Affect- and Personality-Aware Recommender System

Student Name: Kah Gene Leong
Supervisor Name: Sunčica Hadžidedić
Submitted as part of the degree of BSc Computer Science to the
Board of Examiners in the Department of Computer Sciences, Durham University

Abstract—

Background - Contextual information in recent years has found to be significantly important in recommendation systems. Coupled with the growth of personality and affect computing, the area of research for personality-and/or-affect-aware recommender systems have started to expand. Due to the fact that personality and sentiment do show notable correlations with books, an under-researched recommender domain, this project was initiated to fill in the gaps. To the best of our knowledge, this specific research and implementation that we will be conducting would be a novel approach, allowing the possibility of further expansion.

Aims - The project aims to build a functional personality and sentiment-aware book recommender system. By doing so, we will also investigate the impact of incorporating these two contexts on the accuracy and quality of recommendations generated. Additionally, the paper also aims to look at a potential novel method of inferring Big Five Personality scores of users through a non-model-based approach.

Method - Generation of user and book profiles was achieved using a novel LIWC dictionary-based approach and sentiment analysis. Collaborative Filtering was the underlying recommendation technique for this research, namely matrix factorisation via Singular Value Decomposition. Context post-filtering was selected to re-rank and filter candidate recommendations by personality and sentiment. Three configurations of our model were then evaluated (by several metrics) through offline experiments using the Goodreads Dataset.

Results - We had found that personality and sentiment do produce notable impacts on a book recommender system's performance, both positive and negative. The recommendations' level of personalisation from the context-aware models were at least eight times better than the baseline. In contrast, their diversity scores were no better than the baseline. Additionally, the sentiment-aware model had the highest precision and recall, yet models that incorporated personality showed potential of outperforming at a higher number of recommendations.

Conclusion - The use of both personality and sentiment seems promising for recommender systems, especially in the book domain. Sentiment has shown to be the more dominant context in affecting the performance of a recommender, however, personality does complement it well in optimising larger number of recommendations. Additionally, our novel approach of inferring user's personality scores has shown to hold up against machine-learning-based approaches.

Index Terms—Affective computing, Big Five Personality, Books, Collaborative filtering, Recommender Systems, Sentiment analysis, Automatic Personality Recognition

1 Introduction

With a staggering amount of online information today, recommender systems have been widely adopted to assist web users to filter through them to obtain data/items that match their preferences [4]. Recommender systems (RS) are tools that combine machine learning, business and human-computer interaction (HCI) to generate personalised or non-personalised recommendations to users. With non-personalised recommendations, users are usually suggested items based on generic or simple categorizations, such as best-seller books or top-10 billboard music. Personalised recommendations, however, make predictions and ranks them based on their relevance to a user's preferences and constraints [4]. This allows recommendations to be tailored and diverse for different individuals or groups of individuals [4].

Historically, Content-Based Filtering, Collaborative Filtering and Hybrid Filtering were three of the most commonly used approaches to perform item recommendations [4], [17]. These approaches followed the traditional recommendation process as below [17]:

1) User interests are recorded through item ratings
2) Users and items are filtered/modelled based on the similarities of their profiles
3) Relevant items found from step 2 are recommended to the user.

According to Adomavicius et al. [2], these techniques are fairly simple and context-independent, thus may lose their predictive power as useful information about a user or item is lost due to aggregation. Contextual information does play a vital role in RS to generate ideal recommendations as it influences human behaviour in various disciplines, e.g., a user's preference for one item may differ from one context to another [2], [4], [34], [59].

While context commonly refers to ”conditions or circumstances which affect something” [4], it is a multifaceted concept that has been studied across a variety of fields [2]. Hence, there is no one specific definition that can be applied here. However, we can closely follow the classification given by [2] where context is identified as a condition, that can have a structure, where recommendations are provided. An example relevant to our research is the personality factor, which can be defined in terms of Openness, Conscientiousness, Neuroticism, Extraversion and Agreeableness. Furthermore, another key point to note is that contextual factors can be static (remains with time) or dynamic (changes with time) [2] - these two categories will be brought up again in the next subsection.

The adoption of context-aware recommender systems (CARS) have been increasingly prevalent today as more sites/platforms want to generate more personalised and relevant content for their users [4]. This is achieved by the nature of CARS model having the ability to adapt recommendations based on the specific contextual situation of a user [2], [4]. A simple example is a music recommender that generates playlists based on temporal (time) and mood contexts. It could suggest uplifting songs for users in the morning to get them feeling energized for the day while in contrast, calmer songs would be recommended at night when a user wants to relax and go to sleep.

In this day and age, we are witnessing a rapidly emerging field of psychoinformatics that aims to highlight the relationship between psychology and computer science [37]. By observing HCI, variables such as personality traits, user behaviour, and affect can be examined, extracted and used to further enhance machine learning performances [37]. The two specific psychological factors explored in this research were personality and affect. More specifically, they are considered as modal contexts based on the classifications of Adomavicius et al. [2], representing a user's current state of mind and cognitive abilities. The reason behind this selection of contexts was due to their popularity gain in the recommender system space. This likely stemmed from the fact that research has proven notable impacts of these two factors have on a user's preference/decision-making in various disciplines [12], [40], [59].

1.1 Background

The following section will explain the nature of the two modal contexts relevant to this project. It then proceeds to compare the difference between a traditional RS and a personality-and-sentiment aware RS.

1.1.1 Personality

Often unique to an individual, personality is a combination of their characteristics and qualities which remain persistent in various situations [12], [14]. It has also been known to affect their decision-making capability [12], [14]. As user personality is known to not be relatively static, one of the main benefits of performing personality extraction is that it is typically a one-time process.

Research [17] has shown that personality traits of a user can typically be extracted using two methods: questionnaires and Automatic Personality Recognition (APR). This is justified by the fact that related studies in Section 2.2 and 2.3 leveraged either one of these methods in their implementation. While questionnaires are widely adopted for their accuracy and simplicity, it is time and effort consuming to gather a substantial amount of data and users may experience response burden [17]. APR, on the other hand, infers the personality traits from existing user data and hence is a more viable option from a technological research standpoint [56]. This is beneficial to combat common issues faced in this research field, for instance, anonymous datasets and the inconvenience of using a (personal) questionnaire. The 'existing user data' mentioned usually comes in the form of linguistic features of texts [12], multimedia (i.e. image, voice or video) and user behaviour (e.g., purchase habits of a user) [17].

The most common classification system used for personality traits is the very brief measure of Big-Five Personality (Openness, Neuroticism, Agreeableness, Extraversion and Conscientiousness) [21]. Hence, it has become a popular model to use in studies and tests across domains which allows it to be fairly reliable and consistent. There is, however, another well-known ”non-clinical psychometric assessment” model called Myers-Briggs Type Indicator (MBTI) which uses a combination of four values (Extraversion/Introversion, Sensing/Intuition, Thinking/Feeling, Judgment/Perception) to determine a user's personality and hence their behaviour [50]. While a popular model, it is not well explored in terms of its linguistic characteristics as it is a qualitative approach and relies on theoretical contexts [50]. Moreover, the additional time and resources required to exploit such a complex and multi-item scale also justified that MBTI was not a feasible personality model for our research. All questionnaire and APR-related techniques discussed in Section 2 all make use of the Big Five Personality.

