represented by their acoustic features (Pitch, Roll-off, Rhythm and other) and converted into a vector space. Both personality traits and acoustic features were used as input features in a Support Vector Regressor (SVR) model. When the two features were used separately they yielded an accuracy score of 75.2% and 77.9% as well as a MAE score of 0.84, 0.88 for personality and acoustic features respectively. But when the two were combined, the Accuracy increased to 89.7% suggesting a significant relationship between personality traits and acoustic features. Furthermore, the SVR model was able to outperform the Baseline Collaborative Filtering model which scored 1.12 in MSE and 70.2% in accuracy. Since the proposed method is Hybrid, a combination CF and CBF, the researchers claimed that it did not suffer from the Cold Start problem or Data Sparsity since SVM does not rely on the number of features but rather on the 'margin which separates the data'. Another potential criticism is the lack of a baseline model as a point of comparison, to more precisely assess the usefulness of personality traits and their contribution to the model.

2.3.4 Data Fusion in Music Recommendation Systems

Data Fusion is the process of combining different sets of data from multiple sources in order to create a more concrete and substantial dataset. When applied in the context of recommendation systems data fusion approaches can be divided into two categories, **Cross-Platform** or **Cross-Domain** depending on whether the selected domains are of the same or different source-type.

In the context of music recommendation systems, Cross-Platform fusion utilizes two different platforms or sources related to the same domain, namely Music. In this respect, Oramas, Sergio et al., in an attempt to mitigate the cold-start problem of collaborative filtering the researchers combined the audio features from one source and artist's biographies from another [16]. Their rationale was that artist's biographies and interviews are readily available and generally unutilized information for enhancing item-side meta-data and that could be useful for MRS. The proposed recommendation process is divided into two steps. First, a Matrix Factorization method filters the most relevant artists for each user. Secondly a Multimodal Neural Network selects the most appropriate songs based on the artist's textual and semantic embeddings and music track embeddings. Empirically, Multimodal MLP with both artist's semantics and audio embeddings outperformed all other feature and method combinations.

On the other hand, Cross-domain Recommendation Systems (CDRS) aim to identify and utilize additional and useful information from different recommendation domains in order to improve the target domain. CDRS have generally been applied in order to solve Cold-Start of either Item's or User's sides or Systemic Cold-Start - when a RS is new and with a small number of users and few rated items. One of the main challenges of CDRS is to confirm that the knowledge transferred from a different domain is applicable to the target domain. In order to address this problem, Zhang et al., developed a Consistent Information Transfer method that clusters similar groups of users and items from both domains and identifies latent factors between user to item groups [18]. It was tested with 5 different datasets from three different domains (Movies, Books and Music) and was able to increase the predictive Accuracy even with high-sparsity data. Similarly, a Hybrid CDRS (CF + Association Rules) between two different domains (Music and Books) was able to outperform traditional CF and the Association Rules could provide better recommendations for new users with very few ratings (cold-start) [50].

2.4 Reflections on Current Literature

In this section, the research works presented in the Related Work are discussed holistically in order to identify interesting and worth investigating research paths.

Starting off with **Context-Aware MRS**, current research works have demonstrated that integrating contextual information about the users' external environment, such as

	Objective	**Challenge**	**Model**	**Input**
Emotion-Aware				
Rosa et al. [9]	-	User Satisfaction	Correction Factor	User Study & Social Media
Andjelkovic et al. [10]	Rating Prediction	User Satisfaction	Similarity Content-Based	Mood Tags
Deng et al. [11]	Retrieval	-	Collaborative	Social Media
Assuncao et al. [19]	Retrieval	Playlist Generation Desired Mood	Distance-Based	
Context-Aware				
Pichl et al. [38]	Retrieval	Personalisation	Clustering	Social Media
Wang et al. [39]	Ranking	Cold Start & Interpretability	Graph-Based	Mobile Network
Darapisut et al. [40]	Retrieval	Scalability	Tendency-Based Collaborative	-
Personality-Based				
Onori et al. [8]	Rating Prediction	Diversity	Collaborative	Social Media
Paudel et al. [47]	Rating Prediction	-	Collaborative	Social Media & User Study
Lu and Tintarev [48]	Retrieval	Diversity	Factorization	User Study
Cheng and Tang [49]	Rating Prediction	Data Sparsity	Content-Based	User Study
Data Fusion				
Oramas et al. [16]	Ranking	Cold Start	Hybrid Filtering	-
Zhang et al. [18]	Rating Prediction	Cold Start & Data Sparsity	Factorization	-
Zhang et al. [50]	Retrieval	Cold Start & Diversity	Collaborative & Association Rules	Social Media

Table 2.2: Summary table for the literature review

their current location, the time of day they select to listen to music, the weather among other factors can help improve the predictive performance and personalisation of the RS. Listening to music is highly situation specific since people tend to prefer musical items with different attributes depending on their current situation be it the workplace, the gym or relaxing at home. Furthermore, some of the fundamental challenges of RS such as scalability [40], cold start and intractability [39] have been addressed in recent years. Altogether, research on CA-RS for MRS has already reached a mature enough level to the point of being deployed in production even, by highly successful streaming platforms such as YouTube [6] and Spotify [7].

Regarding the recent research centered around **Emotion-Aware MRS**, four out of seven works, were dedicated to experimenting with different sources of accessing users' emotional states or reactions. Ayata et al. [14] used Wearable Devices, Iyer et al. [13] and Gilda et al. [15] used Face Emotion Recognition while Deng et al. [11] used users' Microblog posts and discussed how they can be integrated in existing recommendation models or how to be deployed in applications [13] [14] [15]. Deng et al. [11] also discussed how emotional information can be integrated into user-based collaborative filtering models which proved to outperform all other models in the study. All other models in the study were either CF-based, Graph-Based or Bayesian-based. While the study was able to validate the hypothesis that emotional features can improve the performance of UB-CF, the study's scope and generalisation is limited since Memory-Based CF models are not anymore considered to be among the State-of-the-Art in Recommendation Systems. We can not know how the model would compare against Matrix and Neural Factorization models. Rosa et al. [9], proposed a new metric for assessing users' emotional states and compared it with two other metrics. Their proposed method was supported by a subjective user study in which participants rated their general satisfaction with the generated playlists of the different metrics. Nevertheless, the proposed metric was not incorporated and tested in an existing recommendation algorithm. Andejelkovic et al., in [10], examined how EA-RS can help improve user experience in interactive music players. Additionally, the researchers performed a comparative analysis between various models of which one was a Context-Aware Matrix Factorization tuned with explicit user emotions (tags selected by the users' themselves). This model was able to outperform UB-CF in terms of RMSE but was then outscored by a Factorization Machine-based model that utilized item-side mood data. The researchers also used a hybrid model that calculated 'user mood' to 'artist mood' similarity that was then used as input feature in a content-based filtering model that calculated artist to artist similarity. This model while under-performing in the offline evaluation, it was rated higher in terms of User Satisfaction and User Control in an online user study of 279 participants.

There has been an increased interest in EA-RS for music and the growing body of research work is showing the prospective usefulness of emotional features in MRS as well as various potential sources of accessing users' emotional reactions.

However, what is **lacking** from the research on Emotion-Aware Systems are the following

1. Assessment of how emotional features work with **State-of-the-Art** RS models such as Factorization Machines and Neural Factorization models,
2. Examination of how main challenges of RS, such as **Data Sparsity**, **Diversity** and **Item Cold Start**, specifically affect EA-RS for music.

Furthermore, a noteworthy limitation of recent works is the use of '**idealised**' datasets where no emotional values are missing hence there is no emotional data sparsity. However, in a real-world EA-RS there will be instances that a portion of the users will not be connected with the emotional source, be it Wearable Devices or Social Media posts. This issue is frequently discussed in the literature, but only from the user's point of view. If an individual user is not connected to the emotional source, then she will not be able to receive emotion aware recommendations and thus a traditional approach will have to be applied. Nevertheless, the long-term consequences of this phenomenon are not considered. If multiple cases of missing emotional values are aggregated it could potentially harm the performance of EA-RS due to Emotional Data Sparsity. Furthermore, this could also reinforce an Item Cold Start situation where new and less popular items have not been linked with enough emotional reactions and thus will not be able to be recommended.

Moving on to **Personality-Based** MRS, based on the current research, assessing the potential usefulness of personality traits in MRS could be argued to be inconclusive. First, Onori et al. [8] relied on a user studies of 65 participants for accessing the personality traits of users, training and evaluating the RS. The end results were mixed and the sample size was too small to be able to generalize their findings for Collaborative Filtering methods, known for requiring significant amounts of data to function properly. Furthermore both Paudel et al. [47] and Cheng and Tang [49] used Memory Based Collaborative Filtering and simple distance functions for integrating Personality Traits in finding neighboring user. Nevertheless, Memory-based CF are not in accordance with the current State of the Art in MRS which mostly consists of Factorization Machines, Neural Networks and Hybrid Filtering. Paudel et al. [47], even though personality traits were able to reduce the RMSE to 3.04, from 3.89 of the baseline CF model, a Matrix Factorization model yielded a RMSE of only 0.88. Thus we can not conclusively evaluate the potential usefulness of Personality Traits on State of the Art MRS models, based on these research works. Cheng and Tang [49] utilized a Hybrid Approach, based on a SVM model. In this study, using Personality Traits, slightly improved upon 'Directional Accuracy' but worsen in terms of RMSE when compared to only using Acoustic Music Features. While using both Personality Traits and Music Acoustic Features was able to improve the overall performance, there was no baseline model so as to be able to conclusively evaluate the importance of Personality Traits. Of the four recent research works on Personality-Based MRS, Lu and Tintarev [48] could be argued to be the most rigorous and informative. The researchers developed a Factorization Machine-based (in accordance with the SOTA) model whose performance was enhanced with the use of User Personality re-ranking function. The personality based model was able to outperform its baseline with a 'Hit Rate at 10' of 0.141 compared to the 0.043 of its baseline. Additionally, the model was evaluated in an online user study (of 25 participants) that similarly scored better in terms of both Recommendation Quality and Recommendation Diversity. Nevertheless, the initial training dataset consisted of the musical preferences of only 148 participants. Apart from the mixed results, all works on Personality-Based MRS analyzed and utilized users' personality trait. An idea that has not yet been explored is acquiring the personality traits of music artists and studying the potential relationship between users and the personality of their favourite artists.

Finally, in the case of **Data Fusion-based** MRS, in all recent research works, Data Fusion was used in order to address the challenges of Cold Start while Zhang et al. [18] attempted to mitigate the problem of Data Sparsity and Zhang et al. [50] also discussed the potential usefulness of Data Fusion in improving the Diversity of MRS. Generally, DF-MRS has by now proven to be a potent solution for multiple of the fundamental challenges of RS as well as flexible since it can work in combination with various RS approaches. While Oramas et al. [16] developed a method for profiling artists based on their biographies, an idea that was not present in any of the works, is the idea of Multi-Level profiling, meaning the fusion of data for multiple levels of item and user representation. Specific to the domain of music, utilizing all track-level, album-level and artist-level could proven useful in representing the multifaceted relationship between users and musical items and by extension improve the overall model's performance.

In summation, **Context-Aware** MRS have been researched in more depth and are already being deployed in production in contrast to **Emotion-Aware** and Personality-Based models that are require further research. **Personality-Based** MRS have shown some promise but the results are arguably inconclusive. On the other hand, recent research EA-RS for music, has proven the prospective usefulness of emotional features but there exist under-researched areas in relation to the fundamental challenges of RS such as Cold Start, Diversity and Data Sparsity. One important challenge that has not been addressed is that of Emotional Data Sparsity resulting from Missing Emotional Values that could severely hinder the performance of EA-RS when applied in real-world applications. Furthermore, most proposed Emotion-Aware models did not keep up with current State-of-the-Art practices and models, such as Factorization Machines and Neural Networks and this should be noted and explored. Finally, **Data Fusion** has by now proven a formidable solution to the challenge of Cold Start and in improving the issues of Data Sparsity and lack of Diversity but they have not yet been used in the context of EA-RS.

Considering all of the above, the process of the systematic review has revealed an important research gap and a new and potentially fruitful research path. One centered around Emotion Aware MRS that are able to utilize **Multi-Level Profiling** of items - among them - the personality traits of artists, the emotional expression of song lyrics and the mood of the music, collected from a Data Fusion pipeline in an attempt to study the issue of **Emotional Data Sparsity** and overcome the resulting issues of Item Cold Start and Popularity Bias that have not been studied in the context of EA-RS. Furthermore, this research path involves the introduction of the idea of a '**Cross Platform Audience Reaction'**, where the average emotional reaction of an audience in relation to each specific musical item is collected and used as an input feature in hybrid RS algorithms. The rationale behind this idea is that for items where no previous emotional interactions are known in the target platform, the average emotion of a cross platform audience could work as a proxy and inform the Emotion-Aware model of how users tend to react on this particular item. Thus mitigating the effects of **Item Cold Start** and **Popularity Bias** in EA-RS for music. In essence, these proposals attempt to overcome a previously neglected challenge in the research of Emotion Aware Recommendation Systems by synthesising the most recent research on Hybrid Algorithms, Personality-Based Recommendation and Cross Platform Recommendation Systems. The necessary steps and processes for studying this issue and the development of the aforementioned proposals are described in detail in the next section.

Chapter 3

Developing Emotion-Aware Recommendation Systems for Music

In this chapter, the necessary components are discussed for developing Emotion-Aware Recommendation Systems for Music relying on algorithms with Hybrid architecture, able to integrate both Collaborative and Content-Based Information. This process also includes, the data collection process (Section 3.2) for accumulating a dataset appropriate for recommendation tasks that also pertains the emotional reactions of real-life users in relation to musical items (U-I interactions). It also encompasses the development and deployment of emotion analysis models for extracting said emotional reactions from users' textual data (Section 3.3). And finally, the integration of the extracted U-I interactions and the extracted Item features into Factorization Machines and Two-Tower Neural Networks, two widely used hybrid models (Section 3.4). But before doing so, it is important to clearly define the motivation and rationale for attempting this research endeavour.

3.1 Problem Formulation

As was discussed in the Introduction, the central objectives of this dissertation, is 1) the development of an Emotion-Aware Recommendation System (EA-RS), specialized on the domain of Music, 2) the study of challenges specific to Emotion-Aware Systems and 3) the development and evaluation of practicable solutions. The working hypothesis of this work is that aggregated cases of "Missing Emotional Input" in EA-RS lead to the phenomenon of Emotional Data Sparsity hindering the performance of the systems, affecting both the predictive accuracy and the reinforcing Cold Start situation for new and less popular items. This situation further contributes in the “Popularity Bias” and “Grey Sheep problem” common in Collaborative Filtering approaches.

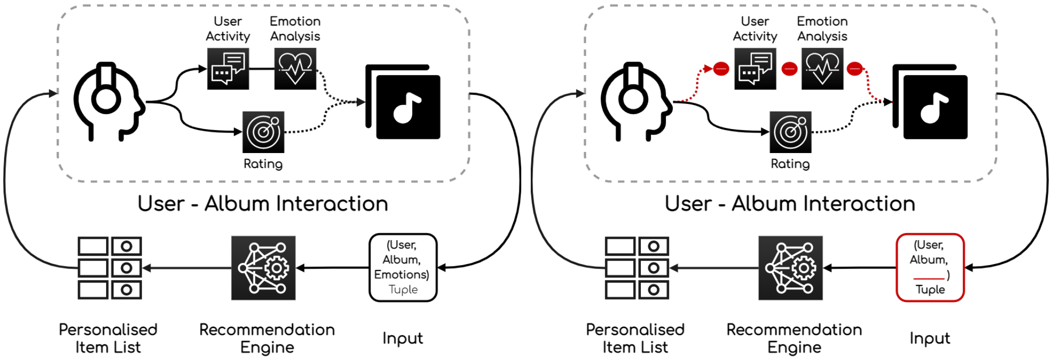


FIGURE 3.1: The expected workflow of an Emotion-Aware System (left) and a case of a missing emotional input (right) that leads into a Missing Emotional Value.

More specifically, as discussed in the Literature Review, EA-RS require an **Emotion Input Source** for recognizing the emotional states and reactions of its users. The input source may be Social Media posts, Face Emotion Recognition, Wearable Devices or User Reviews. However, in real-world applications, for each of the possible input sources there may be a significant amount of missing values. Multiple users may not post frequently on their social media or rarely wear their wearable devices for their activity to be emotionally analysed. When aggregated, these missing values may hinder the performance of the system. This issue is frequently discussed in the literature but only from the User’s perspective [19]. If a User is not connected with the “Emotion Input Source” and the system lacks access to her current emotional state and thus is not able to perform Emotion-informed recommendations. It is commonly agreed upon that this issue can not be overcome and the accepted solution is to perform recommendation based on the user’s past interactions similarly to a conventional recommendation system.

What is lacking in the Literature is examining the issue of **Missing Emotional Values** from the Item’s perspective and its long term consequences. Since an EA-RS is trained on the relationship between **(User, Item, Emotions) N-tuple** interactions, and New and Cold Items - with which the users have not yet interacted - have not been linked with specific emotional responses, can thus not be recommended. The system does not know what emotional responses a New Item stimulates to different groups of users, nor under what emotional states various Users choose to interact with the new Item. This situation is related to the Cold Start of “traditional” Recommendation Systems but specialised on EA-RS.

The proposed solution for the aforementioned challenges involves the experimentation with Hybrid Recommendation Algorithms,Multi-Level Profiling and Cross-Platform

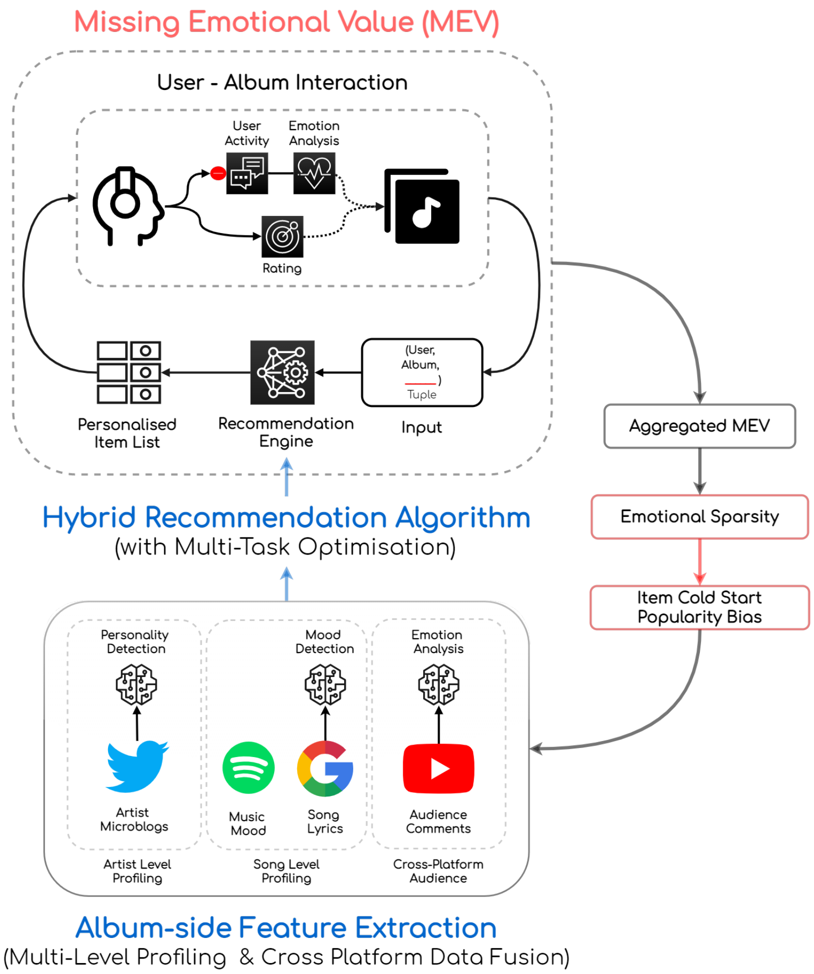


FIGURE 3.2: The consequences of aggregated missing emotional values leading to emotional data sparsity and reinforcing new item's Cold Start and Popularity Bias (Red) and the proposed solution (Blue).