The Big Five Personality model is often times also regarded as the Five Factor Model (FFM) where these factors/traits are shown in Table 1. Traditionally, these traits are formalised as bipolar scales as to aid with computational processing [14]; However, more recent research have represented the traits on a numeric scale (e.g., 1 - 5) [12] [19].

TABLE 1
Big Five Personality traits

Trait	Scale
Openness	insightful vs unimaginative
Conscientiousness	organized vs careless
Extraversion	sociable vs shy
Ageeableness	friendly vs uncooperative
Neuroticism	calm vs neurotic

Personality computing is the interdisciplinary study field which aims to extract personality-relevant information to be used in computing systems. The term was first coined by Vinciarelli and Mohammadi [56] who did a survey on such systems and addressed three fundamental problems found in literature, APR (which will be explored in this paper), Automatic Personality Perception and Automatic Personality Synthesis [56]. The use of personality computing has proliferated into various domains and research directions in the last decade, especially recommender systems and human-computer interaction. This is because research has shown the existence of a substantial relationship between user behaviour and a user's preferences [6].

As highlighted at the start, the incorporation of contextual information has proliferated in various domains as it allows recommendations to be made based on varying specific circumstances [4]. According to Nguyen et al. [40], personality (as contextual data) plays a big part in affecting people's media choices, for example, highly conscientious people like action and adventure. The research conducted by Cantador et al. [12] also supports the idea of the existence of a correlation between a user's personality and their decision-making in various media forms, one of which is books. An example of this is that people who have high openness tend to prefer Poetry and Sci-Fi books while those who are less open go with horror and crime books more [12].

1.1.2 Affect

According to the American Psychological Association [7], affect refers to any experience of feeling or emotion. A user's feelings and emotions can heavily influence their behaviour, and thus affects what a user chooses [59] and how they learn rationally. According to Ishanka and Yukawa [25] and Zheng et al. [63], emotions were first used as a contextual parameter in the recommendation system domain by Gonzalez et al. [20]. Since then, domains such as movie or music recommendations have seen several recommender systems integrating emotions and other affective data into their algorithms.

Like personality computing, affective computing also stands at an intersection between computer science and psychology, just with the additional discipline of cognitive science [10]. Its objective is to track and extract the emotional state of a user [10]. There are a number of concepts that have been extensively studied in this field, namely: affect, emotion, mood and feeling [1]. These concepts are still used loosely today despite the amount of research in these fields and hence are most of the time difficult to distinguish in papers [1]. Cambria et al. [1] further elaborated on this issue by providing the following definitions to illustrate the likeness of these concepts:

i Affect is a feeling or emotion usually expressed through facial or body language

ii Emotion is a mental state that happens suddenly and usually eludes physiological changes

iii Mood is a state of mind or emotion

iv Feeling is an affective state of consciousness which is caused by emotions and sentiments

The above definitions given mostly pertain to these terms being used from a psychological standpoint. However, the nature of our project requires a computational method of studying a person's affective state [1], leading to the adoption of sentiment analysis. This approach mines or extracts a user's emotions, mood, etc from natural language (i.e., text), facial expressions and other visible signs that have been pre-processed [1], [8].

Focusing on the word 'sentiment', it is the underlying feeling, attitude, emotion or evaluation that is implied by a personal opinion towards a certain target [1]. This target is an entity upon which a sentiment had been expressed, and its existence is the main discriminant between sentiment and affect [1]. Since the premise of item recommendation relies on understanding a user's feeling/preference towards the item (target), sentiment was deemed a more appropriate concept to be used for our project.

According to [1], sentiment analysis with affective computing can be performed in two manners: emotion detection or polarity detection. To put it simply, the former refers to understanding users at an emotional level, e.g., anger, joy and sadness, while the latter focuses on the binary classification of positive and negative. There is also a ternary classification where a neutral state (an absence of both positive and negative sentiment) is considered as a baseline [1]. To reduce the complexity of implementation, our research will specifically be looking into the use of sentiment orientation (polarity) for users and items with ternary classification. This alone is sufficient enough to define an entity's sentiment without the need for additional information such as intensity and/or sentiment type [1].

1.1.3 Personality-and-Affect-Aware RS

The only contrast between affect-and-personality-aware recommendation systems and their traditional counterparts is that the former adds user personality traits and emotional states/sentiments onto the user and item profile. Thus, a personality-and-affect-aware recommender produces recommendations that are tailored to a user's rating history and also their personality and sentiment. A key thing to remember is that the traditional filtering techniques, e.g., CF, CBF and hybrid filtering are still adopted in both cases [4], [17]. To put it simply, these recommender techniques would always be applied as the underlying model for generating recommendations in both context-aware and non-context-aware recommenders.

On another note,contextual information can be added to the normal recommender system flow through filtering or modelling [2], [4]. As their name suggests, the former filters the dataset or recommendations based on the user's context (in this case, personality and sentiment) while the latter incorporates them into the recommendation generation process itself [4]. The configuration depends on how the context is being used by the system. A future section will define the three stages in detail and the specific selection made for this research.

1.2 Project Aim

The overall aim of this project was to implement and evaluate a recommender system (RS) that considers user and item personalities and sentiments, aside from their ratings, when generating recommendations. The domain of focus was physical/printed books due to a research gap of similar recommenders in this field (see Section 2.1), and the strong correlation between personality and book genre preferences [5], [12], [39].

Findings from this work were used to answer the overarching research question (RQ): *"Does incorporating personality and sentiment improve the accuracy and quality of a recommender system?"*. More specifically, precision and recall were looked at for the accuracy of recommendations while diversity and level of personalisation were analysed for quality of recommendations. To investigate and answer this RQ, the following three stages of objectives had been laid out.

Starting with the basic objectives:

- Perform substantial literature review and research in order to decide on the domain of application and associated dataset that this study would revolve around.
- Determine the specific types of personality and sentiment classification with adequate justification (e.g., deciding if Big Five Personality or Myers-Briggs Type Indicator would be more suitable for this study)
- Select (through mild evaluations) an appropriate 2D recommender model that would be used as a baseline for the project.
- Implement the chosen 2D recommender model.

These tasks would effectively help clarify and solidify the foundation and direction of the research.

The intermediate objectives for the project were:

- Review and select the personality and sentiment acquisition approach for both user and item to build their respective "profiles" in the RS. This involves exploring both manual and automated techniques such as questionnaires or performing automated sentiment analysis.
- Decide on the method of incorporating personality and sentiment (from either context filtering or modelling) to finalise the architecture of the RS.
- Implement and evaluate a fully-functional personality-and-sentiment aware model. The evaluation process and result analysis are to encapsulate the metrics highlighted in the main RQ.
- Write up a report documenting the research performed, methodologies and findings of this study to allow future expansions.

Given time, a few advanced objectives that would be tackled were:

- Conduct an online evaluation, i.e. user study to obtain primary data in the form of live feedback on the performance of the recommender system and the user's experience.
- Validate and test the chosen personality and sentiment acquisition methods against more advanced or state-of-the-art methods to justify the correctness of the user and item profiles.
- Investigate the possibility and impact of using various configurations of the chosen personality model (if suitable) through additional evaluations. As an example, different combinations of the Big Five model, OCEAN, can be used to build a user's profile which was assumed to impact the performance of the RS.