Audience Reaction. Firstly, a **Hybrid Models** are required, since the employed model should be able to utilize both collaborative interactions, dynamic User-side information - emotional states in particular - and Item-side metadata. Hybrid models have been proven better equipped in dealing with new and less popular items by utilizing the content-based information while maintaining high levels accuracy by leveraging collaborative information [4]. Secondly, **Multi-Level Profiling** is an attempt to capture the multifaceted relationship between Music and its Audience, since a user may drawn upon a musical item for its Musical, Emotional, Lyrical content, its Genre Classification, its Popularity or lack thereof, the projected Personality of the Artist and others. Thirdly, **Cross-Platform Audience Reaction** is the idea of collecting and analysing the emotion response of the audience of a Musical Item from a different Music-Related source and using the extracted average emotional response as an input feature in the Target platform. Both ideas are extensions of cross-platform data fusion techniques which have been shown capable in dealing with the cold start of new items [16]. The fundamental premise is that a Hybrid System will be able to identify complicated patterns and relations between Users, their Emotional Responses, the Item's Multi-Level Features and the Cross-Platform Audience Reaction and will be instrumental in mitigating the Cold Start of New Items. Furthermore, it is hypothesised that the described architecture will increase the overall predictive performance of the model, as well as improve the Personalization and Novelty of the recommendations and alleviate the Popularity Bias.

This Chapter is dedicated to the in-depth description and outline of the followed methodology including (1) the creation of a real-world feature-rich dataset pertinent for Recommendation Tasks (2) collected through a Data Fusion pipeline involving (3) Real-time Emotion Analysis, (4a) the development of Factorization Machine with a Learning-to-Rank objective function and (4b) a Two-Tower Neural Network with Multi-Task optimization, both suitable for utilizing Hybrid Features.

3.2 Data Collection and Data Fusion

There exists various well known and widely used datasets for training, testing and benchmarking recommendation algorithms in the domain of music, such as the Million Song Dataset and Last.fm Dataset [51] but no such dataset exists applicable to Emotion-Aware tasks. All other research works, discussed in the "Related Work" section, had to collect their own datasets, which were not made publicly available. Similarly, in order to fulfill the outlined objectives of the present research work, it is necessary to collect a feature-rich dataset that meets the following criteria:

1. Real-World interactions between Musical Items and numerous Users
2. An adequate amount of User-Item interactions
3. Access to the Implicit Emotional state of Users in relation to Musical Items
4. Access to the Emotional Responses of a Cross-Platform Audience
5. Rich and Multi-Level Item-Side Metadata

Regarding the first two criteria, training and evaluating recommendation systems generally require a significant amount of real user-item interactions since a small dataset would either result into an extremely sparse User*Item matrix or the model would most likely be overfitting on simplistic and unrealistic relationships and patterns. Furthermore, the third criterion is a prerequisite for developing an Emotion-Aware system, since an EA-RS requires the current emotional states or reactions of its users, at least on sufficient portion of the total interactions. Finally, criteria 4 and 5 are required for evaluating RQ-2. Since no individual source could satisfy all five criteria, a Data Fusion pipeline was employed, able to collect and integrate data from multiple sources including "Album of the Year", AZLyrics, Spotify, Twitter and YouTube APIs.

First of all, the website **"Album of the Year"** (AOTY) was selected as the central source of users and music items. AOTY is a review aggregator website, centered exclusively around music albums. Reviews of newly released albums are collected from various well-known and credible music reviewing sources such as music magazines (e.g Rolling Stone), magazines with art-related sections (e.g The Guardian), music databases (e.g All-Music) and famous, verified, independent reviewers (e.g The Needle Drop). Additionally, there is an active community of users rating, reviewing and discussing newly released albums. Both the users' and critics' ratings are aggregated independently, forming two sorted lists of the what each group considers to be the best albums released each year. Collecting data from AOTY, can adequately satisfy criteria 1, 2 and 3. Firstly, it consists of real-world interactions between users and musical items - albums in particular. Secondly, it has an active community of numerous users of which a significant portion expresses their opinions, reactions and feelings in the form of written reviews. Hence, the reviews can be analysed with Sentiment and Emotion Analysis techniques in order to extract the emotional responses and sentiment of the users in relation to the items. In terms of item-side metadata, AOTY offers the album's and its artist's names, the release date and the musical genres that the editorial team selects to categorise the album. Furthermore, we have access to the average critic and user scores as well as the number of reviews per album, features that may be useful in determining the relative popularity of the item. These metadata may be proven useful for the recommendation algorithm but can not fully satisfy the fourth requirement (Multi-Level Profiling) since no information regarding the lyrics, the musical content or the Artist's personality or biography is provided.

Furthermore, for each individual album collected from AOTY, **Spotify's API** was selected for accumulating the pre-computed acoustic mood, in terms of valence and arousal, of each track of the album. The lyrics of each track, if existed, were collected from a **Custom Google Engine** connected to various lyrics databases such as AZ-Lyrics, musixmatch and others. The lyrics were then analysed by an emotion analysis model, which will be discussed in detail in the next section. Moreover, for accessing the emotional reactions of a cross-platform audience, **YouTube's API** was used for gathering and emotionally analysing the ten "most relevant" user comments in relation to each music track. From YouTube's API the amount of views for each track-video is also kept as an indicator of its general popularity. All "track-level" features where then aggregated and averaged onto the "album-level". And lastly, the album's artist twitter profile was collected by using **Twitter's API**, if it existed, and a personality detection model predicted the average personality traits of each artist as expressed in their tweets. Combining all previously mentioned data and their extracted features, we have acquired access to multitudinous real user-albums interactions (Criteria 1 & 2) of which a significant segment is user reviews and by extension we have access to their emotional content (Criterion 3), the emotional reaction of a cross-platform audience (Criterion 4) and a variety of item-side features including musical genres, acoustic mood, lyrics emotions, "in-platform" and "general" album popularity and the artists' personality traits (Criterion 5).

Regarding the practical implementation of the described data collection process, a pipeline was developed centered around web scraping and parsing AOTY with nested requests to the other APIs for each track of every album. The Critic-based sorted list was selected as the source-list, with items of at least five critic reviews - the lowest possible value - so as to ensure that even less popular items would be present in the final collected item list. The workflow of the data collection process can be seen in Figure 3.5.

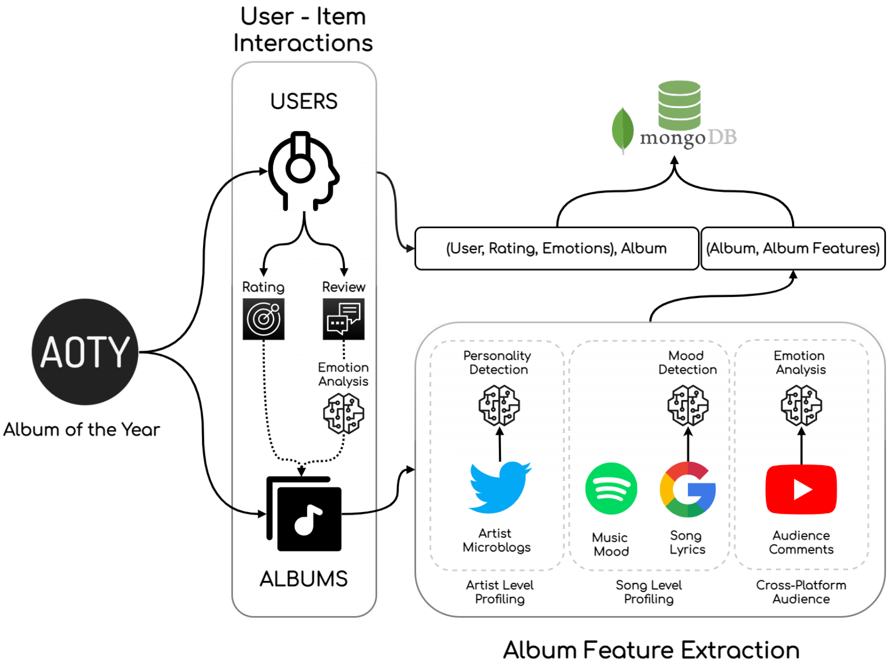


FIGURE 3.3: Data Fusion Pipeline

The process begins by web scraping Albums from **'Album of the Year'** and the users that have interacted with each Album, either by rating or reviewing it. For each review, the textual content is analysed by an emotion analysis model (described in detail in the next section) and each (User, Album, Emotions) interaction is stored in the MongoDB database. If the interaction does not contain a review from which to extract the user's emotions, all emotional values are set to zero. Furthermore, for each accumulated Album, the pipeline begins the process of collecting various album-side features. **Spotify API** returns the valence and arousal scores for each track in the album and then the scores are averaged for the whole album. From the **Custom Google Engine**, the lyrics for each track are scraped and analysed with a Mood Detection model specifically trained for analysing musical lyrics. From **Twitter API**, the 200 most recent tweets of the artist are collected and analysed by a Personality Detection model. The scores for the Big 5 or O.C.E.A.N personality traits are estimated for each artist, as well as their levels of expressed anxiety and avoidance. Additionally, from Twitter API, the total number of followers and tweets are also collected for each artist, representing their online popularity and activity. From **YouTube API**, the video for each track of the album is queried and the 100 'most relevant' user comments are collected, analysed with the same emotion analysis model and then averaged out at the album level.

For each collected Album, the extracted music mood from Spotify, the artists' personality traits from analysed tweets, the extracted user emotions from Youtube comments, the album's genres from AOTY and the various popularity metrics are all aggregated into a python dictionary and then saved in the MongoDB. When the automatic data collection process is finished, the data could be used for training and evaluating various recommendation algorithms. But before describing the selected recommendation models, it is important to describe how the emotion analysis and mood detection models, used in the data fusion pipeline, were developed and employed.

3.3 Mood Detection and Emotion Analysis

As previously noted, an Emotion Analysis model is required for analysing user reviews in order to create an emotion-aware recommendation system. Additionally, a Mood Detection model is necessary specifically for analysing lyrics, in order to partially examine the validity of the idea of Multi-Level Profiling. For both tasks, a framework similar to the proposed workflow of [52] will be utilized. Fundamentally, Chatzakou et al., proposes a hybrid emotion analysis methodology that combines machine learning and lexicon-based approaches. Initially, an emotion lexicon is applied on the text of an already emotionally annotated corpus. Thereafter, both the textual information and the extracted emotional values from the lexicon are given as input features in a machine learning algorithm. This framework necessitates the selection of 1) an emotional lexicon 2) an emotionally annotated corpus appropriate for the target task 3) a method for representing textual information and 4) a classification machine learning algorithm. In this section, the used datasets, the necessary pre-processing steps, the selected machine learning algorithms and their results will be presented and discussed.

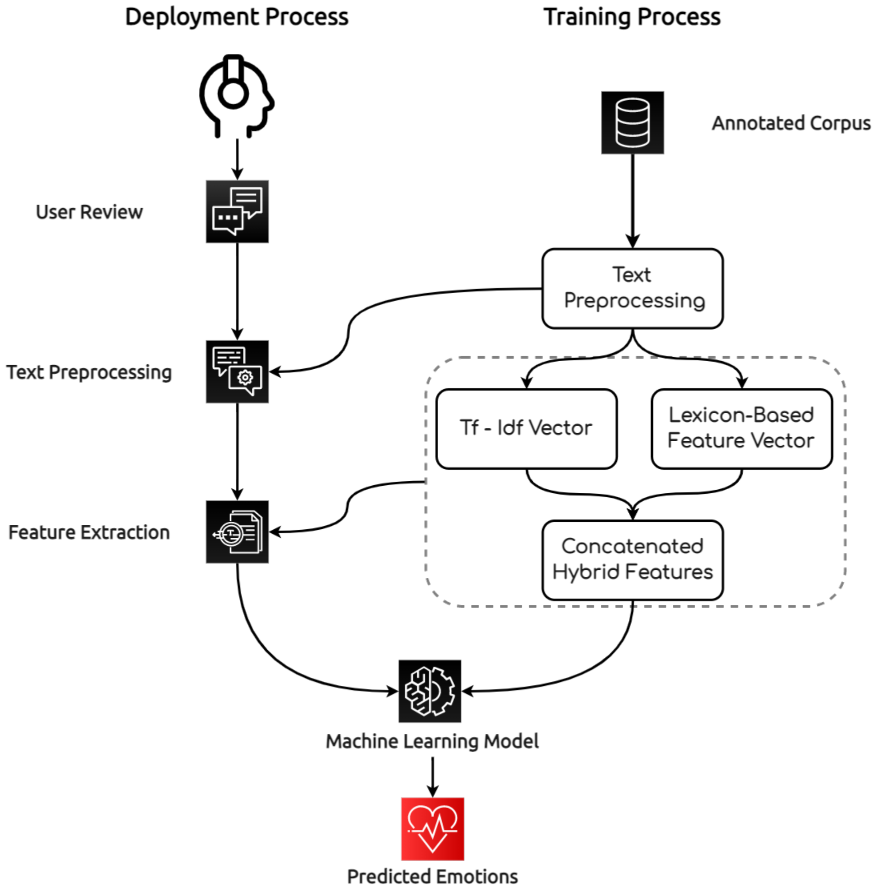


Figure 3.4: The pipeline of the hybrid emotion analysis approach utilizing lexicon and textual features as input in supervised machine learning models.

3.3.1 Emotion Datasets and Preparation

Concerning the selection of appropriate datasets for each task, the dataset form [43] (discussed in the related work section) was selected for the task of **Mood Detection on Lyrics**. The dataset consists of 2595 lyrics manually annotated on the four quadrants of the Circumplex model translating into Q1: Happy (high Valence and high Arousal), Q2: Angry (low Valence and high arousal), Q3: Sad (low valence and low arousal), Q4: Relaxed (high valence and low arousal). The publicly available dataset contains the names of the song, the name of the artist and the annotated mood value. Thus, the lyrics had to be collected anew by using the same workflow described in the previous section for collecting musical lyrics. From the total songs in the source dataset, after filtering out non-English lyrics, the collected lyrics amount to 1279 songs of which 298 are labeled as ‘relaxed’, 407 as ‘happy’, 244 as ’sad’ and 330 as ‘angry’.

For the task of **Emotion Analysis on Reviews**, no publicly available dataset with emotionally annotated labels, specialising in user reviews, was found. Most were labeled in terms of sentiment values. Consequently, the second most appropriate choice, was a Kaggle dataset consisting of 20.000 sentences of 'general emotional situations' of which 6761 were annotated as 'joy', 5797 as 'sadness', 2709 as 'anger', 2373 as 'fear', 1641 as 'love' and 719 as 'surprise'. The dataset was scored as 10/10 in terms of usability and was created by a frequent Kaggle contributor with a 'dataset expert' status collected by following the proposed methodology of [53]. This particular dataset was selected because it encompasses a diverse array of general emotional situations that could potentially better capture the multifarious emotions expressed in user reviews.

Subsequently, it was necessary to find a suitable emotion lexicon for each task. **NRC Word-Emotion Association Lexicon**, or EmoLex, is a widely used emotion lexicon containing more than 14.000 annotated in terms of both sentiments (negative, positive) and emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) [54]. On the other hand, since the MoodyLyrics dataset is labelled in terms mood - valence and arousal transformed into four categorical values - the most appropriate available emotion lexicon is the **NRC Valence, Arousal and Dominance** (VAD) lexicon [55]. VAD is a lexicon dataset for mood detection which includes more than 20.000 English words and their valence, arousal and dominance values in a scale of 0 to 1.

For both tasks, the raw text is passed through a **Pre-Processing** function that first identifies the part-of-speech tags (adjective, verb, noun or adverb) of a sentence and then lemmatizes each word with the use of WordNet Lemmatizer from NTLK. In linguistics and natural language processing, lemmatisation is the process of grouping different forms of a word into a single item, or lemma. For example, the lemma of the words 'working', 'worker', 'worked' is 'work'. This process increases the chances that a lemmatized word will be matched in the emotion lexicon in comparison to its various derivatives and inflections and thus decrease the chance of false negatives. Afterwards, each word in the cleaned input text, is searched in the emotion lexicon. If a word is found in the lexicon, the designated scores for each emotion is added in a different variable, one for each emotion. When the searching process is finished, each emotional variable is divided by the total amount of identified emotional words in the input text providing the average score for each emotion. This process was applied for the totality of both datasets and both were saved anew with their original textual information plus their extracted emotion and mood scores, normalized into a scale between 0 and 1 with the use of sklearn's MinMaxScaler.

After collecting both datasets and extracting their emotion and mood scores respectively, the text of the datasets needs to be preprocessed, cleaned and represented in a form suitable for classification machine learning algorithms. To this purpose, the texts are tokenized, lower-cased, existing punctuation marks and stop words are removed and then the remaining words are stemmed with the use of Porter Stammer by NLTK. Stemming is a process similar to Lemmatization, where the inflections and derivatives of a word are grouped into a single item representing its root form, not by identifying its lemma but by applying heuristic rules. These rules include among others the removal of 'ing', 'ed' and 'ly' suffixes, the conversion of 'are', 'am' and 'is' into 'be'. Furthermore, duplicate entries where removed and the annotated labels of moods and emotions were transformed from text to numbers with the use of LabelEncoder from sklearn.

Furthermore, the textual data were transformed into a 'term frequency - inverse document frequency' form or **Tf-Idf**. Tf-Idf is a statistical method for representing textual data into numerical that calculates the weighted importance of each uni-gram or n-grams in a collection of texts. The first part, the TF statistic, calculates the raw frequency of a n-gram, meaning how many times a word or a combination of words appears in a document D. The second part, the IDF statistic, calculates the logarithmic scaled inverse division of all documents that contain each n-gram. This relationship translates into the importance and the amount of information that a n-gram contains. If specific words, like the stop words 'and' and 'the', appear very frequently across multiple documents their proportional importance is rather low. Inversely, words that occur frequently in specific document but rarely on the whole corpus are considered more important and are weighted higher. Tf-Idf is a relatively simple method but is widely used on numerous information retrieval tasks such as search engines and text-based recommendation systems [56].

Finally, due to the fact that the phenomenon of **Class Imbalance** was present in both datasets - an issue that tends to create problems in machine learning models [57] - it was deemed necessary to experiment with re-sampling techniques and select algorithms that are receive correction parameters that mitigate the effects of an imbalanced dataset. Methods for mitigating the effects of imbalanced datasets can be broadly categorised into 1) re-sampling techniques and 2) class-weight aware algorithms. The first category can be further categorised into under-sampling and over-sampling. Under-sampling is the process of discarding samples of the majority class until the instances from the majority class reaches down to the level of the minority class. This can be done randomly or with specialised methods such as Tomek Links. On the contrary, over-sampling is the process of adding more samples into the minority class so as to reach closer to the majority class. This also can be done by randomly copying and slightly altering existing samples or by specialised methods that create synthetic instances such as SMOTE [58]. The second approach for dealing with imbalanced data is with specific machine learning models that can adjust their internal loss functions relative to each class imbalance ratio. These models are more aware about the imbalance existing in the data and are fine-tuned accordingly. Models with the ability to be adjusted by a class weight parameter include Logistic Regression, Decision Trees, SVM and Neural Networks among others.

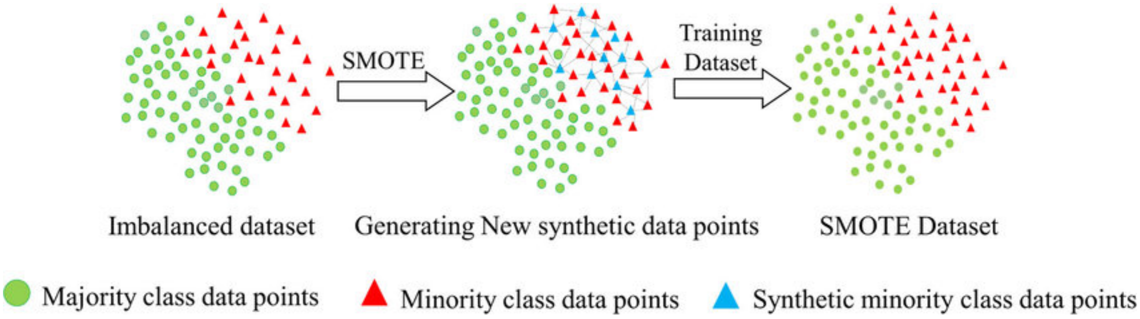

SMOTE
Training Dataset
Imbalanced dataset
Generating New synthetic data points
SMOTE Dataset
Majority class data points
Minority class data points
Synthetic minority class data points

Figure 3.5: SMOTE : Synthetic Minority Oversampling Technique

In the context of this work, since both the Emotion and Mood dataset were of relative small scale, it was deemed more appropriate to experiment with **Over-Sampling** techniques and **Class-Weight Aware** algorithms, so as to minimise the loss of information resulting from under-sampling. More specifically, the 'Synthetic Minority Oversampling Technique' or SMOTE was selected as the over-sampling method. SMOTE works by selecting instances of the minority class, identifying its nearest neighbours and creating a new synthetic instance by combining aspects of two neighbouring instances [58]. After **splitting** the data on a 80 to 20 ratio, the minority classes were over-sampled with the use of SMOTE resulting in 325 songs per class for the MoodyLyrics dataset and 5408 instances from each class for the Emotional Corpus dataset.

Ultimately, in this section, the selection of two textual datasets one annotated on Mood and the other on Emotional values was discussed, as well as the necessarily cleaning, pre-processing, feature representation and balancing between unequal classes that were required. After that, the process of training and evaluating machine learning algorithms was made possible.

3.3.2 Training and Evaluation of Mood and Emotion Models

Provided that the labeled datasets are collected and cleaned, the emotion scores are extracted from emotion lexicons and the textual information are vectorized, the process could proceed to training and evaluating machine learning algorithms for emotion analysis with hybrid features. As was discussed in the previous section, the prerequisites for the selected algorithms were that they should be able to 1) handle multi-class classification 2) accept a class weight parameter for adjusting imbalance datasets and 3) be able to work with sparse textual data of relatively low scale. Thus, suitable algorithms include

Logistic Regression and variations of Support Vector Machines. Neural Networks are also very effective models for multi-class text classification however they tend to require larger scales of datasets. Hence, for the purposes of this work Logistic Regression (LR) and Linear Support Vector Machines (L-SVM) were selected as well as the Multinomial Naive Bayes (MNB) but only for the already balanced dataset from re-sampling since it can not receive a class weight parameter.

For discovering the best performing model and for the fair evaluation of the selected machine learning algorithms, the method of **BayesCV K-Fold Cross Validation** was selected. This is a technique useful for 'tuning' the hyper-parameters of a model, meaning finding the best performing combination of parameters, by training a surrogate model that is monitoring the parameter-space and selects combinations that tend to perform better together and on contrarily avoid bad-performing ones. This method has a significant advantage over the traditional GridSearchCV which is exhaustively testing all possible combinations of multiple values of various hyper-parameters to arrive at the best one. Additionally, BayesCV also performs Stratified Cross Validation. The training set is divided and trained into K-folds (K=3 was selected) where the percentage of each class is approximately maintained, and then the the model is trained with a given combination of parameters on one the training-fold and evaluated on the testing-fold. The same process is repeated K times for each hyper-parameter combination. At the end of this process, the BayesCV retrieves the best performing combination of parameters for that model based on its cross validated performance. Which model is considered the 'best performing' one is decided based on the researcher's defined evaluation metric or metrics. Thereafter the best performing version of each model is then evaluated on our initial testing set that so far no model has encountered.

Among the three selected model, SVM - based on Stochastic Gradient Descent - has the most hyper-parameters that require tuning. Those include five different 'loss' functions, a 'penalty' for the regularisation term, 'alpha' that multiplies the regularisation term, the 'learning rate scheduler', the 'max number of iterations' and the 'class weight' among others. After tuning the algorithm, the best performing combination proved to be the 'hinge' loss (the loss used for linear SVM), a relatively low 'alpha' of 0.0001, 'penalty' of l2 regularisation and a 'max iteration' of 1000 steps. Secondly, Logistic Regression also requires the definition of a regularisation term 'C', as well as a 'solver' which is the algorithm used of optimisation and a finally, a 'penalty' which is the norm of penalisation. The BayesCV search discovered that 'l2' penalty, 'lbfgs' solver and C equal to 1 was optimal. Both for the Emotion and Mood textual datasets the same parameters were found to perform best, probably due to the similar nature and size of both. The only significant difference is that the mood detection is performed on the whole lyrics text while the emotion analysis model applied on each sentence of a review and then an average score is calculated. Thirdly, the Multinomial Naive Bayes, is a simpler model that only requires the tuning of parameter alpha which defines the parameter tuning. The maximum possible value of 1 was found to be optimal. Apart from the internal parameters of each model, tests were made for setting the class weight to be proportional to the imbalance ratio between the classes, so as to mitigate the issue of imbalance, or with using SMOTE for oversampling the minority classes. Finally, the ngram range parameter of the 'Tf-Idf feature extractor' was set for different values for selecting between uni-grams, bi-grams or both at the same time.

Beyond the technical tuning of the algorithms, it was deemed necessary to examine the performance of the machine learning algorithms when different features and feature combinations were provided as input. As was described in the previous section the objective is to build an emotion analysis model with architecture architecture, one that utilizes both machine learning and emotional lexicons. Henceforth, for both Mood Detection and Emotion Analysis tasks, **three different feature combinations** were tested:

1. The emotional scores extracted from the two NRC lexicons
2. The vectorized text with the use of Tf-Idf n-grams
3. The concatenation of both the extracted emotional scores and the vectorized text

This examination would prove if the hybrid emotion analysis architecture approach was the best choice or if a simpler method would suffice.

For the **evaluation** of the classification algorithms, both datasets were split into train and test sets to a fraction of 80% and 20% of the total dataset respectively. In this way the algorithm is trained and cross validated on the training set and at the end, the previously unseen testing set is used for evaluating the model's predictive performance. The process of K-Fold Cross-Validation aside for providing a method of tuning the multiple hyper-parameters of the model, it is also instrumental in order to avoid over-fitting of the model and to more thoroughly evaluate the model's performance. As the metric of evaluation, Accuracy, Precision, Recall and F1 score were selected. Each metrics denotes a different relationship between correctly classified items (True Positives, True Negatives) and wrongly classified items (False Positives, False Negatives). Using all four classification metrics offers an overall perspective of the models performance and behavior.

3.3.3 Results of Mood and Emotion Analysis Models

For the **Mood Detection** model aimed for the classification of Songs' Lyrical content, in terms of Valence and Arousal, after tuning the multiple combinations of hyperparameters, it was identified that the Tf-Idf vector based on uni-grams had the better performance instead of bi-grams or uni-grams and bi-grams together. Furthermore, when SMOTE oversampling was applied, it improved the predictive accuracy of Linear SVM (L-SVM) and Multionomial Naive Bayes (M-NB) while Logistic Regression (LogReg) was unaffected by the (relatively small) imbalance of the dataset. The results are presented only evaluated on the F1 score due to the fact that after balancing the dataset, all four metrics were practically identical. The best possible results that were able to be reached by the three algorithms on the validation phase, are presented in the following table for the previously unseen test set.

	NRC	**TF-IDF**	**NRC+TF-IDF**
L-SVM	0.686	0.840	0.907
LogReg	0.756	0.845	**0.912**
M-NB	0.630	0.799	0.777

Table 3.1: Results from the Mood Detection on Lyrics Models, with different Feature Combinations, in terms of F1 Score.

Logistic Regression was able to consistently perform comparatively better for all three feature combinations against the other two models. Both Linear-SVM and Logistic Regression perform better when both mood and textual features are given, while Multinomial Naive Bayes gives is better when only textual information were given. However, the overall best result was given by **Logistic Regression**, with a score of 91.2%, working with **Hybrid Features**, the concatenated NRC (Valence, Arousal, Dominance) scores and the Tf-Idf vector as input.

For the task of building an **Emotion Analysis** model for user reviews, tuning the algorithms and examining various n-gram lengths and different imbalance handling techniques it was discovered that a combination of both uni-grams and bi-grams worked better for the Tf-Idf vector while setting the 'class weight' parameter to 'balanced' worked better for L-SVM and LogRec while for MNB were necessarily over-sampled with the use of SMOTE. The parameter and techniques combinations that worked best for each algorithm during the cross validation phase were then evaluated on the previously unseen test set. The results in terms of F1 score are presented in the following table.

	NRC	TF-IDF	NRC+TF-IDF
L-SVM	0.335	**0.851**	0.847
LogReg	0.337	0.837	0.843
M-NB	0.320	0.798	0.790

Table 3.2: Results of Emotion Analysis Models with different Feature Combinations, in terms of F1 Score.

For the objective of building an Emotion Analysis machine learning model, the best performing in terms of overall F1 score was **Linear SVM** but when utilizing only the **TF-IDF of both uni-grams and bi-grams** as an input feature and with the class weight parameter set to 'balanced'. In this task, the emotional scores extracted from NRC EmoLex were not able to further contribute to the model's ability to predict the emotional classes. This event validates the choice of not 'shooting straight' at the model with the most complicated and hybrid architecture but also examining various simpler possibilities. Apart from the over-all F1 score, due to the fact that the dataset consisted of six classes, it is useful to also examine the per class F1 scores of each model's best possible performance.

	Anger	Fear	Joy	Love	Sadness	Surprise
L-SVM	**0.89**	**0.85**	**0.91**	**0.78**	**0.93**	0.75
LogReg	0.87	0.84	0.90	0.76	0.91	**0.78**
M-NB	0.85	0.80	0.88	0.72	0.89	0.66

Table 3.3: Per Class Evaluation of Emotion Analysis Models in terms of F1 Score.

Again, the Linear SVM model consistently performs better, in five out of six classes, the other two candidate models, it was thus selected as the model of choice for analysing user reviews as a part of the data fusion pipeline described in Section 3.2. However, since this model was going to be applied in user reviews accompanied by ratings and since the sentiment of the review is likely correlated with the chosen rating, it was also deemed important to maintain the information of Sentiment that was present in the NRC EmoLex by utilizing VADER. VADER is a pre-trained sentiment analysis model able to predict the degree of how positive, neutral or negative is the expressed sentiment of a text. Additionally, the merits of VADER lies in its various heuristics able to understand negations ('not good'), contractions ('wasn't very good'), degree modifiers (words that affect the intensity of expression), polarity shift (with words like 'but'), slang words, emoticons, acronyms and initialisms as well as the adjustment of sentiment with the use of punctuation or capitalised words. All these characteristics make VADER a valuable tool especially for analysing online activity and user reviews that usually consist of multiple of the above characteristics.

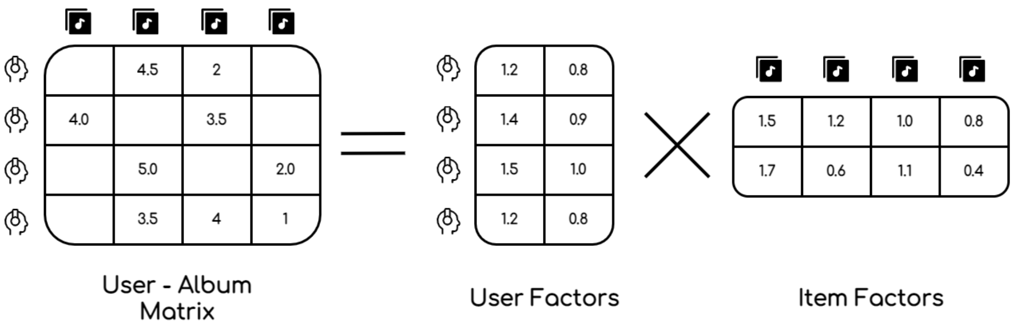
The final stage of this process, was the finalisation of the two models, one for Mood Detection on Lyrics based on Logistic Regression with TF-IDF uni-grams and NRC VAD mood scores as input, and secondly an Emotion Analysis Model for Reviews based on Linear SVM utilizing TF-IDF uni-grams and bi-grams. The models were exported in a ‘pickle’ form and then were integrated inside the data fusion pipeline described in 3.2. Thereafter, both models were integrated into the Data Fusion pipeline inside two appropriate functions that receive the text (lyrics or reviews) apply the necessary pre-processing actions (similar to how the data of each model were pre-processed for training). The model performs a prediction on the input text and the data are then stored into the database.

In this section the process was described of training, evaluating and deploying one Mood Detection model for song lyrics classified as a whole into Happy, Angry, Sad and Relaxed, and one model for Emotion Analysis of online user reviews where each sentence is analysed and the scores on six emotions are averaged for the whole review. Since the integration of both models into the Data Fusion pipeline was complete, the data collection process could begin and thereafter be used for the training and evaluation of the Hybrid Recommendation Algorithms. In the next two section two such models will be presented, namely Factorization Machines and Two-Tower Neural Networks.

3.4 Hybrid Model #1 : Factorization Machines

Factorization Machines (FMs) is a general predictor machine learning model applicable for regression, classification and ranking problems that maps real-valued input features into a factorized latent space. A significant advantage of FMs is their ability to work properly even with high dimensional sparse data while maintaining near linear complexity, making them ideal candidates for click through prediction and recommendation systems. As the name implies, there exists a close relationship with Matrix Factorization (MF), another widely used family of models in the field of RS.

Matrix Factorization (MF) models belong to Model-Based Collaborative Filtering approaches and essentially attempt to factorize the user-item interaction matrix into two or more representative low-rank matrices. This is achieved by identifying latent factors and hidden patterns in the user-item interaction matrix and representing these relationships in the form of low dimensional embeddings. An indicative example can be seen in figure 3.6. Instead of relying on calculating the ‘all users to all items’ similarities,



	4.5	2	
4.0		3.5	
	5.0		2.0
	3.5	4	1

1.2	0.8
1.4	0.9
1.5	1.0
1.2	0.8

1.5	1.2	1.0	0.8
1.7	0.6	1.1	0.4

Figure 3.6: Matrix Factorization : low dimensional latent factors extracted from the sparse User-Item interaction matrix

as Memory-Based Collaborative Filtering models, MF utilizes the factorized low-rank matrices to perform predictions on unobserved interactions and identify similar users or items. The latent factors between users and items are learned by calculating and minimizing the squared error between the predicted and the actual target variable, most commonly ratings. Thus the algorithm is optimized for minimizing rating prediction errors. One popular variation of Matrix Factorization was proposed and popularised by Simon Funk [20] during the Netflix Competition of 2006-2009, also includes user and items biases and regularization terms in order to avoid over-fitting. The regularized squared error is calculated by minimizing the following function:

$$\sum_{(u,i)\epsilon S} (r_{ui} - \hat{r}_{ui})^2 + \lambda(b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2)$$

where u are the users and i are the items in the S sample, r the observed rating, $\hat{r}$ the predicted rating, b are the biases and q the factors and λ the regularization term .The minimization is achieved by stochastic gradient descent between

$$b_u \leftarrow b_u + \gamma(e_u i - \lambda b_u), b_u \leftarrow b_i + \gamma(e_u i - \lambda b_u)$$

$$p_u \leftarrow p_u + \gamma(e_u i \cdot q_i - \lambda p_u), q_i \leftarrow q_i + \gamma(e_u i \cdot p_u - \lambda q_i)$$

where γ is the learning rate and $e_{ui} = r_{ui} - \hat{r}_{ui}$. Matrix Factorization models have been used extensively in a wide range of applications in the field of RS. They are very powerful models however, they come with two significant limitations. One is that they can work only with the user - item interaction matrix and can not utilize Content-Based auxiliary features or Demographic, Contextual or Emotional information. This is also a contributing factor in the model's difficulty to deal with new users and items. Secondly, it is difficult for MF to work with implicit interactions. In cases that ratings are missing and implicit information are available, such as duration of interaction or number of interactions, MF models tend to under-perform and implicit interactions are far more common in real-world applications. Furthermore, as discussed in section 2.1.3, in relation to Steck's work, models solely trained explicit information tend to underperform since 'ratings are not missing at random'. Various variations of MF models have been proposed able to work with implicit interactions by turning implicit information into a binary problem (observed interaction or non-existent) and adding a numeric variable expressing the confidence of preference [59]. After transforming the problem into binary these models are still optimized for error minimization in their predictive ability. This however have a significant loss in not expressing the polarity of the interaction, if the user liked or disliked the item, which translates in poor ranking performance [33].