1.3 Paper Overview

Section 1 has provided a brief look at the overall context of the project and the key concepts involved. In the next section, Section 2, related works involving personality and/or affect aware recommender systems and potential context measurement/collection instruments were discussed. Some of these works were evaluated and used to further justify the motivation for this project. Furthermore, Section 3 discusses in detail our solution and specific approaches taken to implement the personality and sentiment-aware recommender system. This is followed by Sections 4 and 5 which visualise and provide reasoning for the results and findings of our offline experiments. Additionally, the latter will also cover the achievements and limitations of our model, as well as its potential future improvements/work for expansion of research. Lastly, Section 6 will summarise the key findings and contributions of the paper.

2 Related Work

2.1 Similar Book Recommenders

This subsection mainly reviews past research and implementations of book recommenders, most of which have been found to not have incorporated any of the modal context used in this research. Research into RS for books is still ongoing, with the latest substantial papers released about one year ago. To name a few, [15], [30], [41] were recent studies that used Collaborative Filtering (CF), both memory and model-based, or an enhanced CF approach to improve the book recommendation process. Older studies such as [46] and [35] showed higher quantities of research into Content-based Filtering (CBF) for books or some hybrid methods between CBF and CF. This suggests that CF is gaining in popularity among book recommenders.

Another point to note is that a survey of past book recommenders [5] had highlighted that there was no substantial piece of work that explored the use of psychological factors such as a user's mood to improve recommendations. This is partially true based on our investigation as it appears that there are no known sentiment-aware recommenders for books; conversely, there have only been two personality-aware book recommenders. The first was SuggestABook [11], which analysed users' Facebook profiles (i.e. their interests, likes and books that they have read) to predict their personalities. Additionally, they had also mapped personalities to book genres to build the book profiles. While evaluation had been performed on their CF and CBF implementation, the results were not mentioned in the study. Moreover, the paper [11] failed to provide any details on the employed personality model as well as the specific methodology used to incorporate the user and item personalities into the recommendation generation process.

On the other hand, the solutions and findings of Hariadi and Nurjanah [22] were significantly better documented. Their approach was to improve on an existing hybrid attribute-based recommender model called Most Similar Visited Material to the most Similar Learner (MSV-MSL). It essentially looked at both the interpersonal interest similarity (which was considered a form of personality) and interpersonal rating behaviour similarity to predict the rating score of a book for a user [22]. Hariadi and Nurjanah

[22] had extended this to incorporate personality similarity by creating a neighbourhood for a user based on scores calculated from Collaborative Filtering and two personality factors, interests and rating behaviour similarities. The concept behind incorporating personality into the rating prediction technique is notable and would serve as a potential method of implementation for this research. Moreover, the fact that a more quantitative and statistical model like Big Five Personality was not utilised in both these studies justifies our research to fill in the gap for this RS domain.

2.2 Personality-Aware or Affect-Aware Models

To compensate for the lack of modal-context-aware recommenders in the book domain, this subsection will explore the use of personality and affect in other recommender domains. This would allow us to get a better insight into the possible techniques available to incorporate personality or affect into the recommendation process.

Nalmpantis and Tjortjis [38] experimented with a personality movie recommender system with two different weightings between the user's personality traits and their movie interests: (i) 50/50 and (ii) 80/20. The base system that they had used was CF. They incorporated the Big Five Personality test while using the user's movie reviews to determine their taste and interests. From the results of their user study, the 50/50 personality recommender system was found to be the top preferred model, outperforming K-NN and the 80/20 model. One of the gaps in research mentioned was that more metadata (e.g., gender, age, etc.) could be incorporated in the future to group users to make more accurate predictions [38]. Balakrishnan and Arabi [9], on the other hand, had introduced a hybrid personality-aware movie recommender which involved Big Five Personality traits and the user's demographic information (e.g., age, gender, etc). The model used a hybrid implementation of demographic filtering (DF), CBF, CF and personality. According to [9], this model had potential to be expanded to consider emotions as there was evidence that showed having more personalised information will lead to better recommendations.

Tkalčič et al. [54] have implemented a personality-aware image recommender system that incorporates the Big Five Personality and a CF algorithm. These images were sourced from the IAPS database and were meant to invoke certain emotions in a viewer. The personality traits were acquired through the IPIP questionnaire while a user study was performed to gather preferences on 70 images from the same database [54]. The paper illustrated the use of three different user similarity measures to evaluate CF, rating-based (baseline), Euclidean big five based, and weighted Euclidean big five based. It was found that the techniques involving personality had better recall and F1-measure than the baseline but were similar in terms of precision.

Similarly, Elahi et al. [18] implemented a "places of interest" mobile recommender system for the Alto-Adige region in Italy. The model used the TIPI questionnaire to gather the Big Five Personality from users upon registration which is then fed into a matrix factorization algorithm for contextual pre-filtering [18]. Their evaluations showed that the personality-based model had the lowest mean absolute error against three non-context-aware models.

In comparison to most recommendation system implementations reviewed thus far, the emotion-aware news article recommendation by Mizgajski and Morzy [36] was the most complex and of industrial-grade. They gathered millions of user impressions from three different news sites using a self-assessment widget which defines user emotions towards each article; they also implemented knowledge-based and CF algorithms along with their own affective user similarity formula in Elasticsearch. While Mizgajski and Morzy [36] did not run any offline evaluations, their live traffic evaluation had yielded the finding that making recommendations by targeting selected emotional reactions had improved click-through rates.

Lastly, Wang et al. [59] combined user emotion with two other contextual information: listening location and listening time, to build the user profile for their music recommendation system. They adopted Meyers' emotional content-based classification method to determine the emotion of the song; All Music Guides was used to classify the listening location while a concept hierarchy was built to link the songs to listening time [59]. As a result, the modal had scored the lowest mean absolute user error and the highest classification accuracy (Precision, Recall and F1 Score) values against non-context-aware models.

2.3 Personality and Sentiment-aware Models

While affect and personality computing could be seen as proliferating in the recommender system domain, we have found minimal research or implementations that incorporated both affect and personality awareness. Ishanka and Yukawa [25] were the only study found to have accomplished so in a proper manner while the vast majority of previous studies did not even suggest the fusion of two modal contexts in their future works.

Ishanka and Yukawa [25] had incorporated both personality and emotion into their recommender model, with a specific focus on the travel destination domain. All emotion and personality traits were inferred from user tweets to build the user profile [25]. Ishanka and Yukawa [25] used sentiment classification technique to derive the user emotions (Plutchik's eight emotions and sentiment polarity) while APR was used to extract (Big Five) personality traits from the tweets. The paper stated that Collaborative Filtering technique was used for the recommendation process, applying Pearson correlation as a similarity measure to find the relationship between each user and item profile [25]. Their model had been shown to outperform a non-context-aware model by about 2% in precision and 8% in mean average precision, illustrating a positive impact in classification accuracy with the incorporation of the two modal contexts.

2.4 Past Methods to Acquire Context

Aside from past implementations, we also investigated the various methods used by the studies above to acquire personality and sentiment. This subsection will consider these approaches and discuss their findings.