This is the context that Factorization Machines were proposed in. There exists an association between MF and FMs, however, FMs could more accurately thought of as the intersection of Matrix Factorization and Support Vector Machines. In fact, when FMs were proposed in 2010 by S. Rendle, they were directly compared with both MFs and SVM [25]. While MF models are trained, optimised and used for a single purpose task on an all to all user to item interaction matrix, FMs represent user-item interactions in n-tuples between users, items and the target values - ratings or implicit information - as well as additional real valued features or one hot encoded vectors. The required input data formulation is practically akin to training with SVMs and other traditional machine learning models. FMs have been shown able to mimic and contest the performance of Matrix Factorization models such as FunkMF and SVD++ when only the interaction matrix is supplied [25]. However, their advantage lies in the ability to work with auxiliary features matrices if they are supplied. FM's function expresses the relationship for every pairwise feature interaction and combination.

$$f(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} (v_i \cdot v_j) x_i x_j$$

where w are the calculated weights of each feature and $(v_i \cdot v_j)$ is the dot product between two features represented onto the latent factor space embeddings. This formulation is very similar to Logistic Regression with the significant difference of calculating the inner product of the two features instead of calculating each term independently. By factorizing the relationship between each feature combination, FMs are able to significantly decrease amount of parameters that require calculation compared to Linear models and Support Vector Machines. If for example a small recommendation dataset consists of 100.000 users and 10.000 items, a polynomial regression model would require $User + Item + User \cdot Item$ would require the calculation of over 1 billion parameters. On the other hand, FMs of F = 10 factors would require the calculation of $Users + Items + Factors \cdot (Users + Items)$, approximately 1.2 million parameters, three orders of magnitude lower than linear models. This part of FMs explains how FM are able to solve the first problem of MF models - their lacking ability of utilizing auxiliary features - while at the same time being able to work with sparse high-dimensional data, in contrast to SVMs. This however does not solve the second issue, ranking preferences when dealing with implicit interactions. Two main concerns have been recently raised in relations to the limitation of FMs and their potential lack of expressiveness and their ability, or lack thereof, of scaling to hundreds of sparse and dense features [23]. This issue however, will be examined empirically in the context of EA-RS.

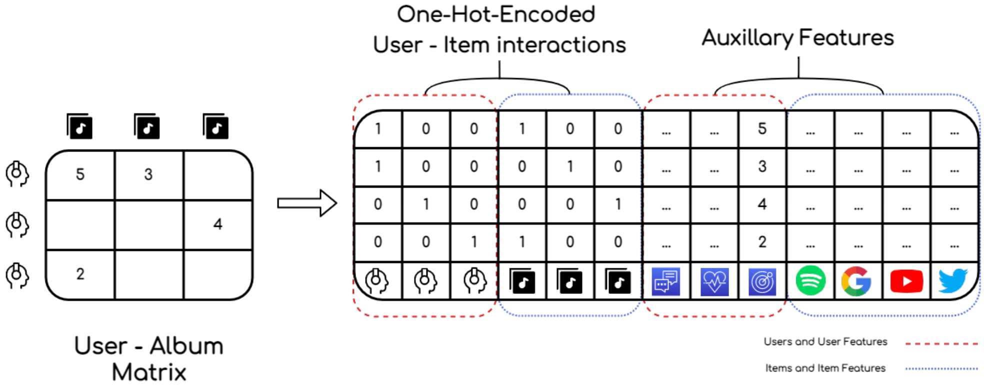


FIGURE 3.7: Data transformation from a sparse User-Item matrix to a one hot encoded formulation enabling the employment of auxiliary content-based features in FMs.

The second important innovation in the context of FMs is their ability to employ Learning-to-Rank (LTR) optimization techniques. LTR techniques are being used in State-of-the-Art information retrieval systems, search engines and recommendation systems [60]. In contrast to MF models that are optimized for minimizing prediction errors, FMs are trained on both observed interactions and unobserved interactions. Two widely used LTR methods in the field of Recommendation Systems are Bayesian Personalized Ranking (BPR) and Weighted Approximate Pairwise Rank (WARP). First, **Bayesian Personalized Ranking** is founded on the assumption that each 'observed' interaction between a user and an item is preferred over an 'unobserved' interaction and that all pairwise interactions are independent [61]. Then, the optimal ranking sequence is calculated for each user by the maximal posterior probability between n-tuples of (user, observed interaction, random unobserved interaction) given by the following function:

$$Max_\theta : ln[p(rank_u|\theta)p(\theta)]$$

where $rank_u|\theta$ is item ranking for user u as predicted by the model. Thereafter, the ranking based on the assumption that implicit feedback (observed interactions) (i) is preferred by the user over an unobserved interaction (j) can be expressed with the following function:

$$p(>_u|\theta) = \prod_{(u,i,j)\epsilon S} \sigma[f(u,i|\theta) - f(u,j|\theta)]$$

where σ is a sigmoid function for transforming items onto [0,1] range. $f(u,i|\theta)$ and $f(u,j|\theta)$ are calculated by using FMs function. By adding a regularisation term, the final function that the model attempts to maximize becomes :

$$Max_\theta \sum_{(u,i,j)\epsilon S} ln(\sigma[f(u,i|\theta) - f(u,j|\theta)]) - \lambda||\theta||^2$$

In the same vein, **Weighted Approximate Pairwise Rank** relies on the same fundamental assumption as BPR and relies on (user, observed item, unobserved interaction) triplets [62]. However WARP does not select unobserved interactions randomly. Instead, it is sampling multiple unobserved interactions for each user and selects items that the model has not yet learned to rank accurately. This process is comparable to methods of active learning where the most difficult cases are selected for examination. WARP results in a more focused gradient and is especially important in mitigating the popularity bias since it actively seeks to learn from 'difficult cases' such as cold and less popular items [63]. Actually, BRP could be considered as a special case of WARP, one where the maximum sampling limit is equal to one and the gradient is passed through a sigmoid function that normalises the results in the (0,1) range. WARP algorithm works as following :

1. Select a specific *(user, observed interaction)* tuple.
2. Sample an unobserved interaction at random.
3. Performs prediction for both items.
4. If the models ranks the unobserved interaction significantly higher (decided by a margin parameter), than the interacted item, the gradient should be updated so as to inverse the prediction's direction and rank higher the observed item over the unobserved one.
5. If the ranking is not violated, then the process is continued until a violation is found.
6. If the negative example was found during the first few samples, then perform a large gradient update.
7. If the negative example was found after examining multiple samples, then the model is close to convergence. Perform a constraint gradient update.

In summation, Factorization Machines, improve upon Matrix Factorization models by utilizing Learning-to-Rank optimization and by allowing for the use of auxiliary item- or user-based features. FMs are capable of working with high-dimensional and sparse data by factorizing features into latent feature space while maintaining at the same time the flexibility of SVMs and the ability to work with multiple types of features, within linear time complexity. FMs auxiliary features can either be static or dynamic, user-centric or item-centric. Their architecture make them an ideal candidate for working with both Context-Aware RS (e.g utilizing the times of day each interaction occurred) and Emotion-Aware RS that rely on dynamic user-centric features. Furthermore, by combining them with WARP they are a promising contributing factor for mitigating the Popularity Bias of Collaborative Filtering models, one of the purposes of this research work.

Many Factorization Machines modules have been developed and are readily available including fastFM, LibFM and XLearn. However, for the purposes of this work, LightFM, developed by Miciej Kula in 2016 was selected [17]. The reason was that LightFM offers a flexible API allowing for extensive experimentation with various datasets, features combinations. Furthermore, the architecture of LightFM allows for easily integrating both item-based features as well as dynamic user-based features such as emotion reactions, a crucial criteria for the purposes of this work. Finally, LightFM offers an implementation of both discussed Learning-to-Rank functions, BPR and WARP, promising better overall performance than optimizing for error minimization in rating prediction and a potential to mitigate the issue of popularity bias.

3.5 Hybrid Model #2 : Two Tower Neural Networks

Artificial Neural Networks (ANNs) is a family of models that in recent years have gained huge popularity in multiple research fields and applications, from Natural Language Processing, Speech Recognition, Computer Vision, Time Series Prediction and among others, Recommendation Systems. ANNs offer flexibility and powerful building blocks for undertaking challenging problems. Firstly in terms of flexibility, in the context of RS, ANNs can easily be adapted for working with either explicit or implicit user feedback or for both - with multi-task optimisation. Additionally, they can effortlessly be optimized either for retrieval, rating prediction or ranking tasks just by being optimized for multiple tasks at once and by slightly modifying and combining the loss functions and the optimization metrics. Furthermore, domains that require sequence-based recommendation, for example generating musical playlists, can be better handled by using ANN variations known as Long-short Term Memory Neural Networks. Secondly, ANNs can utilize rich auxiliary content-based and user-side information and - even more importantly for specific purposes - can integrate multi-media data such as images, videos, music and text, without the need for training separate models in different modules and then creating complicated integration pipelines. Thirdly, ANNs offer the capability for transfer learning methods by using pre-trained models on different datasets, of related tasks, for extracting useful features or for fine-tuning the model on the target dataset. ANNs had been used in content-based Recommendation Systems, however, since 2017, there has been a flourishing of Neural Network inspired by Collaborative architectures.

Neural Collaborative Filtering (NCF), was an attempt to consolidate ANNs' ability to work with complex auxiliary features with the powerful predictive power of Matrix Factorization models and their ability to identify rich latent factors [23]. The proposal of X. He et al substituted the inner dot product of Matrix Factorization with a Multi-Layer Perception (MLP) and the use of a Log Loss function in order to discover latent factors for both users and items. First, NCF receives two hot encoded vectors, one for the user and one for the items she has interacted with. The item vector consists of 1 for observed interactions and 0 for unobserved ones. The input vectors are then passed into the 'embedding layers' that represent users and items onto a dense latent space in the form of embeddings. Thereafter, the embeddings layers are fed into one or multiple MLP layers which attempts to identify hidden patterns inside the data. Finally, the output layers produces the predicted probability for each User-Item pair, which translates in the likelihood that a 'user U will choose to interact with item I'.

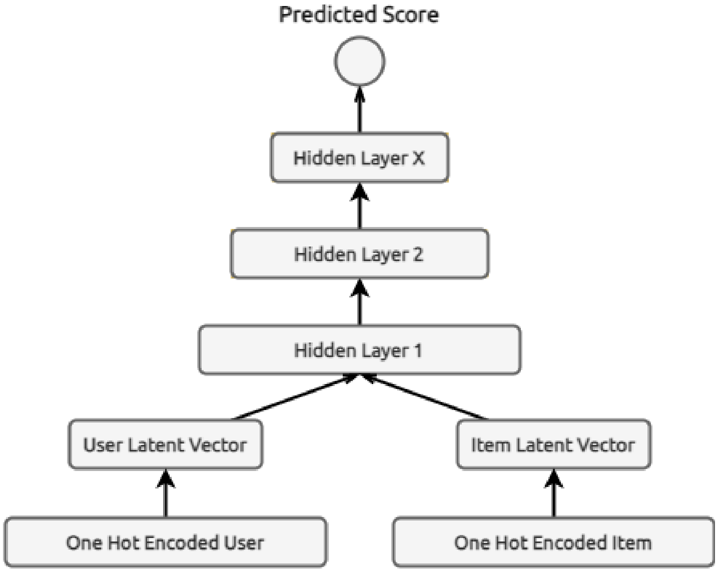


Figure 3.8: The architecture of a simple neural collaborative filtering model.

In order to deal with Implicit interactions and the lack of negative examples, the proposed Pointwise Log loss function utilizes the framework of rating prediction but of the task of classification in combination with negative sampling. The minimization of the mean square error between the predicted rating and the actual target value is being calculated for a specific user-item interaction but the result is passed through a sigmoid activation function turning the rating prediction into 0 or 1, a positive or negative example. Then the model attempts to correctly predict the observed interactions as positive and the randomly sampled unobserved interactions as negative. Apart from Pointwise loss, Pairwise and Adaptive Ranking Loss functions, as were described in Factorization Machines for WARP, can also be used with ANNs. In these cases, the model is trained on n-tuples of (user, observed interaction, random unobserved interaction) and tries to minimise the 'Loss = 1 - sigmoid (Positive Prediction - Negative Prediction)' where sigmoid S is

$$S = \frac{1}{1 + e^{-}x}$$

Another approach for training Neural Networks for Recommendation Systems with the logic of Collaborative models is **ANN Multi-Class Classification**. This approach treats recommendation as an extreme case of multi-class classification and using the softmax classifier in the output layer. In its general form, the softmax is expressed as :

$$\sigma(\overrightarrow{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

where z_i is a value taken from the $\overrightarrow{z}$ input vector that requires 'softmax' normalisation in order to be in accords with a proper probability distribution between K classes. Z_i is set as the power of e which always returns above zero values but it will be close to zero for negative values and continuously higher proportional to the input as it increases. The sum at the denominator is the 'normalisation term' that ensures that all z_i values will add up to one. In the context of RS, the use of softmax was initially proposed, by P. Covington et al from YouTube's research team [6] and was formulated as follows :

$$P(w_t = i|U, C) = \frac{e^{u_i u}}{\sum_{j \epsilon V} e^{u_j u}}$$

In this formulation, the altered softmax function attempts to predict the probability that a user U is going to choose to interact with item i (from the totality of the dataset), at a specific point in time w_t and a specific context C. In actuality, u is the sparse vector representation - or embeddings - of a user's 'query' created by the user's past interaction history which the ANN is trying to learn from. Then, the softmax function in order to classify the items in the corpus expressed as a probability between 0, 'user U is not going to interact with item i' and 1, that she will.

Naturally, if there are tens of thousands or even millions of items, classifying every single user towards every item in the corpus would be extremely time and computationally consuming. Additionally, apart from being infeasible and time consuming it would also be vacuous and inconsequential since, in all likelihood, most items would not interest most users and thus would be negatively rated since most users showcase specific inclinations, even if they are relatively diverse, towards certain types of items. Furthermore, on a more practical level, most recommendation systems present, at most, a few thousand items towards each user, therefore predicting the relationship between all-to-all users would not be useful. For these reasons, P. Covington et al, proposed a method of Negative In-Batch Sampling to work in combination with the Softmax function. To that purpose, the proposed approach samples a few hundreds or thousands negative items from the same mini-batch based on the assumption that 'since a user's preferences are learned, there no reason sampling more items'. Therefore the recommendation process is split into a Retrieval and a Ranking phase. In the first, a 'Candidate Generation' Neural Network receives the user's past interaction history and their demographic and contextual characteristics, it produces the dense vectors for the user and learns to retrieve a few thousands 'candidate' items, among millions, that the user will most likely be interested in. Second, the ranking model receives the embeddings of the candidate items as well as the items' features and ranks them in order to predict the items that will be of most interest to the user and be present to her in the most appropriate order. The final ordered list, consists of the Top K recommendations and is the result of the approximate nearest neighbor's calculated from the dot product between user and items embeddings. Despite its improved performance, the reliance on a fixed item vocabulary has proven limited in dealing with a dynamic item set and the continuously evolving nature of recommendation systems [64].

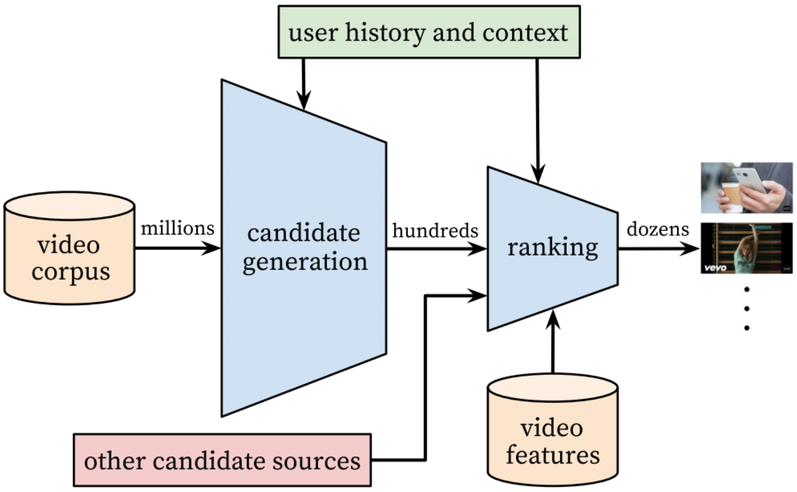


Figure 3.9: Architecture of YouTube's recommendation system (P.Covington, 2016)

Another similar neural-based architecture for RS, is known as **Two-Tower Neural Networks** (2T-NN). These models consist of two identical Multi-Layer Perceptron

networks, one responsible for representing user query embeddings, one for item embeddings and the output of the model is the inner product between the two embeddings $s(x, y) = \langle u(x, \theta), v(y, \theta) \rangle$ where θ is the learned map model parameter. An indicative model with a 2T-NN architecture can be seen in figure 3.10. This architecture is more commonly used in Natural Language Processing, for calculating text similarity, textual information retrieval, chat-bots and more generally problems with a 'query - candidate' nature but in the context of RS, they are adapted for working with high-dimensional sparse data. Variations of 2T-NNs have recently been deployed by YouTube, Google Play [27] and AirBnb [64]. One notable advantage of this method is its flexibility, since it can easily be adapted for Content- and Context-Aware tasks. The input for the user tower can be user ids, search queries, explicit ratings or implicit interactions, interaction length, contextual information such as the time of interaction or more relevant to this study, emotional reactions. Similarly, the item tower can receive a diverse array of meta-data, including raw multi-media data such as images, music and videos, all integrated into one single model, without the need for training separate models. with negative sampling from the mini-batches.
Additionally, 2T-NN can overcome the reliance on a fixed item vocabulary of multi-classification ANNs by working with hashing item dictionaries, which make it easier to work in a dynamic and continuously evolving environment, frequent in RS. This means that candidate items do not have to be classified and negatively sampled from the the totality of the corpus but rather can be selected from an efficiently produced, filtered subset that satisfies a required criteria. For example, the filtered items of the last N days, so as to ensure the freshness of the candidate items in a specific subset of the recommendation system.
Finally, an additional advantage of 2T-NNs is that item embeddings can be pre-computed at the training stage while the new user embeddings can be estimated at the serving stage, allowing for better capture the contextual situation of the user while maintaining efficiency. Again, 2T-NNs generally treat recommendations as 'extreme multi-class classification' utilizing the Softmax function

However, one significant limitation of models relying on in-batch negative sampling, is that training data tend to follow a power law distribution, most users have disproportionally interacted with a few most popular items, and sampling negative items from the training mini-batches will reflect this bias and continuate the popularity bias or alternatively will over-penalise popular items. This issue has been shown to be mitigated by '**Sampling Bias Correction**', a proposal that estimates item frequency in the mini-batches and adapts their selections accordingly. The correction factor is an adaptive parameter, learned through gradient descent in parallel with the optimisation of the main RS task. [Sampling-Bias-Corrected Neural Modeling for Large Corpus Item

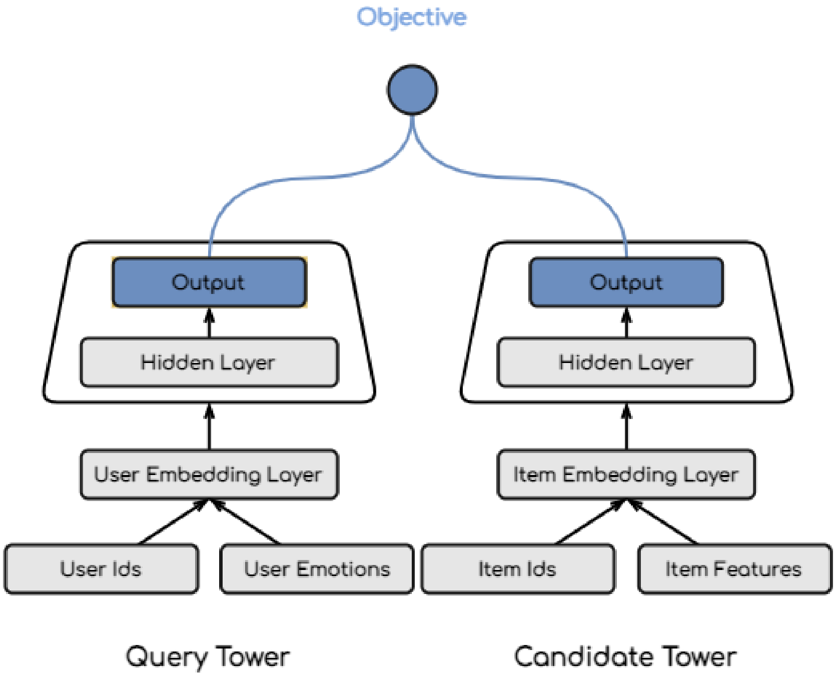


FIGURE 3.10: Architecture of a Two Tower Neural Network.

Recommendations] Even further, in-batch negative sampling, even more severely affects cold items - items that no user in the training set has interacted with - will by definition, never be selected from the training mini-batches, thus the model will not be able perform informed prediction on them. For dealing with this problem, **Mixed Negative Sampling** (MNS) has been proposed, where negative items are selected both from in-batch by taking account item frequencies and from sampling negative samples from the uniform corpus thus increasing the probability that a cold item will be selected as a negative candidate [26]. In MNS, a parameter selects the ratio between in-batch and uniform negative selections, giving the ability to select the optimal or desired balance between Accuracy and Diversity. In both approaches, the softmax will still attempt to maximize the affinity for known (user, item pairs) while minimizing the affinity between unobserved (query, candidate) pairs belonging to other queries in the batch.

For the purposes of this work, the 2T-NN architecture was selected due to its flexibility and the capacity to straightforwardly receive both user-side and item-side features, an important prerequisite for the needs of this study. Even further, the use of ‘sampling bias correction’ or ‘mixed’ in-batch sampling could be proved instrumental in mitigating the challenges of item cold start and popularity bias, the main issue addressed by this diploma. The developed model is divided into two towers, the first receiving the user name from individual user-item interactions and their emotional reactions in the form of (User, Emotions) tuples where Emotions consists of (Sentiment, Anger, Joy, Love, Sadness, Surprise). The second tower receives the Album Name and if selected, multiple item-side features, including Genre, Music Mood, Lyrics Mood, Artist Personality, Cross-Platform Audience Emotions and various Popularity Metrics. It is important to note that although the described model is trained on users' emotional reactions, meaning that the interaction precedes the emotional measurement, this architecture could easily incorporate the current emotional state of the user, where the emotional state precedes the selected interaction. Then, a pre-processing layer is responsible for *Normalising* the continuous features in the range of (0,1) of both User-side and Album-side values, *Vectorizing* textual features such as the Genre into uni-grams and for creating a *StringLookup* hashing dictionary for both users and items. The hashed user and item names are then given as input onto their respective embedding layers and finally the embeddings are concatenated with the corresponding normalised features.

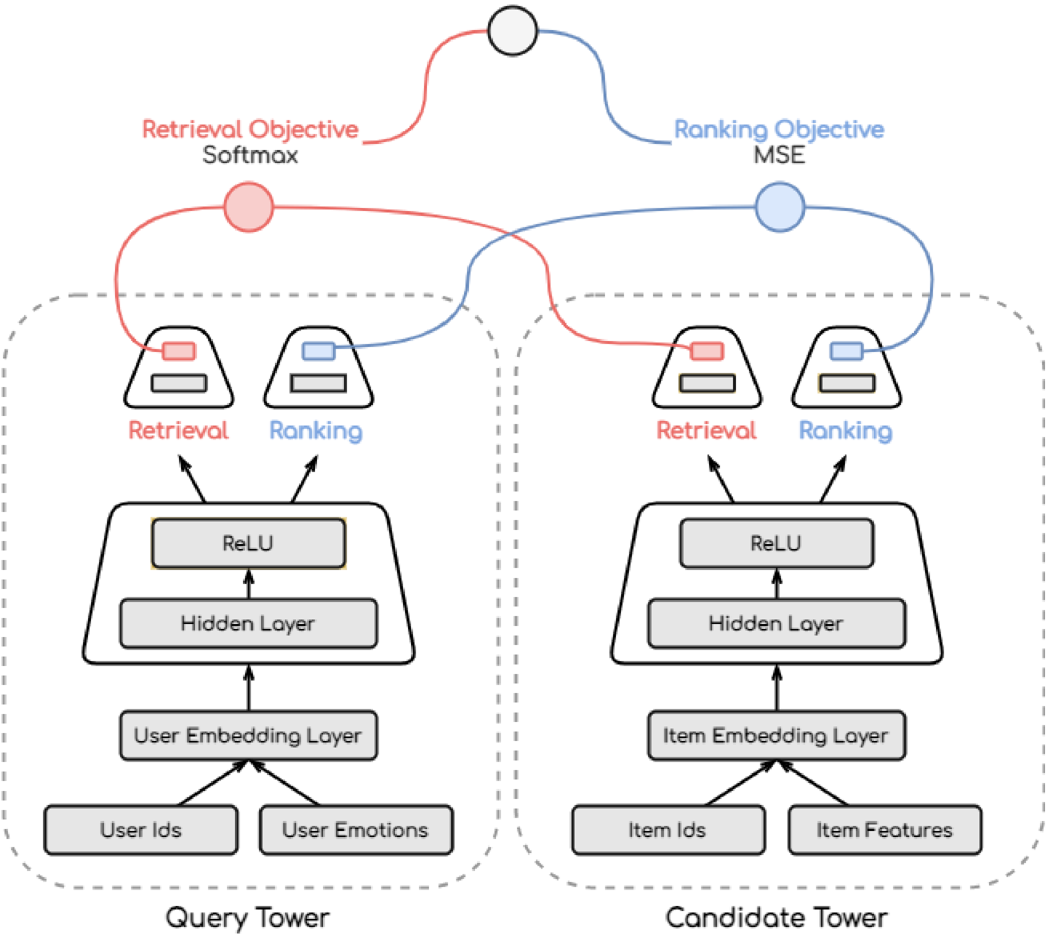


FIGURE 3.11: Architecture of a Multi-Tasking Two Tower Neural Network optimised for both ranking and retrieval.

The model is then trained based on Multi-task optimisation, a Retrieval objective and a Ranking objective, similarly to figure 3.11. For the Retrieval Task, the Softmax classifer was used with Categorical-Crossentropy as the loss function, appropriate for cases where the classification targets are in one-hot-encoded form. Moreover, the model is optimised for two objectives in parallel. The Retrieval Task attempts to minimise the categorical-cross-entropy error based on the Categorical Accuracy@K. Due to the fact that the retrieval model uses implicit interactions, sampling negative interactions was necessary in order to learn the positive and negative preferences of the users. To this end, a mini-batch of 128 candidates was randomly sampled from the uniform item dataset instead of the interaction tensor, that follows a power-law distribution. On the other hand, the Ranking Task attempts to minimise the loss in terms of Mean Squared Error between predicted and actual explicit rating score, based on the Root Mean Squared Error metric. The final output, the predicted items, is the inner product between the computed user and item embeddings.

Chapter 4

Experimental Design and Results

This Chapter begins by presenting the framework of the conducted experiments, whose purpose is to examine this study's central hypothesis. Section 4.1 is separated into three subsections discussing the experimental design, the dataset description and the selected evaluation metrics. Afterwards, the empirical results are presented and analysed in Section 4.2 which is similarly separated into three sections for quantitative 'Accuracy' metrics, qualitative 'Beyond Accuracy' metrics and finally, a comprehensive discussion of the ensuing insights.

4.1 Evaluation and Experimental Design

Examining the central hypothesis of this diploma and evaluating the proposed solutions, require a carefully planned experimental design. First, it must be proven that Collaborative Filtering models will indeed be negatively affected by Missing Emotional Values and by extension the increased Emotional Sparsity. Moreover, it needs to be shown that this situation further negatively impacts new and unpopular items, meaning that it contributes to the already present challenges of CF known as Cold Start and Popularity Bias. Thereafter, the proposed solutions to these challenges must be evaluated thoroughly and in ways that indicate the causal relationships between a proposed solution and the change in performative outcome. In this section, the whole process of experimentation as well as the methods and metrics of evaluation are discussed.

4.1.1 Experimental Design

In order to validate the study's central hypothesis a experiment design metaphorically resembling a *'case-control'* study is being performed, one were the emotional sparsity is the *'dependent variable'*. Initially, all models will have to be trained in two separate versions of the dataset. The first set, that will be referred to as the **'Ideal dataset'**, there will be no missing emotional values and hence no Emotional Sparsity. The models trained on this dataset can show how the models would behave in a ideal situation where no emotional sparsity was not present. This sampled set consists only of the U-I interaction containing written reviews from where the emotional reactions of the users' could be extracted. Based on central hypothesis and following the existing research on EA-RS, the performance of different models are expected to be improved in some way, most probably in terms of RMSE, Accuracy and Personalisation.

Thereafter, all models will be trained on a **'Real-World' dataset**, one were there are missing emotional values and by extension increased emotional sparsity. This dataset resembles the situation that emotion-aware models deployed in production will most probably have to face. Users write reviews for few of the items they interact with and it is hypothesised that most users will similarly not constantly wear their wearable devices nor will they continuously want to use face emotion recognition, so as to extract their emotional reactions. If the study's central hypothesis holds true, it is expected that pure Emotion Aware Collaborative Filtering models will face significant reductions in terms of Diversity, reinforcement of the Popularity Bias and Item Cold Start. It remains to be empirically examined how the two different types of hybrid models will behave under the increase in emotional sparsity. The detailed description of both datasets will be presented in the next section.

Model	**Parameter**	**Values**
Collaborative Filtering	Distance Measure	(MSD, Cosine)
	K Neighbors	(10, 50, 100)
	Type	(User-based, Item-based)
Matrix Factorization	Learning Rate	(0.0005, 0.001, 0.01, 0.1)
	Factors	(10, 20, 30)
	Epochs	(5, 10, 20)
Factorization Machines	Loss Function	(BRP, WARP)
	Epochs	(5, 10, 20)
	Components	(10, 30, 50)
	Learning Rate	(0.01, 0.05, 0.1, 0.15, 0.2, 0.3)
Two Tower Neural Networks	Embedding Size	(16,32, 64)
	Batch Size	(512, 1026, 2048)
	Layer Size	([16], [32], [32, 64], [32, 64, 128])
	Learning Rate	(0.01, 0.05, 0.1, 0.15, 0.2, 0.3)

Table 4.1: Hyper-parameters of each model

Apart from evaluating each model on the two dataset, in terms of conducting a fair comparative evaluation between heterogeneous models, careful **hyper-parameter tuning** is very important for getting the best possible performance out of each model. For this reason, all models are trained on an exhaustive Grid-search between multiple values for multiple parameters and then are evaluated on the Validation set in order to identify the best possible hyper-parameter combination. For each model, on every possible feature combination, on both dataset, the best possible hyper-parameter combinations are selected so as to finally evaluate them on the previously unseen Testing Set. The examined hyper-parameters for each model can be seen in figure 4.1. For the Factorization Machines were the loss function ('BPR or WARP), the number of epochs (5, 10, 20), learning rate (0.01, 0.05, 0.1, 0.15, 0.2, 0.3) and components (10, 30, 50). For Two-Tower Neural Networks, the learning rate (0.01, 0.05, 0.1, 0.15, 0.2), the embedding size (16, 32, 64), batch size (512, 1024, 2048) and the layer size ([16], [32], [32, 64], [32, 64, 128]) where each item in the list would be added as another hidden layer. The number of epochs was set to 30 and not tested further, since an Early Stopping callback was used, that if the model would not improve for 5 consecutive epochs in terms of validation accuracy, then the training process would be terminated and the model's weights with the highest validation accuracy would be retrieved. For distance-based collaborative models were the distance metrics (mean squared displacement, cosine similarity), the K nearest neighbors (10, 50, 100) and user- or item-based. Finally, for Matrix Factorization were the learning rate (0.0005, 0.001, 0.01, 0.1), factors (10, 20, 30) and epochs (5, 10, 20).

Model	**Feature Combinations**
Collaborative Filtering	Baseline
	User Emotions
Matrix Factorization	Baseline
	User Emotions
Factorization Machines	Baseline
	User Emotions
	Album Features
	Hybrid Features
Two-Tower Neural Network	Baseline
	User Emotions
	Album Features
	Hybrid Features

Table 4.2: Possible Feature Combinations for each model. Baseline relies on U-I Interactions, User Emotions relies on U-I Interactions and Emotions etc.

Finally, it is important to examine how each model performs with different **features combinations**. To this end, each model will have a baseline, a version solely relying on Explicit or Implicit Interactions without any Emotional Features so as to be compared with the model's performance when also using emotional information. Furthermore, in Hybrid algorithms, able to integrate Item-side features, there should be an 'item features' only test and one where both item features and users' emotions are present. This evaluation framework will be able to indicate how the algorithm works when emotional features are added and moreover, what item-side features may contribute to the model's performance. This means that the two hybrid models will have four different feature combination versions meaning 1) baseline : interactions only, 2) interactions + user emotions, 3) interactions + item features, 4) interactions + users' emotions + item features. All four version will exist for each dataset, resulting into eight different versions for each hybrid algorithm, as can be seen in figure 4.2.

4.1.2 Dataset Description

Regarding the collected data from the pipeline described in section 3.2, the user album interaction from Album of the year, between 2018 and 2020, were accumulated. This process amounted to 1997 unique albums and unique 12.981 users. After cleaning the data and dropping duplicate entries, the total amount of known interactions between users and albums were 369.770. From those, 63.298 were written reviews (including ratings) from which the users' emotions could be extracted, while 306.472 were ratings only. Although the known number of interactions are of the one hundred thousand scale, all possible interactions amount to 25.923.057, five orders of magnitudes larger as can also be seen in figure 4.1. This translates into 98.85% sparsity. Moreover, among the known interactions only 20.65% are 'emotional' meaning that we have access to the users' emotional reactions which means that the total **'emotional sparsity'** in the dataset is 99.7%.

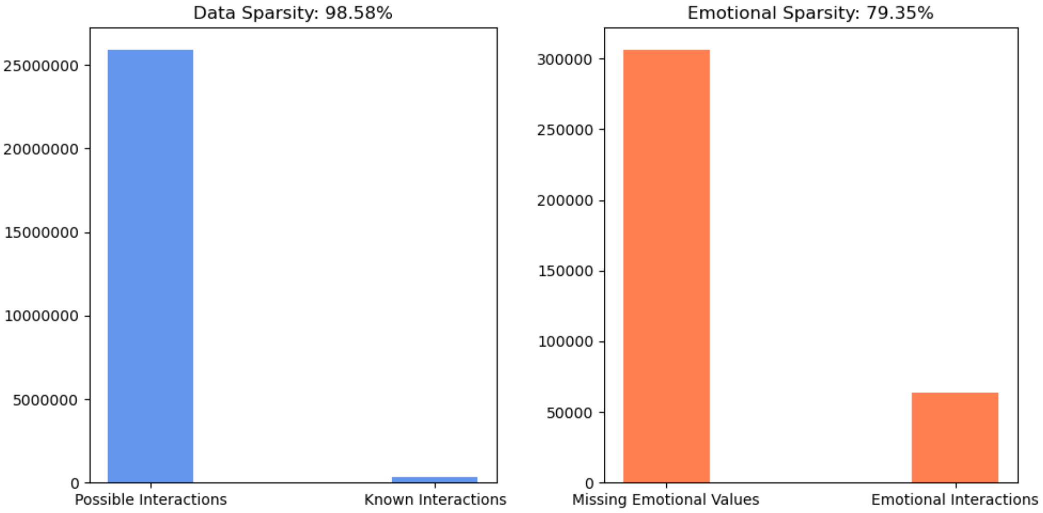


FIGURE 4.1: Total and Emotional Data Sparsity

Furthermore, 50% of the total interactions were made between the 148 most popular items. This indicates, that the interaction data follow a **Power Law distribution** as can be seen in figure 4.2. This is a very common phenomenon in recommendation-related tasks, known as the 'long tail of recommendation systems' and is also a contributing factor to the popularity bias. This is due to the fact that the majority of users have interacted with a very small minority of items, while most items have received very few interactions. Consequently, models relying on collaborative information will tend to reflect this bias [30].

Regarding the division of the dataset into three non-overlapping samples for training, validating and finally testing each model, a technique known as **stratified user-side splitting** was utilized. This method splits the dataset in such a way that that each user will appear at least N times in both training and validation set, but with different item interactions. With this approach, no accidental 'cold' users could appear only in the validation or only the test set. This was decided due to the fact that, one of the central issues of this study was the phenomenon of item cold start and it was deemed important to not also have to deal with accidental cold users, thus making it even harder for the model to perform informed predictions. For the user-side stratified splitting, completely cold users (with only one total interaction) were filtered out and then the three samples

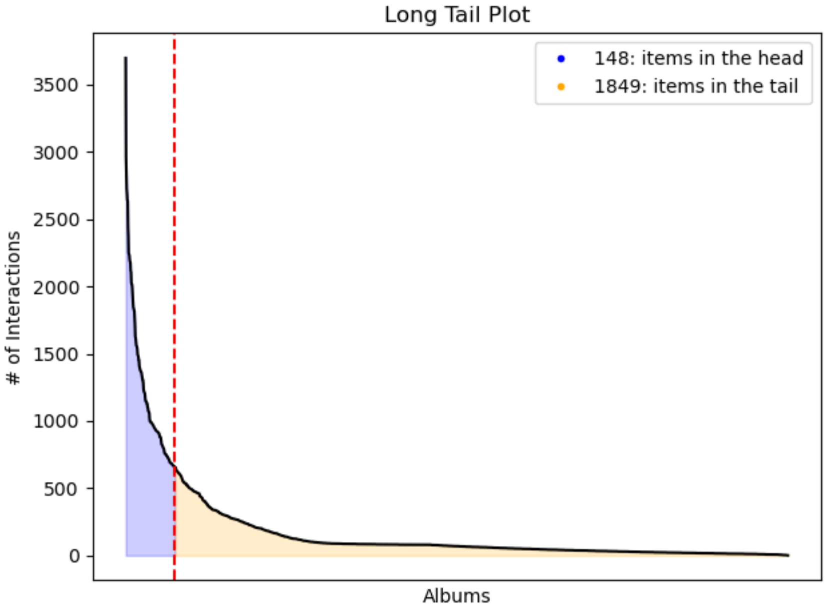


Figure 4.2: Power Law distribution of User-Item interactions

were spit by a 70 / 15 / 15 ratio into training, validation and testing set respectively. The same ratio was selected for both the 'Ideal' and the 'Real-World' datasets. The first, initially consisting of 63.298 emotional interactions was split into 46542, 8884, 7302 while the second consisting of all known interactions (369.770) were split into 265592, 54291, 45689 for train, validation and test sets respectively.