2.4.1 Personality

In terms of personality, our research has shown that the two primary methods of obtaining this modal context were questionnaires and APR. A prominent questionnaire adopted was the Big-Five Inventory - 10 (BFI-10) that was used in [6], [9] and [54] to classify user personality data. This was because it was a short quiz, making it more convenient for processing during the implementation phase [17]. Similarly, the Ten-Item Personality Inventory (TIPI) was also seen to be widely favoured in [18], [27] for its short length and high usability. By using either of these questionnaires, the associated studies have seen a noteworthy improvement in both error rate (e.g., Mean Squared Error) and classification accuracy of their models. On the other hand, a survey on personality-aware recommender systems [17] had highlighted that longer questionnaires do exist as well, such as NEO-Personality-Inventory Revised (240 items) and BFI with 44 items. These are seldom preferred in this research field as it was assumed that users were likely to get bored and may not provide truthful/correct answers [17].

Alternatively, research in APR was shown to proliferate in three domains: text, behaviour and multimedia-based [17]. Only the first two were explored in the following as they were relevant to our research. Referring to the views of Vinciarelli and Mohammadi [56], most APR methods tend to adopt a lexicon approach with one of the prominent tools being Linguistic Inquiry and Word Count (LIWC) as they have found a significant correlation between its categories and Big Five Personality traits.

The Personality Recogniser was a popular openly available Java command-line interface that was implemented based on the research conducted by Mairesse et al [33]. The study had performed rigorous and diverse experimentation on classification and ranking approaches for APR using a number of machine learning models, e.g., SVM, Linear Regression, etc [33]. Above all, their results also suggest the combined use of LIWC and MRC has produced good results too against the baseline. Hence, these findings and results from Mairesse et al. [33] proved to be useful as a comparison for models adopting a similar approach. Likewise, Carvalho and Louwerse [13] had also performed experiments to investigate the performance between two lexicon-based approaches, top-down (LIWC) and bottom-up (topic modelling), as well as a grammar-based approach for APR. Its findings compellingly show that these four models are no better than one another and hence justify that any of the proposed methods would be sufficient and reliable for personality prediction.

Furthermore, Yang and Huang [61] created a game recommender system that performed text mining on user messages and reviews to identify their Big Five personality traits. In terms of calculations, they had followed the findings of another study on M5 Regression trees, which was shown to produce the best results, and used categories from the LIWC dictionary [61]. Results from their online evaluation showed that their personality-aware model achieved a higher average score. Additionally, the aforementioned travel destination recommender [25] had leveraged an APR tool called IBM Watson Personality Traits Service which automated the process of calculating personality trait scores. This service is unfortunately not available anymore but did enable Ishanka and Yukawa [25] to produce promising results.

Finally, behaviour-based APR refers to relating analyzed behavioural patterns with dominant traits [17]. As evidence, Ng et al. [39] analyzed and proved the prominent relationship between book preferences with a user's personality by using user-generated book tags from a famous book website and the user's personality from a user's Facebook activity (browsing activity, likes, etc). Cantador et al. [12] had carried out a similar study on a wider scope, covering user genre preferences in movies, TV, music and books. It leveraged the then-famous myPersonality dataset [51], a Facebook psychometric test application, which allowed the research to produce substantial findings that showed correlation scores.

2.4.2 Sentiment

We have found that more recommender systems tend to opt for emotion as a context in contrast to just sentiment orientation. This was likely due to the fact that emotions consist of more specific categories [1] which was assumed to better improve the performance of a recommender. Considering that sentiment orientation and emotion are classified under the same term, ”sentiment”, [1] we explore the following.

The most popular method of inferring sentiment in this field was discovered to be sentiment analysis, especially on text-based content. Ishanka and Yukawa [25] had utilised the sentiment lexicon created by the National Research Council of Canada (NRC) to extract user emotion from their tweets. Qian et al. [45] calculated the sentiment values using a non-supervisory learning method (according to SentiWordNet 3.0) on user reviews; this resulted in a large improvement of error rate and classification accuracy against the non-context-aware baselines. Moreover, research on analysing movie emotions and sentiment polarity from textual content [28] had utilised NRC, AFINN and ERO in its implementation. Comparing their results against IMBb had produced notable findings which suggested a strong correlation between sentiments and some emotions towards movies. Despite promising results, many of these datasets are considered knowledge-based techniques and they can fail in understanding nuances during semantic inference as the data is fixed and flat [1].

Leung et al. [32] took a step further and built a ”Tweets Affective Classifier” model which adopted a novel model-based approach to analyze and extract emotions from tweet messages for a movie recommender system. While it was found that their model did tailor recommendations to a user's emotions and preferences, no quantitative offline evaluations were performed to justify its performance. Referring to the views of [1], statistical (e.g., support vector machines) and deep learning models are the most prominent approaches today as systems trained on a large corpus can learn the polarity of affect keywords and the other keywords around it (word co-occurrences).

On the other hand, there were studies [8], [52] that leveraged sensors and visual capturing devices such as video cameras to implicitly detect a user's sentiment. This was in view that many people tend to mask their true emotions by not showing obvious reactions, for instance, facial expressions [8]. The former study [52] had found that detecting emotions from spontaneous face expression videos (i.e., emotions were not forced) resulted in their recommender's low classification accuracy. Conversely, the model that inferred emotions through physiological signals [8] showed high accuracy. This contradiction in findings illustrates the complexity of performing multimedia sentiment analysis.

2.5 Summary

Overall, all analyses on personality and/or affect-aware recommenders by the surveyed literature have produced positive results. This suggests that these two modal contexts do improve a recommender's performance, namely prediction and classification accuracy. Moreover, the fact that there has only been one substantial paper on a recommender system that combines both contexts [25], is evident of the research gap in the field. This is further justified by the future work of Balakrishnan and Arabi [9], and Zheng et al. [63] that recommended further analysis on the correlation of different contextual factors, such as personality and sentiment, and their association with other data like user ratings, item features, etc.

Furthermore, the book domain had also been found to lack proper adoption of modal contexts in its recommendations despite the advancements in personality and sentiment acquisition methods seen above. Therefore, in consideration of all the points highlighted, this project aims to fill the research gap by implementing a personality and sentiment-aware book recommender system. Aside from prediction and classification accuracy, this system would also be evaluated on diversity and level of personalisation (metrics that were not mentioned in any reviewed literature) as it is generally understood that good accuracy does not guarantee a satisfying user experience [4].

3 Methodology

To address the research questions, this section discusses and evaluates possible methods to implement a personality and sentiment-aware recommender system that recommends books. The following subsections will detail the dataset, considered CARS schemes, 2D recommender models, and methods to incorporate personality and sentiment into the recommendation process. Figure 1 illustrates the overall top-down architecture of the implemented system, focusing on the crucial data and main processes involved.

In a nutshell, our model leverages a context-independent, model-based Collaborative Filtering technique (matrix factorisation) to first generate a set of recommendations before it is further personalised (i.e., re-ranked and filtered) to a user's personality and sentiment. This was the primary reason why the profiles of users and books were observed separately (as seen at the top of the diagram). Likewise, this was also in view of the research carried out by [12], [34], [40], etc that personality and sentiment correlate to the book types that a user might prefer/select. The variable N seen in the final block at the bottom refers to any number (e.g., 5, 10, etc.) of recommendations.