Dataset	Training	Validation	Testing
Ideal	46.542	8.884	7.302
Real-World	265.592	54.291	45.689

Table 4.3: Both datasets split with 75 / 15 / 15 ratio for training, validation, testing.

Furthermore, item with fewer than 2 interactions in the ideal dataset and 20 for the real-world dataset, were considered as Cold Item and all related interactions were filtered into a separate sample, used only for testing the issue of item cold start. The cold thresholds were not chosen arbitrarily but rather to ensure that at least one thousand interactions would be available for evaluating the issue of item cold start. Finally, in is important to note that the same 'random seeds' were used in each experiment for reasons of reproducibility and fair comparison.

4.1.3 Evaluation Metrics

Evaluation Metrics for recommendation systems, as was discussed in section 2.1.3, can be broadly divided into two categories, 'accuracy'- and 'beyond accuracy'- related metrics. First, the predictive accuracy of recommendation systems can be evaluated in terms of their rating prediction (applicable only when explicit ratings are present), retrieval and ranking abilities and performance. For all three categories, all users' interactions present in the testing set are iterated sequentially, with the target value being 'hidden' from the model. In **rating prediction**, each (user, item) tuple is given to the model and the predicted rating for each tuple is returned. Then, all predicted ratings are collected and compared with the actual explicit user rating. For calculating the error between the predicted and the actual ratings, a commonly used metric is the **Root Mean Squared Error** or RMSE which is given by the following formulation,

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\hat{r_n} - r_n)^2}{N}}$$

where N is the total amount of (user, item) pairs in the testing set, $\hat{r_n}$ the predicted rating and r_n the actual rating. RMSE shows the overall deviation between the predicted and the actual rating. The lower the RMSE values, the better the model's predictive ability in rating prediction. The calculation of RMSE is only relevant to model's that utilize the explicit ratings of the users' such as Matrix Factorization and the multi-task Two-Tower Neural Network, described in the previous section, but not for LightFM's Factorization Machines which are trained on implicit interactions.

For evaluating the **retrieval** accuracy of each model, every (user, item) pair in the testing set is iterated and the item that is considered the ground truth is hidden from the model. Then, the trained model is asked to perform K recommendations for the current user. All recommended lists are collected and compared with the ground truth. For this task, the **Top-K Categorical Accuracy** metric was selected which calculates the percentage of True Positives in the predicted list divided by the number of items in the ground truth list. However, for the purposes of this study, each interaction is evaluated separately, thus the N number of items in the ground truth list is always equal to one. This is due to the fact that emotion-aware recommendations are trained on (user, item, emotions) tuples and different emotional values may, supposedly, result in different recommended lists. Therefore, evaluating on the aggregated user's items would miss this valuable information.

For evaluating the performance of each model in 'Beyond Accuracy' issues, the metrics of Item Coverage, Novelty and Personalisation were selected. Firstly, **Item Coverage** is simply the totality of unique items I in the recommended item lists (from the retrieval evaluation phase) divided the total amount of items N in the dataset, formulated as $\frac{I}{N} * 100$ [35]. Despite its simplicity, it can display the degree of bias of a model towards certain items or the ability to perform diverse recommendations. On its own, a high Coverage score may be unimportant if the model's accuracy is very low. Indicatively, sampling items at random may result in a coverage score of 100% however the recommendations would be meaningless to the users.

Secondly, **Novelty** measures the ability of the recommendation system to suggest unconventional and less popular items. It was proposed by T. Zhou et all, 2010 and is uses the 'self-information' of each item in the recommended item list averaged across all recommended item lists in the testing phase.

$$Novelty = \frac{1}{|U|} \sum_{\forall u \epsilon U} \sum_{\forall i \epsilon K} \frac{\log_2 \frac{count(i)}{U}}{|K|}$$

where U is the absolute number of users, K the amount of items in each recommended list and count(i) is the amount of total interactions that each item i has received [34]. Novelty in combination with Item Coverage can measure the degree of Popularity Bias present in a recommendation model. If both metrics are relatively low it can be inferred that the Popularity Bias is present.

Thirdly, **Personalisation** is the degree of dissimilarity between the recommended item lists, calculated with the use of cosine similarity, indicating how 'generalist' or 'individualistic' are the recommendation of a model. Although higher scores translate into higher dissimilarity and hence better personalisation, on its own, this metric says little about the quality of the recommended item lists but it can be useful when paired with a decent score in terms of predictive accuracy. Moreover, while it is not the optimal metric, it can also be used to infer the extent to which Grey Sheep Users may be handled by the model.

Furthermore, while training and evaluating each model the **computational time** of each process is calculated in order to examine the relative scalability of each model in different scales of data. This is important for accessing the potential usage of such a model in production where efficiency and scalability are really important.

Finally, another significant question, is how to properly examine the issue of item cold start. A common approach (e.g Maciej Kula 2015) would train a model on two separate dataset samples, one where no cold items were present, neither in the training nor in the testing set and one where cold items would exist in both sets. The difference in terms of predictive accuracy for the two different datasets would indicate how well or badly each model reacts to cold items. In this study, since the dataset is already devided into 'Ideal' and 'Real-World', training all models on all possible feature combinations would tremendously increase the experimentation load. Furthermore, i believe this approach to be relatively redundant. Alternatively, it is suggested that a more efficient approach is filtering out the cold items from the training and testing sets and retaining them on a separate sample. After the main training and evaluation, it is requested for each model to perform K-item recommendations for the (user, item) interactions in the cold sample, with the ground truth item being hidden. This task can also be evaluated by using the categorical accuracy at K. Naturally, this is a very demanding test since expecting a model to recommend a previously unseen item in the Top K list and correctly predict this single interaction is very unlikely. For this reason, a relatively high K value should be used so as to 'give the chance' to the item to be recommended even if at low-ranking positions in the list. Still, a True Positive cold item would indicate that the model is able to recommend cold items to appropriate users, even if at small rates. Accordingly, the categorical accuracy at top 100 is being used for examining the issue of item cold start, and will be referred to as **'Cold-Accuracy@100'** for the remaining of the text.

Category	**Metric**	**Task**
Accuracy	RMSE	Rating Prediction
	Accuracy@K	Retrieval
Beyond Accuracy	Novelty	Popularity Bias
	Item Coverage	Popularity Bias
	Personalisation	User Satisfaction
	Cold-Accuracy@K	Cold Start
	Computational Time	Scalability

Table 4.4: The selected evaluation metrics and their evaluation objective.

4.2 Results

In this section the results are presented of the evaluation process for all models, in all possible features combinations, for both the 'Ideal' and 'Real-World' datasets in terms of both 'Accuracy' and 'Beyond Accuracy' metrics. Each model will have a **'Baseline'** where only the explicit or implicit U-I interactions were being utilized by each respective model. When the users' emotional reactions are added to the model, alongside the collaborative U-I interactions, this feature combination will be referred to as **'User**

Emotions'. Similarly, when the rich item-side features are added alongside the collaborative interactions, the feature combination will be referred to as **'Album Features'**. Finally, when all three are present (interactions, users' emotions and album features) the model will be characterised as **'Hybrid'**.

It is also important to note that in all following figures, lines of the same color represent the same model (e.g 2T.NN are always orange) and have a different point shape (squares for 2T.NN). Moreover, the use of continuous lines signify the use of the 'Ideal' dataset, where no emotional sparsity is present, while intermittent lines are for the 'Real-World' dataset, where a high level of emotional sparsity is present. Some values are missing in order to de-clutter the figures , but in those cases, a minor grid is added to estimate the missing value. Finally, the evaluation process will be presented by evaluation metrics and not by model, so as to more easily compare the performance of the various models.

4.2.1 Accuracy Results

Starting off with the measure of **RMSE**, only Funk's Matrix Factorization (fMF) and Two Tower Neural Networks (2T.NN) were utilizing explicit user ratings and hence could be compared on this task. Apart from fMF, other 'traditional' Collaborative Filtering models were trained and evaluated. Two distance-based models, one centered around item similarly and one on user similarity. However, their performance was significantly worse than fMF, and was thus emitted from this analysis.

Studying the continuous lines in figure 4.3 (the Ideal dataset) the *Baseline* of 2T.NN (0.148) slightly outperforms fMF (0.156). However, when *User Emotions* are employed, we notice an abrupt decline in predictive error for fMF, falling at 0.132, the lowest score in the Ideal Dataset. This phenomenon is in accords with the discussed related studies in EA-RS, who have shown that emotion-aware collaborative models can improve the model's predictive accuracy. In this version of fMF, only the sentiment scores were utilized, more emotions could not be integrated 'successfully' and 'productively' in its objective function. It can be inferred that users' sentiment in relation to items can indicate the polarity and intensity of their reaction and thus inform the rating prediction learning. On the contrary, the behavior of 2T.NN is more static, anchored between 0.145 and 0.148, a slight difference at the third decimal point, indicating that adding user or item features do not significantly contribute to the model's rating prediction accuracy.

Focusing on the intermittent lines of 4.3 - meaning the 'real-world' dataset (RWDS) - fMF's *Baseline* performance drops to a 0.132 score from the previous 0.156. However,

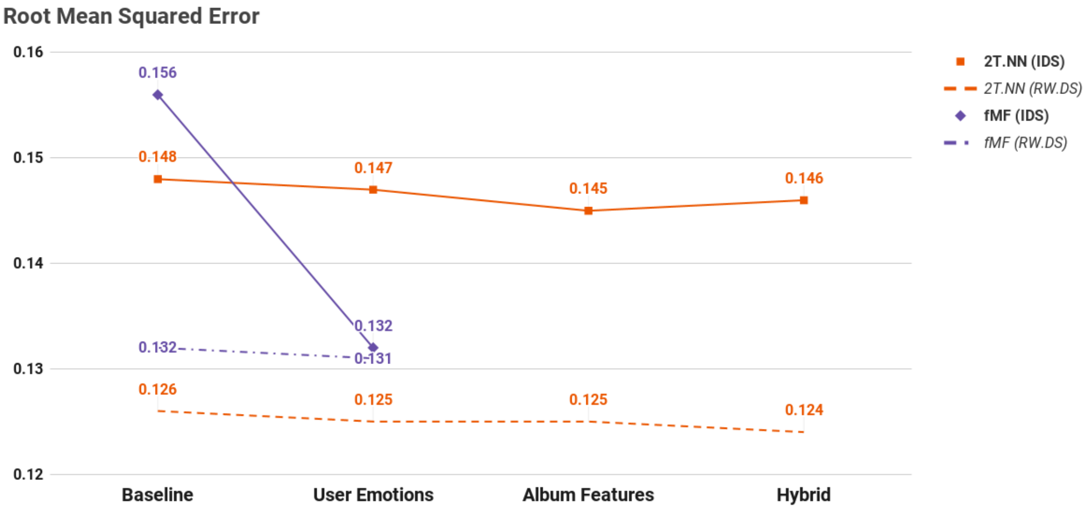


Figure 4.3: RMSE metric for evaluating errors in explicit rating prediction. The lower the value, the lower the error between predicted and actual ratings. Only Matrix Factorization (fMF) and Two-Tower Neural Networks (2T.NN) utilized explicit ratings.

the model's behavior does not imitate its performance on IDS. The *Baseline* score stays almost unaffected when *Users Emotions* are added, scoring 0.131 and not further lowering the model's errors as in the IDS. It can be argued that that the high emotional sparsity of RWDS, the 'dependent' variable, is responsible for this effect. Similarly, we notice a significant fall in 2T.NN's prediction error, from a median of 0.1465 down to 0.125 across all four feature combinations. *Hybrid* features have a minor advantage, of only 0.02 against the baseline, reinforcing the idea that auxiliary features do not meaningfully affect the accuracy of explicit rating prediction of 2T.NN. Moreover, the overall drop in terms of RMSE for 2T.NN can more likely be attributed to the increase in data scale. The ideal dataset consists of approximately 65.000 user-item interactions while the real-world datasets of 350.000. Neural networks are known for requiring a large amount of data in order to perform optimally and their improved performance in the RWDS can be explained as such.

Continuing with the evaluation of the models in terms of **Categorical Accuracy**, a 'bare-minimum' score is being set by performing 100 random recommendation, resulting in 5.3% and 4.9% Acc@100 for the IDS and RWDS respectively. Similarly, recommending the 100 most popular items to all users resulted in 44.4% and 39.2% for IDS and RWDS.

As can be seen in figure 4.4, there exists a significant divergence between the three model's performance. For the IDS, across all feature combinations, fMF has a mean performance of 24.6%, 2T.NN has 34.5% while Factorization Machines (FMs) have 60.9%. Similarly, for the RWDS, fMF have a mean score of 24.5%, 35.57% by 2T.NN and 56.4%

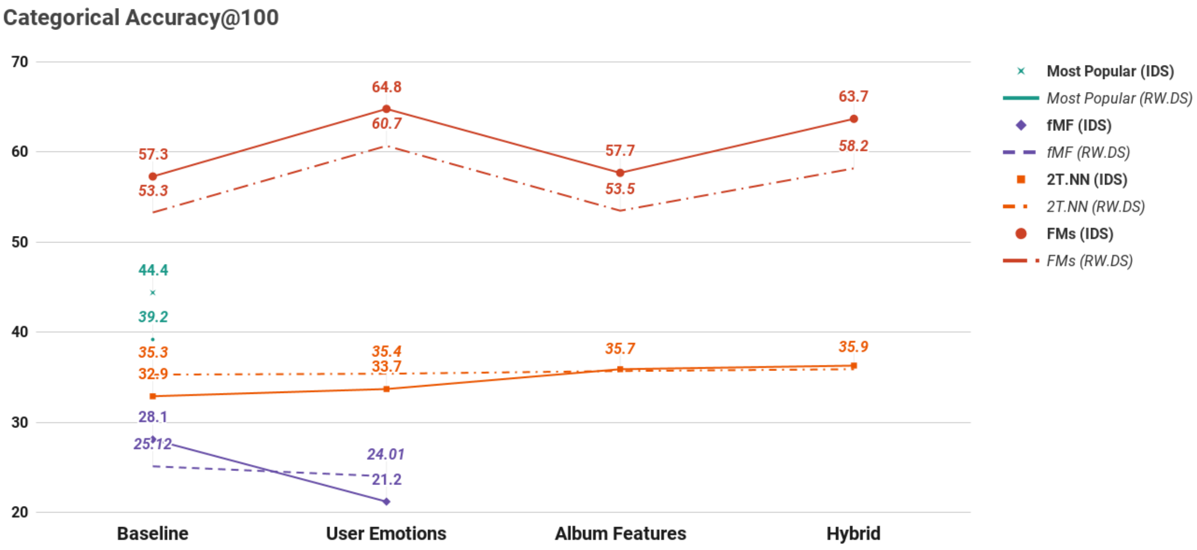


Figure 4.4: Categorical Accuracy at 100 recommended items for both the Ideal Dataset (IDS) and Real-World Dataset (RW.DS).

by FMs. FMs have the best average performance in terms of Acc@100, by a large margin, when compared to the other models and the 'most popular' recommendations on both datasets. Surprisingly, both fMF and 2T.NN perform lower than simply suggesting the 100 most popular items. This shows the scale of the power law distribution and the ingrained bias towards the most popular items.

By taking a more focused look, we can see that fMF's performance on the IDS has a notable drop when the *User Emotions* are utilized. This occurrence is in contrast to fMF's improved performance in terms of RMSE. As has been previously discussed in the 'Theoretical Background and Fundamentals' section, improvements in rating prediction do not necessarily translate to improvements in retrieval scores. This could be a criticism of many recent studies in the field of EA-RS, that solely relied upon RMSE for evaluating their model's performance. Similarly, we can also see that fMF yields slightly higher scores in *Baseline* (25.12%) than in *User Emotions* (24.01%) on the RWDS. In both datasets, *User Emotions* were not able to improve the retrieval accuracy of fMF.

More closely examining 2T.NN's performance, we can see that in IDS the *User Emotions* (33.7%) scores slightly better than its *Baseline* (32.9%), followed by a further increase in *Album Features* (35.6%) followed by *Hybrid* features (35.8%), the highest score of 2T.NN in the IDS. This 'progressive' improvement in Acc@100 indicates the potential usefulness of exploiting Users' Emotions and even more, rich Item features in datasets where no emotional sparsity is present. Although the same pattern can be identified on the RWDS, the improvements in performance are notably smaller, at the scale of +0.1%, +0.4% and +0.6% for *User Emotions, Album Features* and *Hybrid* respectively, when compared with their *Baseline* (35.3%).

Lastly, FMs, having the best average performance in terms of Acc@100, follow a different pattern than 2T.NN. In the IDS, a *Baseline* (57.3%) is followed by a slight increase with Album Features (+0.4%), followed by *Hybrid* (+6.4%), followed by the highest score of *User Emotions* (+7.5%) an impressive 64.8%. While on the RWDS, the *Album Features* have the same score as the *Baseline* (53.3%) who are followed by *Hybrid* (+4.9%), followed by the highest score of *User Emotions* (+7.4%), an impressive 60.7%. Remarkably, FMs' *User Emotions* performance remains very high on RW.DS despite the high levels of emotional sparsity. FMs show the best mean accuracy@100 when compared to all other models and *User Emotions* in particular notably enhances the model's performance. Based on these results, the trajectory of utilizing Factorization Machines for Emotion-Aware systems is very promising.

4.2.2 Beyond Accuracy Results

As was previously noted, apart from evaluating the predictive accuracy of the selected models it was deemed important to examine their performance in terms of Diversity, Cold Start and Scalability.

The metric of ***Item Coverage*** shows how many unique items a model selects, indicating the spectrum between Popularity Bias and Diversity of recommendations. Initially, the 'floor' and 'ceiling' scores are estimated by performing 'Most Popular' and Random recommendations yielding 0.5% and 99.8% respectively. Naturally, sampling random items from the complete catalog leads to a near 100% score, meaning that practically all items are suggested at least once. We also saw that random recommendations yielded only around 5% in Accuracy@100. We can conclude that Item Coverage on its own does not tell as about how correctly targeted the diverse suggested item lists are. But complemented with a satisfactory high Retrieval score, Item Coverage can indicate a model's ability to perform both accurate, informed and diverse recommendations. On the other hand, Most Popular recommendations had an adequate retrieval accuracy but its coverage is practically zero, since it is suggesting the same 100 popular items out of the 1997 album in the complete corpus.