Taking a closer look, it can be noted that personality and sentiment acquisition for books and users all differ in methods. This was an intentional implementation design following past substantial research [12], [33], [44] that justify the suitability of these methods on other models similar to our approach (see Section 3.3). Aside from ratings, the user profile contains the Big Five Personality scores obtained from APR, i.e, calculating based on relevant LIWC dictionary categories [62]; sentiment was asked from the user upon logging in to the system. Book profiles, in contrast, consisted of sentiment from conducting sentiment analysis, and genre personality scores based on research conducted by Cantador et al [12].

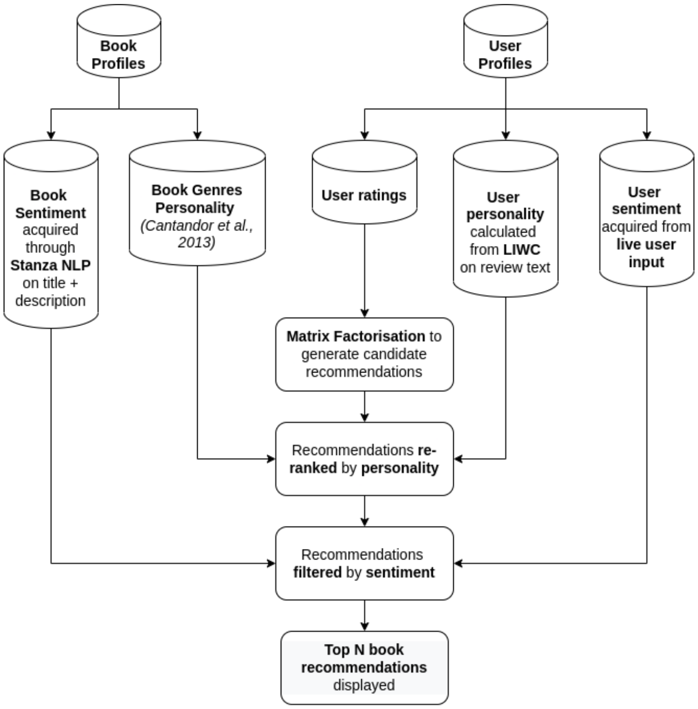

Book Profiles
User Profiles
Book Sentiment acquired through Stanza NLP on title + description
Book Genres Personality (Cantandor et al., 2013)
User ratings
User personality calculated from LIWC on review text
User sentiment acquired from live user input
Matrix Factorisation to generate candidate recommendations
Recommendations re-ranked by personality
Recommendations filtered by sentiment
Top N book recommendations displayed

Fig. 1. High-level architecture of Personality and Sentiment-aware recommender system

3.1 Dataset selection

As matrix factorization is a form of Collaborative Filtering, it thrives when there is a substantial amount of users and items in the system [4]. One of the many pitfalls that this approach tends to face is the "cold-start" problem, a common RS issue where a new user who enters the system without any previous ratings nor sufficient interactions would likely be given irrelevant or inaccurate recommendations due to the lack of shared knowledge by the system [4]. Hence, a dataset which contains a substantial amount of entries would have to be acquired for model training and evaluation before it can be deployed to ensure that it can perform up to par.

To date, there have been no datasets which contain users, items and ratings as well as the personality and sentiment of each entity. The best possible candidate dataset was Personality2018 [40], a personality-annotated rating dataset made by GroupLens. While it does contain calculated FFM scores for each user, it does have some limitations. A few of the major ones are that it: (i) contains movies instead of books (ii) does not have any sentiment data or any means to perform sentiment analysis (iii) has a fairly small user base - about 1834 users. In light of this, the focus of our dataset search shifted to popular book datasets that contained textual data which would allow the capability of performing our own APR and sentiment analysis.

It was found that the Goodreads datasets[1] [57], [58] was the most suitable as it fulfilled the dataset feature requirements for our research. It is a notable ”library” of datasets based on scraped data from users’ public shelves on the Goodreads website. For this research, three main datasets have been chosen. The first dataset consists of 1.3 million book reviews (with ratings from 0 - 5) for 25,475 books (ids only) and 18,892 users. The second and third datasets enriched the data further by providing the title and description of each book id, and set of genres for each book respectively. Information from these datasets allowed us to generate our own personality and sentiment data.

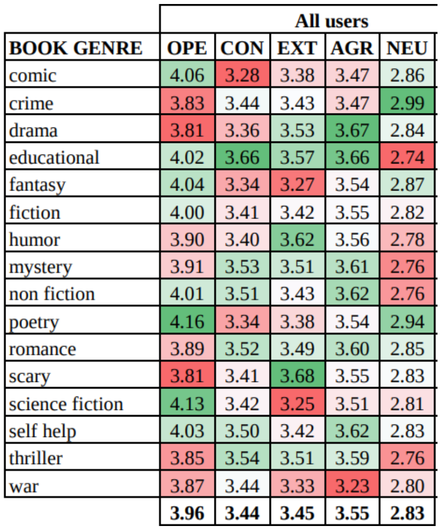
Another key point is that a personality-annotated dataset was required to create a model that would perform APR on user book reviews. The largest dataset that contained both textual data and scores for each of the Big Five Personality traits was from myPersonality [51]. However, the large burden of time and effort required to maintain the dataset eventually led to the halt of its distribution four years ago [51]. In other words, we had to implement an APR algorithm for users from scratch by using the LIWC dictionary as the only data source. Section 3.3.1 explains this process in detail and shows how the resultant user personality scores were verified as reliable synthetic data for our model.

Lastly, a table consisting of book genres and their respective FFM scores (see Figure 2), was also utilised for context filtering in this research. This data is from [12] which illustrated how user personality was related to genre preference. They had based their research on the myPersonality dataset and were able to analyze a significant amount of Facebook user profiles and their entertainment interests, resulting in their findings of a meaningful correlation between the Big Five Personality traits and a user’s preference in multiple domains, one of which was books. Section 3.3.2 will go into more detail about the usage of this table of data.

3.2 Data Pre-processing

As pointed out in the preceding section, there were three main datasets from Goodreads selected for this research. Through pre-processing, the goal was to clean up the reviews file and build the user and book profiles from these datasets. It was found that the average number of ratings per user was 73 with a median of 36. Hence, users who had less than 20 ratings were removed to ensure sufficient data points for the model to learn each remaining user. As a result, the risk of the model experiencing the ”cold-start” problem when performing training and evaluations would likely be mitigated. This removal process brought the user base down from *18,892* to *12,611* and the number of reviews + ratings from *1.5M* to *1.3M*.

For user profiles, the only pre-processing task required was to concatenate all review texts (with stopwords removed) of each user as a single ”document” per user. This was carried out for our APR algorithm to have a substantial amount of information on each user to calculate FFM scores reliably and accurately. No preparations were needed for user sentiment seeing as it was collected live from the user (see Figure 1). Furthermore, as the LIWC Python module only uses the shorthand notation of the LIWC categories, additional information had to be scraped from the LIWC 2015 manual [42] which shows both the full and abbreviated category names, to conduct mapping of values during APR.