Inspecting the *Baseline* of all models in figure 4.5, for the IDS, Item Coverage is consistently low, with fMF (4.8%), followed by 2T.NN (9.9%) and followed by FMs (11.3%). This is in accordance with current literature on Collaborative Filtering and its bias towards selecting the most popular items. However, adding the *User Emotions* leads to a significant increase for both fMF (20.4%) and 2T.NN (39.4%) while FMs have only

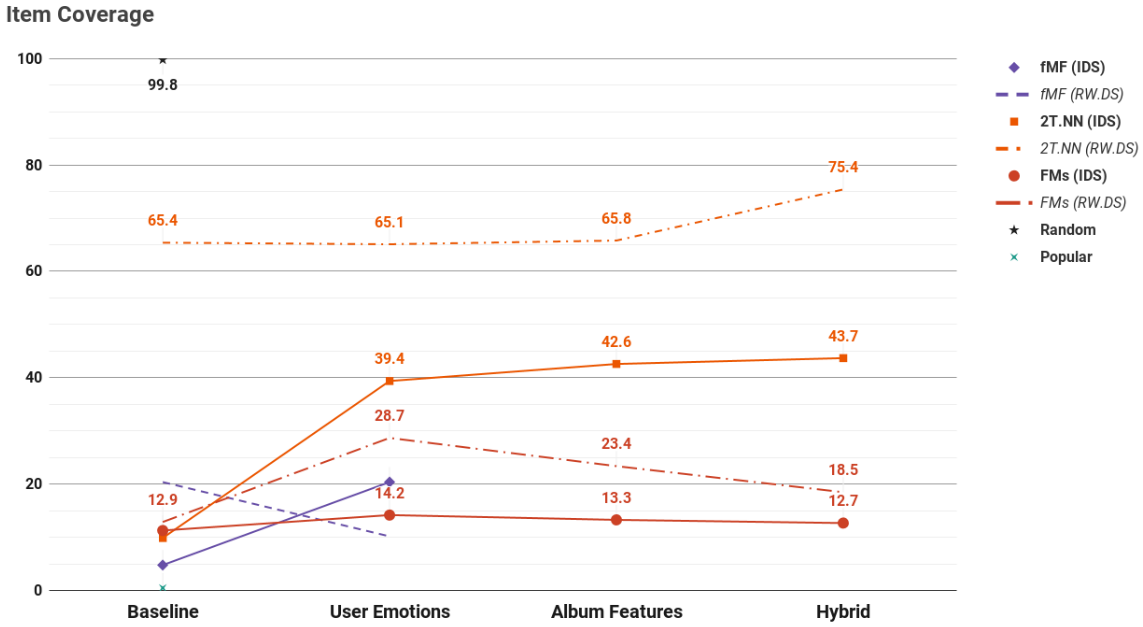


Figure 4.5: Coverage of total unique items in the recommended lists for both the Ideal Dataset (IDS) and Real-World Dataset (RW.DS).

slight increase, reaching 14.2%. From this point on, FMs' Coverage scores are reducing while 2T.NN with *Album Features* scores improve up to 42.6% and *Hybrid* yields the highest score for IDS, up to 43.7%. Moving on to the RWDS, fMF starts of with an improved 20.4% *Baseline* score but falls again down to 10.2% when *User Emotions* are used. This occurrence again may be the result of high levels of emotional sparsity. On the contrary, despite the increased emotional sparsity, FMs have an increased score of 28.7% with *User Emotions* compared to its *Baseline* 12.9%. But the highest score in terms of Item Coverage for the RWDS is reached by 2T.NN (75.4%) with 'Hybrid' Features outperforming all three feature combinations. Interestingly, 2T.NN's *Baseline*, using only collaborative information, is at the same high levels (65.4%) as 'User Emotions' (65.1%) and 'Album Features' (65.8%)

Moving on to the ***Novelty*** metric, indicating the ability of a model to suggest unconventional and less popular items, we have two 'floor' scores of 3.31 and 2.27 from 'Most Popular' for IDS and RWDS respectively and 'ceiling' scores of 9.23 and 7.51 for Random recommendations. Similarly to Item Coverage, there can be an inverse relationship between accuracy and novelty but finding a balance between high Accuracy and an acceptable rate of Novelty could ensure the mitigation of Popularity Bias. By examining figure 4.6 we can initially conclude that 2T.NN have the highest overall performance in terms of Novelty, with a mean score of 7.14 for IDS and 6.15 for RWDS. In IDS, a *Baseline* of 6.92 is superseded by 7.2 with *Album Features*, 7.27 with *Hybrid* features and

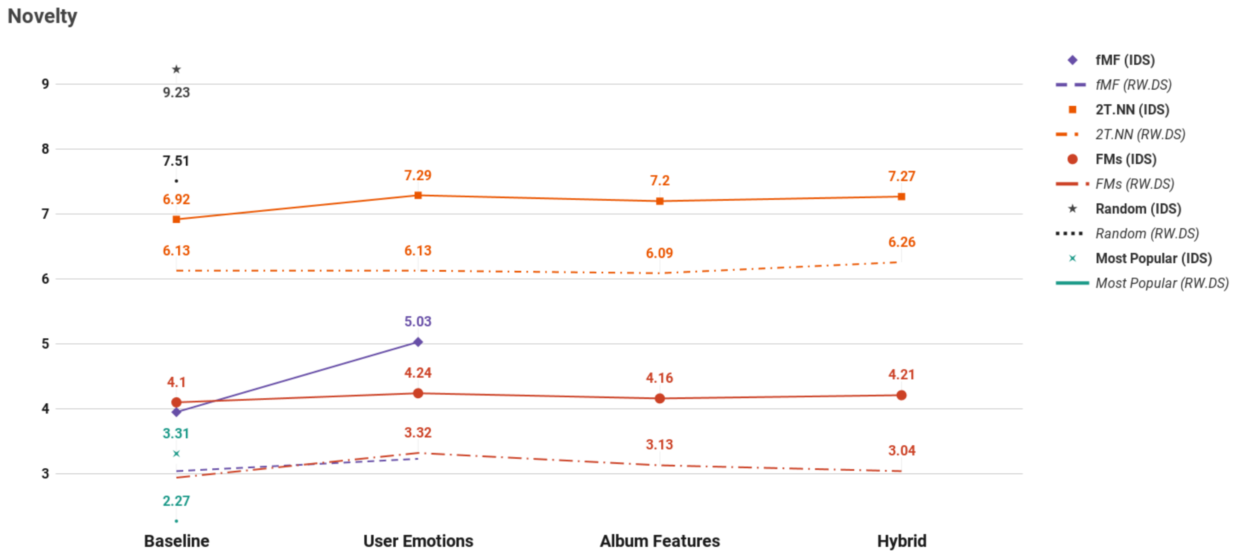


FIGURE 4.6: Novelty scores for both the Ideal Dataset (IDS) and Real-World Dataset (RW.DS).

7.29 for *User Emotions*, the highest overall score in IDS. In RDWS 2T.NN's *Baseline* scores the same as *User Emotions* at 6.13, higher than the 6.09 of *Album Features* but lower than the 6.26 of *Hybrid*, the highest score in RWDS. The high scores of 2T.NN indicate the model's ability to overcome Popularity Bias and perform both interesting and relevant suggestions. On the other hand, FMs have lower mean Novelty scores of 4.17 for IDS and 3.1 for RWDS, designating that model is prone to bias towards popular items and the bias intensifies with the increase of data size and the emotional sparsity. Although, fMF has a low novelty score of 3.95, 3.04 for its Baseline on IDS and RWDS but an increase with 'User Emotions' in the IDS, which again falls in RWDS. However, this coincides with the model's drop in Accuracy@100, hence it can not be considered a net gain.

Personalisation as a metric shows the dissimilarity between the recommended item lists or how customized they are to each specific user. A 'floor' score is again defined by performing the 100 most popular recommendations resulting in 0 personalisation since every user receives the same items, in the same order. On the contrary, random recommendations score 99.99%. An impressive 99% across all four feature combinations is reached by 2TNN for the RWDS, and 98% for the IDS with *Album* and *Hybrid* Features followed by *User Emotions* (97%) and the *Baseline* (90%). Despite the high levels of emotional data sparsity in the RWDS, the emotion-aware system is able to maintain the model's high score. FMs' behavior follows a similar pattern in both datasets, with the highest score being reached with *User Emotions* as input feature, at 89% for IDS and 85.1% for RWDS. In IDS, fMF have a low personalisation 50.4% for the *Baseline* and

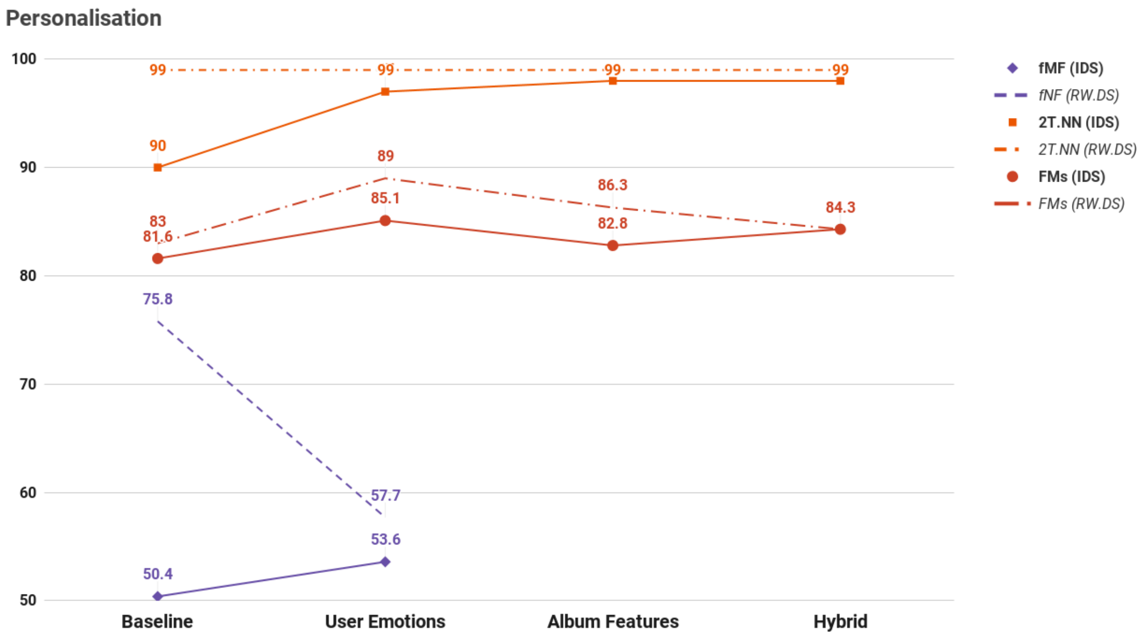


FIGURE 4.7: Personalisation scores for both the Ideal Dataset (IDS) and Real-World Dataset (RW.DS).

an insubstantial increase with *User Emotions* at 53.6%. In RWDS the fMF is able to reach 75.8% with its *Baseline* however, the score falls at 57.7% when *User Emotions* are utilized.

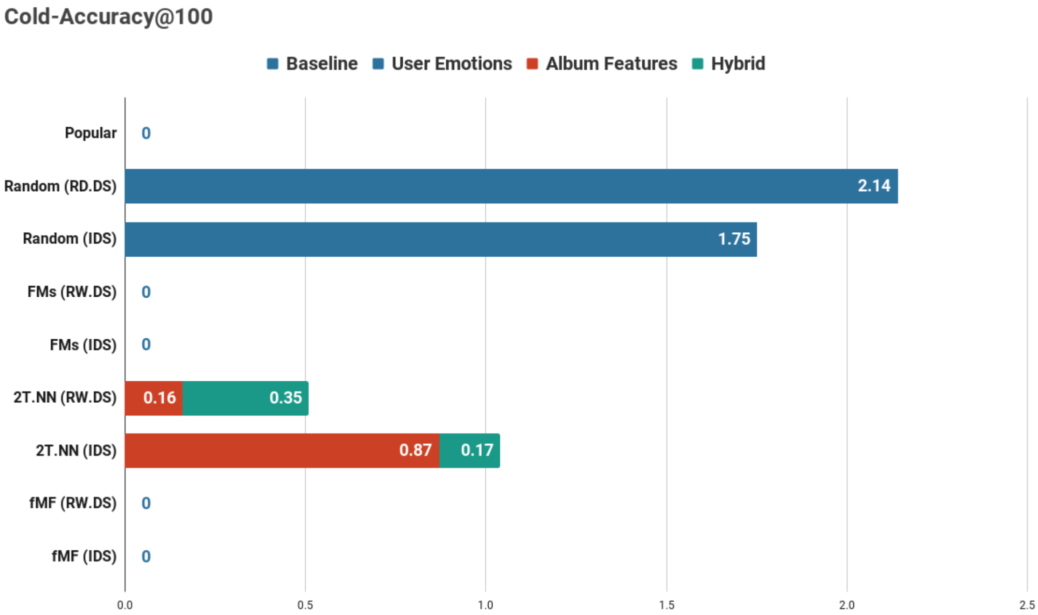


FIGURE 4.8: Accuracy@100 on Cold Items for both the Ideal Dataset (IDS) and Real-World Dataset (RW.DS).

Proceeding with the evaluation of each model in terms of their ability to perform accurate recommendation on purely ***Cold Start Items*** it is important to stress how difficult of a task that is. From the IDS, 570 interactions were filtered out with items that had fewer than 2 interactions, resulting in 387 cold items. Similarly, 321 items were filtered from the RDWS with 4198 total interactions in the cold set. Those items' names were present in each model known corpus, but there was no known interaction with those items for the models. Thereafter, when presented with the user names in the filtered cold-item set, each model had to predict the one cold item out of 1997 possible candidates, with no collaborative information being available in relation to those items. Even when performing completely random suggestions, which by definition have no bias towards popular items, the maximum scores are 2.14% and 1.75% for RWDS and IDS respectively. Naturally, suggesting the 100 most popular items, by definition scores 0% in both sets. But we can see in figure 4.8 that fMF and FMs are also completely unable to correctly predict cold interactions. Only 2T.NN are able to reach an above zero score when utilizing *Album Features* or *Hybrid Features* with a 0.16% and 0.35% respectively for the IDS and 0.87% , 0.17% for the RDWS. These scores are very low, but keeping in mind the difficulty of this task, even such low scores are a significant indication that 2T.NN are able to identify useful patterns in content-based features even when no collaborative information are available. Furthermore, even with the high levels of emotional sparsity present in the RWDS, using hybrid features (user emotions and album features) the model is able to retain the benefits of the former in the emotion-aware field.

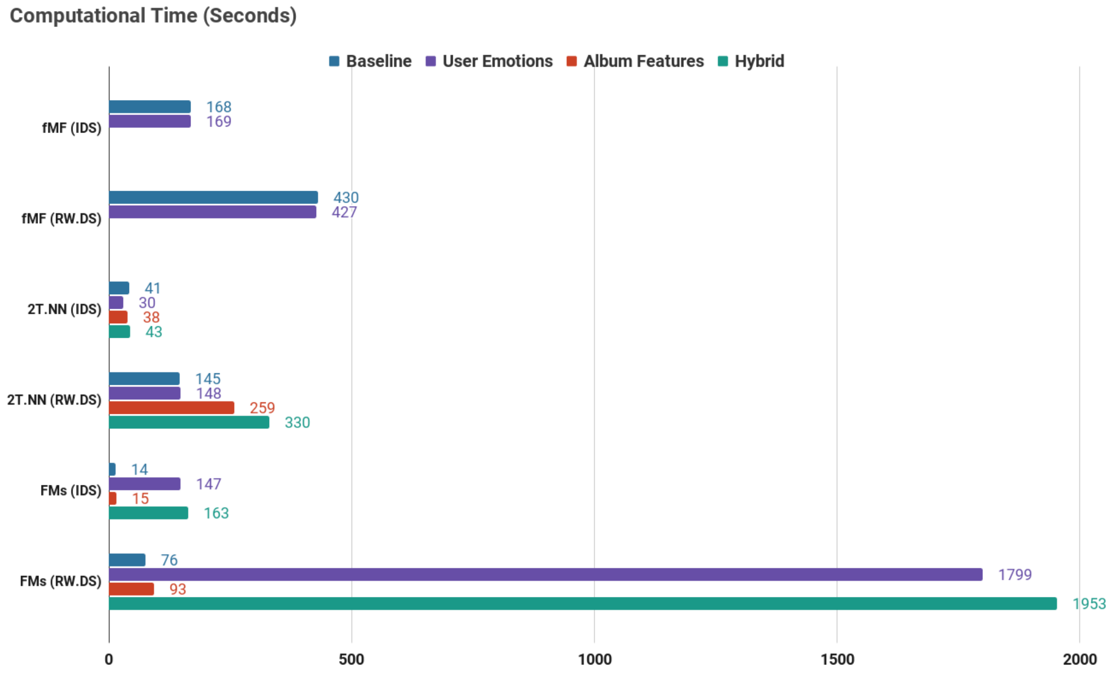


Figure 4.9: Computational Time in seconds for both the Ideal Dataset (IDS) and Real-World Dataset (RW.DS).

Finally, all models were evaluated in terms of their required computational time and by extension their **scalability**. IDS and RWDS are datasets of different scales, the former consisting of 70.000 interactions and 350.000 for the latter, which is approximately 5 times larger. Comparing the computational time that each model requires for training and performing predictions, offers an indication, even if limited, of how scalable they may be. In figure 4.9 we can see that fMF have practically no difference between *Baseline* and *User Emotions* on the same dataset, and has a 2.5 times increase in computational time when moving to RWDS. From 168.5 up to 428.5 seconds on average. Two tower neural networks have a lower average for both datasets with only 38 seconds on the IDS and 220.5 on the RWDS. But there is a noticeable difference when using *Album Features* with an increase from 145 for the *Baseline* up to 259 seconds for the *Album Features* and 330 for *Hybrid Features*. Still the overall computational time is lower than fMF even when both User and Album Features are employed. Lastly, Factorization Machines, while having an even lower computational time for *Baseline* and *Album Features* for both IDS and RDWS, they have an enormous spike in the required computational time when utilizing *User Emotions*. In RWDS, from a very low 76 *Baseline* (lower than the other models) it escalates up to 1799 seconds for *User Emotions* and 1953 with *Hybrid* Features. This can be attributed to the data formulation required by Factorization Machines, meaning that each feature has to be one-hot-encoded. This formulation is very efficient for static information such as demographic information or item features and this is reflected in the very low times required for ‘Album Features’ even in the RWDS. However, user emotions - and more generally contextual information - are dynamic and linked to particular items. Thus, one-hot-encoding these relationships leads to extremely sparse embeddings. As seen by their high accuracy@100 when using *User Emotions* FMs are able to deal with high dimensional and sparse data but in this case with the significant trade-off of increased computational times. This concludes the performance of each model on every possible feature combination on both the Ideal and the Real-World datasets on all seven evaluation metrics. In the next section, the individual results are compounded and discussed on from ‘big picture’ perspective.