	All users				
BOOK GENRE	OPE	CON	EXT	AGR	NEU
comic	4.06	3.28	3.38	3.47	2.86
crime	3.83	3.44	3.43	3.47	2.99
drama	3.81	3.36	3.53	3.67	2.84
educational	4.02	3.66	3.57	3.66	2.74
fantasy	4.04	3.34	3.27	3.54	2.87
fiction	4.00	3.41	3.42	3.55	2.82
humor	3.90	3.40	3.62	3.56	2.78
mystery	3.91	3.53	3.51	3.61	2.76
non fiction	4.01	3.51	3.43	3.62	2.76
poetry	4.16	3.34	3.38	3.54	2.94
romance	3.89	3.52	3.49	3.60	2.85
scary	3.81	3.41	3.68	3.55	2.83
science fiction	4.13	3.42	3.25	3.51	2.81
self help	4.03	3.50	3.42	3.62	2.83
thriller	3.85	3.54	3.51	3.59	2.76
war	3.87	3.44	3.33	3.23	2.80
	3.96	**3.44**	**3.45**	**3.55**	**2.83**

Fig. 2. Book Genres x FFM Scores Table by Cantador et al. [12]

Furthermore, all book IDs in the preprocessed rating/review dataset were combined with their associated title, genres and description from the other datasets. An additional pre-processing step executed here was to convert the format of how book genres were stored. Table 2 illustrates this change. Dictionary keys that were combination(s) of genres were split as some of these individual genres were also used by [12] to produce Figure 2. To exemplify, a common genre tag used in the dataset was ”mystery, thriller and crime” yet it is clear in Figure 2 that these three genres have separate sets of personality trait scores; hence the importance of splitting them up.

TABLE 2
String format of genres before and after preprocessing

Before Pre-processing	After Pre-processing
{”fantasy, paranormal”: 31, ”fiction”:8, ”mystery, thriller, crime”: 1, ”poetry”: 1}	”fantasy|paranormal|fiction |mystery|thriller|crime|poetry”
{”fantasy, paranormal”: 12, ”young-adult”:8, ”fiction”: 1, ”children”: 16}	”fantasy|paranormal| young-adult|fiction|children”

3.3 Context Acquisition

This subsection focuses on the methods used to infer personality and sentiment for users and books based on the preprocessed data.

[1] https://tinyurl.com/goodreads-dataset

3.3.1 User Personality

Following the classifications given in Section 1.1.1, text-based APR was the best choice to infer user personalities for this research as the intention was to perform analysis on user reviews rather than carrying out an initial user survey to collect primary data. This was coupled with the fact that text-based APR tends to have the highest accuracy in predicting personality traits [17]. LIWC is a text-analysis program (and offline dictionary) widely used in academia (refer to Section 2.4.1) that classifies texts into psychologically relevant sets [17] which can be used to predict personality traits based on term frequency in each set. For this reason, we have leveraged it in our APR implementation.

The research conducted by Yarkoni [62] formed the basis of our APR algorithm. He performed a large-scale exploratory analysis of Big Five personalities on the word usage in numerous blogs to investigate the correlation(s) that exist between the two. The paper [62] elaborates on the use of the LIWC 2001 dataset to perform a category and word-based analysis of each blogger's work. For our research, only the result for the category-based analysis on LIWC's 66 categories was considered as this approach had provided beneficial results for other past studies as well [13], [33], [43]. Table 3 shows a snippet of the results from Yarkoni [62].

TABLE 3
Abstract of the findings From Yarkoni [62]

LIWC Category	NEU	EXT	OPE	AGR	CON
Total pronouns	0.06	0.06	-0.21	0.11	-0.02
1st pers singular	0.12	0.01	-0.16	0.05	0
1st pers plural	-0.07	0.11	-0.1	0.18	0.03
1st person	0.1	0.03	-0.19	0.08	0.02
2nd person	-0.15	0.16	-0.12	0.08	0

It should be highlighted that the LIWC dictionary used for this research is the 2015 version, however, this was not an issue as majority of the categories remained the same over the years. The APR algorithm accounts for this discrepancy by only considering categories that appear in both results and the acquired dictionary [62]. Detailed below is a novel yet straightforward approach of calculating user's FFM scores that was created for this research.

1) The LIWC Python library was used to parse the dictionary and perform frequency counting for each category on each combined user review.
2) The score for each LIWC category was calculated by dividing each category's frequency by the total number of words in the document [43], [62] - this was presumed to be a form of normalisation.
3) Thirdly, we proposed the following step of multiplying each normalised category frequency (see Table 3) with its respective correlation weight for each of the five personality traits. Though not mentioned in [43], [62], this was thought to be a logical step to take given the diverse correlation scores between LIWC categories to personality traits. To elucidate, the "affect" category has a positive correlation of 0.07 to Neuroticism, in contrast to Conscientiousness which has a correlation of -0.06.
4) Then, all values of each LIWC category were summed up for each personality trait to get the FFM score for a user.
5) Values were scaled between one and five to standardize with the scoring range used by Cantador et al. [12]. This was achieved via interpolation of values per personality trait as the assumption made was that the scores of each trait were independent of each other.

To justify the correctness and reliability of this new APR approach, we evaluated it on its accuracy score, precision and recall. Model-based approaches (e.g., SVM, Naive-Bayes, etc) from past research [13], [33] mentioned in Section 2.4.1 were also evaluated on these specific metrics and had shown to produce promising results. Therefore, the evaluation results of our APR approach would be compared against the aforementioned advanced approaches.

For accuracy, the Essay corpus by [43] was used as it traditionally has been known to be the best benchmark for validating APR-based methods [13]. This corpus is a collection of student essays and the respective student's Big Five Personality trait "intensities", represented as binary values (high/low). Precision and recall, on the other hand, were evaluated based on comparing the predicted favourite genres of a user from the Goodreads dataset against their actual genre preferences. The evaluation and result analysis can be found in Section 4.1 where our novel APR approach is shown to hold up against some of the model-based methods. As a result, the user personalities generated for this study were considered reliable and would not significantly skew the results of our personality and sentiment-aware recommender model.

3.3.2 Book Personality

On the contrary, personalities of books were not required to be inferred through our APR method. Instead, their genre(s) were used as a way of generalising this context data since a book's title and description did not constitute enough text (unlike the combination of user reviews) to reliably infer each personality trait. The concept of generalised context-filtering was introduced by Adomavicius et al. [3] to address this 'sparsity' problem and was first used in pre-filtering. Despite that, Adomavicius et al. [3], [4] have mentioned that this concept could easily be extended for post-filtering methods too.

As a result, the table of data (Table 2) from Cantado [12] was selected as a method of assigning quantitative values for each personality trait to each genre. The specific manner in which these table values are used will soon be made clear in the section about context post-filtering. This form of multimedia APR was considered suitable for our study since books are a form of media and the use of results from a well-cited study allowed this research's experiments to involve more reliable data.

As a brief overview, the table consists of 15 book genres and the respective five personality trait scores. Each row of the table is considered a vector and each cell value can range from one to five [12], [38]. These cell values are of an average of the users who have liked the corresponding genres [38]. Furthermore, the cell colouring is dictated by the degree of each score, i.e. the higher the value, the greener a cell is; the lower the value, the redder the cell becomes [12], [38]. Admittedly, the study had also produced results for each gender which would have provided additional granularity to improve our recommender system. However, we were not able to leverage this data as the Goodreads dataset lacked gender/sex features.

3.3.3 User Sentiment

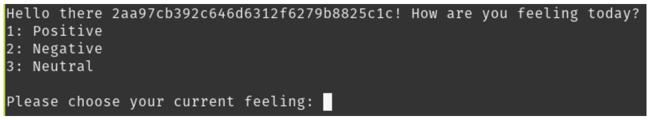
For sentiment, the system will be using the three main stages: negative, neutral and positive to create the user and book profiles.

Due to the dynamism of user sentiments, [53] and [52] have both created a framework that visualizes the role of emotions in three stages of a user's interaction with a recommender system: the entry stage, consumption stage and exit stage. The entry stage is when a user just begins using the recommender system, this would be considered the entry mood where the cause of this current mood is unbeknownst to the system [52]. The consumption stage, on the other hand, reflects the affective state(s) that the user experiences as they consume a recommended item while the exit stage observes the state of the user upon completion of said item consumption [52].

We had decided that our implementation would collect a user's sentiment at the entry stage, i.e. when a user begins using the system. The entry stage of sentiment consumption worked best for this model as its primary aim is to generate book recommendations for users to help manage their mood as soon as they begin using the system. This is achieved via a contextual post-filtering approach which is addressed in section 3.6.2. To that end, having the system collect a user's sentiment while they are reading a book from the system's recommendations or when they leave the system defeats the aim of our implementation. Moreover, according to Tkalvcivc et al. [53], retrieving emotions/sentiments during the consumption stage is a domain that has mainly been researched for implicit affective tagging of items to generate richer metadata for a dataset rather than to build a user profile. Admittedly, the collection of exit sentiment of users would be useful as a form of system feedback to evaluate their satisfaction [52], [53] with the recommendations made. This conversely would require long periods of user study to be performed to gain relevant results and thus was not considered for this research.

The collection of user sentiment was achieved through a Three-Point Likert Scale question, the three points being positive, negative and neutral. This form of live input was required in contrast to performing sentiment analysis on users' past reviews due to the dynamism of human emotions [1]. An illustration of this process can be viewed in Figure 3. Unlike the typical psychometric questionnaires, e.g., TIPI for Big Five Personality, that use five or seven-point scales [17], [49], only three points were required to gather user sentiment as that was the number of sentiment orientations. This follows the views of Toor [55] where an appropriate number of scales is when each option is able to differentiate itself ”as much as validly possible”. Moreover, the decision to include a neutral option is also justified by Toor [55] that suggested middle alternatives should always be provided as users who choose this option are not always avoiding the question or do not know.

The sentiment obtained from this simple question will only be temporarily stored in the system until the user logs out. In other words, once the user enters the system again, the same question will be prompted once more as the system does not assume the same sentiment is kept regardless of the time between each user's session(s).



```
Hello there 2aa97cb392c646d6312f6279b8825c1c! How are you feeling today?
1: Positive
2: Negative
3: Neutral

Please choose your current feeling: █
```

Fig. 3. Asking a user during log in for their sentiment

3.3.4 Book Sentiment

Highlighted in [53] and section 1.1.2 of this paper, the rapid advancements in affective computing research has enabled several automated emotion and sentiment detection techniques to be created. Since the books in the Goodreads dataset did not come with any relevant sentiment labels or tags, we decided to perform an analysis of the books' title and description/summary to build their sentiment profiles. The system utilises the Stanza Python library, a popular natural language processing (NLP) package that is built with highly accurate neural network pipelines to efficiently perform various textual analyses, for this task [44]. Other options considered were the NLTK library, NRC Lexicon and creating a model from scratch - the final option was to only be considered if time permitted. Nevertheless, Stanza was considered the most suited as it had been well-trained with a handful of known data sources, namely the Stanford Sentiment Treebank and Sentiment Labelled Sentences Dataset that covers reviews from popular websites such as Yelp and Amazon [44]. Moreover, manually checking the predicted sentiments from each of the three Python libraries on several books also concluded that Stanza was the most accurate model.

The model assigns negative sentiment as 0, neutral as 1, and positive as 2 [44]. Diagram 4 illustrates the percentage of the 25,475 books that were classified under each sentiment polarity. It is evident that there supposedly exists a disproportionate number of neutral-sentiment books in the dataset, having more than both negative and positive books combined. Manual inspection of each book's sentiment does suggest that were numerous books that were not correctly predicted in terms of their sentiment - this likely stemmed from the fact that the sentiment analysis model is not able to understand the underlying tone/nuance of the book descriptions and title. Section 5.2 dives further into this issue.

3.4 CARS Scheme

As mentioned in Section 1.1.3, there are three main ways that personality and sentiment can be included and affect the recommendations process: pre-filtering, post-filtering and contextual modelling [4]. Context pre-filtering is where the contextual data is used as input to drive data selection or

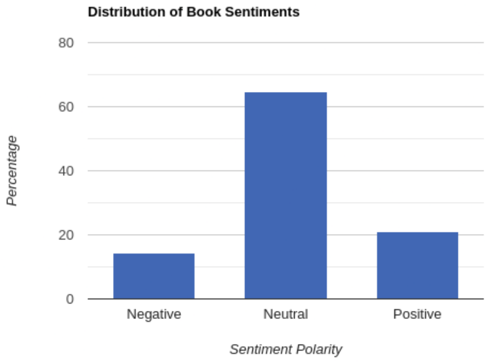


Fig. 4. Distribution of book sentiments

construction [4], [23]. The user ratings are then predicted using an underlying 2D recommender. In contrast, contextual post-filtering is the reverse as contextual data is only considered after a set of recommendations (typically known as the candidate recommendations) are produced from the 2D recommender [4], [23]. Lastly, contextual modelling refers to the contextual information being used directly in the modelling technique to predict the ratings [4], [23].

For this study, we have opted for contextual post-filtering where the specific process applied is detailed in Section 3.6. To determine the advantage of choosing post-filtering over pre-filtering, we performed some experimentation and comparison on these two filtering approaches.

Unlike in post-filtering, we have opted to filter books by the top five genres of a user based on their personality scores for pre-filtering as [25] did not detail their approach despite mentioning the use of personality pre-filtering. It was found that, for 100 randomly selected users, pre-filtering results in a 74% loss of quantity of ratings and books in the dataset even before it is fed into our 2D recommender. Breaking it down further, a little over 50% of the loss derived from filtering by sentiment (a later section will explain this). While post-filtering does also filter books by sentiment, it was only carried out after candidate recommendations were generated, meaning the underlying 2D recommender had a substantial amount of data to make more accurate rating predictions at the start. Coupled with the fact that a wider selection of books could be suggested since personality was only used to re-rank recommendations, post-filtering was deemed to be the more suitable option of the two context filtering methods.

A contextual modelling approach was not considered for this project due to time constraints. To elaborate, as we are using both personality and sentiment for this research, there would be a total of eight additional features that would have to be included in the rating prediction algorithm. To the best of our knowledge, this was a very novel approach as there has been no research on contextual modelling for these two context data and thus, longer periods of experimentation and research would be required. Additionally, related studies ([25], [38]) have only used either of the context filtering approaches as well.