4.2.3 Discussion

In this section are discussed the findings and drawn insights from the behavior and results of each model in the context of emotion-aware systems and the consequences of emotional sparsity and the potential usefulness of hybrid algorithms in mitigating their effects. Initially, examining the behavior of Collaborative Filtering, specifically Funk’s **Matrix Factorization**, it was determined that the predictive error, in terms of RMSE, was decreased when using users’ emotions, with a -0.024 improvement. This finding was in accordance with the majority of current research on emotion-aware recommendation systems as discussed in the literature review. However, this improvement was not similarly present in terms of Retrieval Accuracy, with a significant decrease of -7%, validating the criticism of solely relying on error minimization in rating prediction for evaluating recommendation systems [33]. Furthermore, when a high level of emotional sparsity was present in the 'real-world dataset', the emotion-aware collaborative model had a decrease in terms of Item Coverage (-10%) and Personalisation (-15%) indicating the enhancement of popularity bias and decrease in performance when user emotions were utilized. This phenomenon validates the central hypothesis of this study, that aggregated missing emotional values, and by extension emotional sparsity, would negatively affect emotion-aware systems solely based on collaborative filtering. Moreover, in all cases, the cold-item-accuracy is equal to zero, regardless the existence of emotional sparsity or lack thereof. The two aforementioned issues indicate that traditional collaborative filtering methods are not appropriate for emotion-aware systems and more complex algorithms should be considered for real-world applications.

Secondly, **Factorization Machines** trained on implicit interactions with Learn-to-Rank objective function (warp), were able to reach impressive levels of accuracy100, with an average of 55.3% for the baseline, when no additional feature was considered. Even more impressive, was the model's performance when Users' emotional reactions were utilized, reaching 62.75% on average for both datasets, yielding the highest score compared to all other models. Although, taking advantage of additional album-side features was not able to further improve the model's performance. User emotions were also able to improve the model's performance in terms of Personalisation, scoring an average of 87% (+5% compared to the model's Baseline) and scoring significantly higher than Matrix Factorization (59.3% total average). Conversely, the FMs' performance was not as formidable in terms of Novelty and Item Coverage. On the former it scored a total average of 3.65, even lower than MF's 3.78, and 17% for the latter slightly higher than MF's 13.75%. Similarly to MF, FMs were not able to address the issue of item cold start, with 0% accuracy in cold-and. Finally, despite User Emotions yielding the highest accuracy100 score, a notable trade-off was present in terms of the increase in computational time, a 1750% increase for the ideal dataset and 2267% increase in the real-world dataset. This dramatic increase can be attributed to two factors. Firstly, in the hyper-tuning parameter process, parameter combination that yielded the best results included more training epochs and more factorized components. Lowering those values, could still retain high levels accuracy100 in less required time. Indicatively, lowering the factorized components from 50 to 10, in the real-world dataset, the emotion-aware FM could reach 57.68% in 495 seconds, an 78% decrease from the best performing parameter combination but still an a 551% increase compared to the Baseline. The second attributable reason implicates the FMs' central architecture. Factorization Machines rely on sparse one-hot-encoded representations of the data. Doing so on in relation to users' emotional reactions which in turn are related to an individual item, creates an extremely high dimensional matrix, than even in its sparse representation hinders the model's performance. Consequently, we can conclude that in spite of FMs' accomplished performance in terms of Accuracy100, they can not overcome the item-cold start issue, nor mitigate the popularity bias and furthermore, the trade-off of increased computational time would have to be taken into consideration when building real-world applications.

Lastly, **Two-Tower Neural Networks**, were able to surpass Matrix Factorization both in terms of rating error minimization (RMSE), with an average of 0.125 over 0.131, and retrieval accuracy100, with an average of 35.6% over 24.56% for the 'real-world dataset'. However, they Accuracy was still lower that simply recommending the 100 Most-Popular (41.8% on average) items and certainly lower than Factorization Machines (58.75% on average). On the flip side, 2T.NN were able to reach the highest scores across Item Coverage, Novelty and Personalisation compared to the other two models. In the Emotion-Aware version, these scores remained high even when high levels of emotional sparsity were present in the dataset with 65.1% Item Coverage, 6.13 in Novelty and an impressive 99% Personalisation score. Hybrid Features were able to further improve Coverage (+10.3%) and Novelty (+0.13). Moreover, 2T.NN when utilizing Album and Hybrid Features was the only model able to score an above-zero score in terms of Cold-Accuracy100. These results indicate that 2T.NN are able to somewhat mitigate the **Cold Start** of new items and significantly overcome the **Popularity Bias**, since it was able to score very high on 'Diversity Metrics' while maintaining a decent level of Predictive Accuracy as well as a very low Rating Prediction Error (RMSE). All these while maintaining the lowest computational times in the real-world dataset for hybrid features, 330 seconds compared to 427 of MFs and 1953 of FMs. All of the above, give an important precedence to 2T.NN for EA-RS applications.

Microscopically, in the context of Emotion-Aware Recommendation Systems both hybrid models could be useful in industrial applications but for different tasks. Factorization Machines' high categorical accuracy would be ideally applicable as the main emotion-aware engine, recommending items that are most certainly relevant to the users' current preferences and closely resemble their 'favorite' types of music. On the other hand, 2T.NN could be very useful in recommending more diverse, novel and new items that are relatively relevant to the users' interests, striking a balance between 'Diverse but Relevant' and 'Novel but not Random', similarly to Spotify's Discover Weekly Playlists but in the context and advantages of emotion-aware applications.

Chapter 5

Conclusions

The previous chapter presented the experimentation framework and the derived results of the present project. This Chapter summarizes the learned insights from the conducted research and suggests further improvements and potential directions for future work in developing emotion-aware recommendation systems for music. Firstly, the central hypothesis of this study was that Emotion-Aware Recommendation Systems (EA-RS) for Music applied in a real-world situation, where high levels of emotional data sparsity was present, it would negatively affect their performance in terms of Popularity Bias and Item Cold Start. This phenomenon was identified in Memory-Based Collaborative Filtering and Matrix Factorization models.

The examined solutions to these challenges involved the developed of Hybrid Recommendation models that utilized both collaborative information and content-based features. More specifically, Factorization Machines with Learning-to-Rank objective function and Two Tower Neural Networks (2T.NN) with Multi-Task optimisation were selected. It was concluded that Emotion-Aware Factorization Machines while having the highest levels of Predictive Accuracy, they were not able to mitigate their Popularity Bias and Item Cold Start. On the contrary, 2T.NN were able to retain high levels of Item Coverage and Novelty even when high rates of emotion data sparsity was present in the dataset. This translates into higher levels of diversity and by extension into lower levels of popularity bias while maintaining decent rates of predictive accuracy and very low rates of rating prediction error. Furthermore, 2T.NN, when utilizing Album Features in combinations with Users' Emotions, were able accurately recommend previously unseen items and thus mitigate the cold start of new items. These outcomes situate 2T.NN as a credible solution for both challenges in the context of Emotion-Aware Recommendation Systems.

Despite the ability of 2T.NN to alleviate issues of Popularity Bias and Item Cold Start, some additional adjustments could be considered in order further improve the model's performance. Initially, Neural Networks are known for demanding substantial amounts of data in order to generalise well. In the present study approximately 350.000 user-item interactions were utilized in the 'real-world dataset' that could be considered limited in the context of training neural networks. Hence, continuing the data collection process and enriching the dataset with more user-item interactions could potentially considerably ameliorate the models performance. In case that collecting more data was not feasible, Transfer Learning techniques could be considered instead. The model could first be trained on a general dataset for music recommendations such as the Million Song Dataset [51] and then be fine-tuned on the emotion-aware dataset. Moreover, in this study the ability of Neural Networks to utilize multi-modal inputs was not explored. Instead of using a pre-trained emotion analysis machine learning model, the raw text could be integrated in the neural network's pipeline and trained to recognize user emotions end-to-end. Lastly, since Factorization Machines utilizing WARP, a Learning-to-Rank objective function, had significantly higher accuracy rates, it would a worthwhile attempt to integrate learning-to-rank into 2T.NN. Otherwise, mixed sampling [26] and sampling bias correction [64] have been shown to be beneficial techniques for 2T.NN and that could also be the case for EA-RS.

Regarding the corollary insights that can be drawn from this study, re-evaluating the potential usefulness of EA-RS, it can be deduced that utilizing users' emotional reactions can reduce the error of rating prediction in Collaborative Filtering models and significantly improve the Predictive Accuracy and Personalisation of Factorization Machines while Two Tower Neural Networks were not significantly improved when utilizing the users' emotions on their own. Subsequently, from the results presented in the precious chapter, we can also deduce that Hybrid Filtering was beneficial for EA-RS, first for increasing the accuracy of Factorization Machines and secondly for overcoming the popularity bias and cold start in 2T.NN. More specifically, content-based features involved the use of two novel proposals: 1) Multi-Level Profiling and 2) Cross Platform Audience Reaction that were proven appropriate for the task.

In terms of how the present study can assist future research works, the following ideas could be considered :

1. Utilizing both pre- and post- interaction emotional responses. Tracking the emotional reactions before a user chooses to interact with an item and re-tracking after the end of the song could more precisely estimate the effect that a musical item

has to each user. This in turn could potentially be instrumental in building more robust EA-RS for music.

2. Emotional-awareness could also be incorporated in the generation of musical playlists. This could possibly create more dynamically changing playlists that would adapt in relation to the user's emotional responses and thus improve user satisfaction and situational personalisation.

3. Even further, combining the two aforementioned ideas could lead to 'emotion-directed systems'. The user could explicitly select their desired change in terms of emotional state and the recommendation system would make attempt to reach that goal. For example, a user currently feeling 'stressed' could select to reach a more 'relaxed' emotional state, thus the emotion-aware playlist generation would attempt to progressively reach this objective by tracking the user's reactions in real-time. This idea has been previously explored in [19] but it could be further improved by utilizing contemporary State-of-the-Art techniques.

4. Finally, it should be noted that while Emotion-Aware systems, as a research field, may have a bright future ahead of them, it is considered important to also thoroughly examine the moral dimensions of using these systems. Having access to the psychological states and traits of individual users should raise important ethical questions in terms data privacy and protection since it involves sensitive personal information. Future researchers should take those seriously under consideration especially when developing emotion-aware applications.

Bibliography

[1] Martin J Eppler and Jeanne Mengis. The concept of information overload-a review of literature from organization science, accounting, marketing, mis, and related disciplines (2004). *Kommunikationsmanagement im Wandel*, pages 271–305, 2008.

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*, 2019.

[3] Sandeep K Raghuwanshi and RK Pateriya. Recommendation systems: techniques, challenges, application, and evaluation. In *Soft Computing for Problem Solving*, pages 151–164. Springer, 2019.

[4] Erion Çano and Maurizio Morisio. Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, 21(6):1487–1524, 2017.

[5] Marwa Hussien Mohamed, Mohamed Helmy Khafagy, and Mohamed Hasan Ibrahim. Recommender systems challenges and solutions survey. In *2019 International Conference on Innovative Trends in Computer Engineering (ITCE)*, pages 149–155. IEEE, 2019.

[6] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[7] Kurt Jacobson, Vidhya Murali, Edward Newett, Brian Whitman, and Romain Yon. Music personalization at spotify. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 373–373, 2016.

[8] Melissa Onori, Alessandro Micarelli, and Giuseppe Sansonetti. A comparative analysis of personality-based music recommender systems. In *Empire@ RecSys*, pages 55–59, 2016.

[9] Renata L Rosa, Demsteneso Z Rodriguez, and Graça Bressan. Music recommendation system based on user's sentiments extracted from social networks. *IEEE Transactions on Consumer Electronics*, 61(3):359–367, 2015.

[10] Ivana Andjelkovic, Denis Parra, and John O'Donovan. Moodplay: Interactive music recommendation based on artists' mood similarity. *International Journal of Human-Computer Studies*, 121:142–159, 2019.

[11] Shuiguang Deng, Dongjing Wang, Xitong Li, and Guandong Xu. Exploring user emotion in microblogs for music recommendation. *Expert Systems with Applications*, 42(23):9284–9293, 2015.

[12] Yongfeng Qian, Yin Zhang, Xiao Ma, Han Yu, and Limei Peng. Ears: Emotion-aware recommender system based on hybrid information fusion. *Information Fusion*, 46:141–146, 2019.

[13] Aurobind V Iyer, Viral Pasad, Smita R Sankhe, and Karan Prajapati. Emotion based mood enhancing music recommendation. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 1573–1577. IEEE, 2017.

[14] Deger Ayata, Yusuf Yaslan, and Mustafa E Kamasak. Emotion based music recommendation system using wearable physiological sensors. *IEEE transactions on consumer electronics*, 64(2):196–203, 2018.

[15] Shlok Gilda, Husain Zafar, Chintan Soni, and Kshitija Waghurdekar. Smart music player integrating facial emotion recognition and music mood recommendation. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 154–158. IEEE, 2017.

[16] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. A deep multimodal approach for cold-start music recommendation. In *Proceedings of the 2nd workshop on deep learning for recommender systems*, pages 32–37, 2017.

[17] Maciej Kula. Metadata embeddings for user and item cold-start recommendations. *arXiv preprint arXiv:1507.08439*, 2015.

[18] Qian Zhang, Dianshuang Wu, Jie Lu, Feng Liu, and Guangquan Zhang. A cross-domain recommender system with consistent information transfer. *Decision Support Systems*, 104:49–63, 2017.

[19] Willian Garcias de Assuncao and Vania Paula de Almeida Neris. An algorithm for music recommendation based on the user's musical preferences and desired emotions. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, pages 205–213, 2018.

[20] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[21] Ferdos Fessahaye, Luis Perez, Tiffany Zhan, Raymond Zhang, Calais Fossier, Robyn Markarian, Carter Chiu, Justin Zhan, Laxmi Gewali, and Paul Oh. T-recsys: A novel music recommendation system using deep learning. In *2019 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–6. IEEE, 2019.

[22] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 101–109, 2019.

[23] Xiangnan He and Tat-Seng Chua. Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 355–364, 2017.

[24] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[25] Steffen Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000. IEEE, 2010.

[26] Ji Yang, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Xiaoming Wang, Taibai Xu, and Ed H Chi. Mixed negative sampling for learning two-tower neural networks in recommendations. In *Companion Proceedings of the Web Conference 2020*, pages 441–447, 2020.

[27] Malay Haldar, Prashant Ramanathan, Tyler Sax, Mustafa Abdool, Lanbo Zhang, Aamir Mansawala, Shulin Yang, Bradley Turnbull, and Junshuo Liao. Improving deep learning for airbnb search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2822–2830, 2020.

[28] Ervine Zheng, Gustavo Yukio Kondo, Stephen Zilora, and Qi Yu. Tag-aware dynamic music recommendation. *Expert Systems with Applications*, 106:244–251, 2018.

[29] Diego Sánchez-Moreno, Ana B Gil González, M Dolores Muñoz Vicente, Vivian F López Batista, and María N Moreno García. A collaborative filtering method for music recommendation using playing coefficients for artists and users. *Expert Systems with Applications*, 66:234–244, 2016.

[30] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The impact of popularity bias on fairness and calibration in recommendation. *arXiv preprint arXiv:1910.05755*, 2019.

[31] Dominik Kowald, Markus Schedl, and Elisabeth Lex. The unfairness of popularity bias in music recommendation: A reproducibility study. In *European Conference on Information Retrieval*, pages 35–42. Springer, 2020.

[32] Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2):75–79, 2007.

[33] Harald Steck. Training and testing of recommender systems on data missing not at random. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–722, 2010.

[34] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10): 4511–4515, 2010.

[35] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 257–260, 2010.

[36] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116, 2018.

[37] Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.

[38] Martin Pichl, Eva Zangerle, Günther Specht, and Markus Schedl. Mining culture-specific music listening behavior from social media data. In *2017 IEEE International Symposium on Multimedia (ISM)*, pages 208–215. IEEE, 2017.

[39] Ranran Wang, Xiao Ma, Chi Jiang, Yi Ye, and Yin Zhang. Heterogeneous information network-based music recommendation system in mobile networks. *Computer Communications*, 150:429–437, 2020.

[40] Sumet Darapisut, Ureerat Suksawatchon, and Jakkarin Suksawatchon. The constant time of predictive algorithm for music recommendation with time context. In *2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 63–68. IEEE, 2015.

[41] Thanh Huy Ly, Song Toan Do, and Thi Thanh Sang Nguyen. Ontology-based recommender system for the million song dataset challenge. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 236–241. IEEE, 2018.

[42] Kazuyuki Matsumoto and Manabu Sasayama. Lyric emotion estimation using word embedding learned from lyric corpus. In *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pages 2295–2301. IEEE, 2018.

[43] Erion Çano and Maurizio Morisio. Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pages 118–124, 2017.

[44] Christopher Beedie, Peter Terry, and Andrew Lane. Distinctions between emotion and mood. *Cognition & Emotion*, 19(6):847–878, 2005.

[45] Paul Ed Ekman and Richard J Davidson. *The nature of emotion: Fundamental questions.* Oxford University Press, 1994.

[46] Bruce Ferwerda, Marko Tkalcic, and Markus Schedl. Personality traits and music genre preferences: how music taste varies over age groups. In *1st Workshop on Temporal Reasoning in Recommender Systems (RecTemp) at the 11th ACM Conference on Recommender Systems, Como, August 31, 2017.*, volume 1922, pages 16–20. CEUR-WS, 2017.

[47] Abhishek Paudel, Brihat Ratna Bajracharya, Miran Ghimire, Nabin Bhattarai, and Daya Sagar Baral. Using personality traits information from social media for music recommendation. In *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*, pages 116–121. IEEE, 2018.

[48] Feng Lu and Nava Tintarev. A diversity adjusting strategy with personality for music recommendation. In *IntRS@ RecSys*, pages 7–14, 2018.

[49] Rui Cheng and Boyang Tang. A music recommendation system based on acoustic features and user personalities. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 203–213. Springer, 2016.

[50] Yin Zhang, Xiao Ma, Shaohua Wan, Haider Abbas, and Mohsen Guizani. Crossrec: Cross-domain recommendations based on social big data and cognitive computing. *Mobile networks and applications*, 23(6):1610–1623, 2018.

[51] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. 2011.

[52] Despoina Chatzakou, Athena Vakali, and Konstantinos Kafetsios. Detecting variation of emotions in online activities. *Expert Systems with Applications*, 89:318–332, 2017.

[53] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, 2018.

[54] Saif M Mohammad and Peter D Turney. Nrc emotion lexicon. *National Research Council, Canada*, 2, 2013.

[55] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, 2018.

[56] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[57] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *2008 Fourth international conference on natural computation*, volume 4, pages 192–201. IEEE, 2008.

[58] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[59] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 263–272. Ieee, 2008.

[60] Tie-Yan Liu. Learning to rank for information retrieval. 2011.

[61] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[62] Jason Weston, Hector Yee, and Ron J Weiss. Learning to rank recommendations with the k-order statistic loss. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 245–248, 2013.

[63] Steffen Rendle and Christoph Freudenthaler. Improving pairwise learning for item recommendation from implicit feedback. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 273–282, 2014.

[64] Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 269–277, 2019.