3.5 Collaborative Filtering

The underlying 2D recommender model for our implementation uses a Collaborative Filtering (CF) technique. It is used to leverage the ratings of other users with similar book interests to the active user to predict ratings for unrated books [4], [38]. This would allow a more diverse content recommendation to expand users' interests which would encourage their loyalty to the system. Our literary research conducted in Section 2.1 indicated that more recent works in book recommenders have opted for Collaborative Filtering methods due to the promising results from their evaluations. Moreover, two of the recommenders which incorporated personality [11], [22] had chosen Collaborative Filtering, similar to personality-aware models from other domains [25], [38]. Considering that no papers compared and assessed the performance of two or three of the recommendation approaches, CF was selected for our implementation due to its popularity in similar research fields.

CF is typically implemented via a memory-based (i.e., user- or item-based) or a model-based technique (e.g., matrix factorisation) [4], [11]. While relevant studies [11], [22], [38] have used a memory-based approach, such as K-Nearest Neighbour, recent works by Chertov et al. [16] and Kovačević et al. [30] have shown that a model-based approach like SVD or SVD++ does produce better results in terms of error rate and accuracy in the book domain. Therefore, we evaluated popular CF models such as SVD, SVD++ and Neural Collaborative Filtering (NCF) [4] on rating prediction accuracy to find the best performing one. Content-based Filtering (CBF) with TF-IDF was also evaluated as a baseline comparison as CBF methods were first widely used in book recommender systems (refer to Section 2.1). The results, detailed in Section 4.3.1, show that SVD was most suited for our chosen dataset.

SVD (a type of matrix factorisation) is a state-of-the-art approach that aims to map users and books on a shared latent factor space and tries to predict missing ratings in a user-item matrix by categorizing them based on various factors from inferred user feedback [4]. Moreover, Adomavicius et al. [4] had also highlighted the fact that SVDs are commonly used due to their scalability and accuracy, especially for sparse matrices - like the one used in this implementation. For our research, we have made use of a pre-built SVD model from Surprise, a famous Python library which provides a handful of ready-to-use RS models (built based on respectable sources) and dataset handling tools [24]. This is useful for our research as our main focus is on the impact of incorporating personality and sentiment, and not the implementation of the 2D recommender.

The SVD algorithm was popularised by Simon Funk when he won third place in the Netflix Prize [24], [29]. To get a predicted rating of an item (in our case, a book) for user u, the following equation, Equation 1, was used [29]. Here, μ represents the overall mean rating of all users while $q_i^T p_u$ is a dot product (of the item vector and user vector) which captures the overall interest of the user u in item features. According to Koren and Bell [29], user and item biases frequently occur in CF data, for example, a user may tend to rate items higher than others, hence the SVD approach takes this into account by including these biases into the rating prediction, b_u and b_i.

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u \tag{1}$$

To learn the value of the biases, and user and item vectors, the following regularised square error is minimised [24], [29]:

$$\sum_{r_{ui} \epsilon R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda(b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2) \tag{2}$$

With minimisation, a basic stochastic gradient descent (SGD) algorithm was put in place which loops through all training data and generates training cases by calculating a predicted rating and the error value, i.e difference between the actual and predicted rating. The unknown parameters can then be modified ”by moving in the opposite direction of the gradient” [29].

In terms of hyperparameter tuning, a randomised search cross-validation approach [24] was performed on the model to determine the optimal number of latent factors ($n\_factors$), training epochs (n_epochs), learning rate (γ) and regularisation term (λ). This was performed in-place of a grid-search based cross-validation as the former is more efficient at finding suitable values [24], which is beneficial for the SVD model as the training data used for evaluation was fairly large. As a result, the configuration found to provide the best model performance was: (i) $n\_factors = 10$ (ii) $n\_epochs = 20$ (iii) $\gamma = 0.007$ (iv) $\lambda = 0.2$

3.6 Context Incorporation

As context post-filtering was selected as our method of context incorporation, candidate recommendations generated by our SVD model had to be augmented by personality and sentiment. This subsection outlines the specific methods in which this was achieved to further personalise a user’s recommendations.

3.6.1 Personality

Contrary to the method’s name, books from the candidate recommendations are not filtered out based on an active user’s personality, they are re-ranked based on their relevance towards a user’s personality trait scores. The approach used in this implementation, called the 50-50 approach, is largely referenced from Nalmpantis and Tjortjis [38] as they had performed a user study (with 30 people, male and female included) which compellingly showed that more people preferred this approach over non-personality-aware recommender. Additionally, Nalmpantis and Tjortjis [38] was also the most substantial study to have used personality in a post-filtering setting, hence why it is referenced fairly frequently throughout this paper.

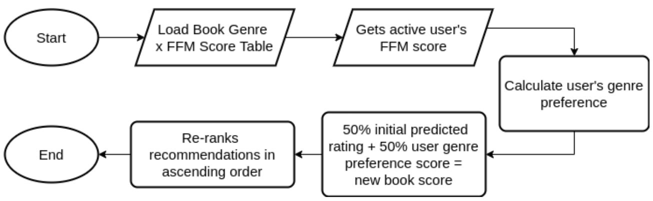

Start
Load Book Genre x FFM Score Table
Gets active user's FFM score
Calculate user's genre preference
50% initial predicted rating + 50% user genre preference score = new book score
Re-ranks recommendations in ascending order
End

Fig. 5. Post-filtering with Personality

TABLE 4
Example user FFM score

OPE	CON	EXT	AGR	NEU
3.55	3.34	1.66	2.15	2.57

TABLE 5
FFM score of Comic and Crime [12]

GENRE	OPE	CON	EXT	AGR	NEU
Comics	4.06	3.28	3.38	3.47	2.86
Mystery	3.91	3.53	3.51	3.61	2.76

The 50-50 approach is conceptualised as a flowchart in Figure 5. The system first calculates the genre preference of an active user using their FFM score and the Genre x FFM Score table shown in Figure 2. For this, a distance metric for each genre was calculated by summing the absolute difference between each personality trait score and the active user’s trait score [38]. It is trivial that a lower distance suggests that a user has a larger preference for a specific movie genre. As an example, given a user with the set of personality trait scores in Table 4 and two randomly selected genres in Table 5, it can be calculated that the distance metric of this use is 3.90 towards the comic genre and 4.05 for the mystery genre. Therefore, we can say that the user slightly prefers comics over mystery books.

Once the same distance calculation has been performed on all genres shown in Figure 2, the system will then incorporate these values into the new score (i.e. weight) calculation for each book. As Nalmpantis and Tjortjis [38] have shown that a 50/50 weighting split between the distance scores and 2D recommender predicted rating yielded a compellingly good result (through offline and user study evaluations), our model adopted this ratio and calculates the new score by totalling half of a book’s predicted rating from the SVD model and half of all its associated genres’ distance value for the active user to calculate its final weight. As not all genres from Goodreads were present in the research by [12], those not seen in Table 2 were either excluded from the calculation or classified under an existing genre (e.g., historical fiction was considered as fiction). The system finally reorders the candidate recommendations in ascending order, as distance score is part of the weighting, and returns a new list of candidate recommendations for the next step of contextual post-filtering.

3.6.2 Sentiment

Unlike personality, sentiment was used to directly filter out the candidate recommendations rather than re-weighting and re-ranking them. This step was performed only after personality had been incorporated into the recommendation process as it would remove a large number of books from the final list of recommendations. We have seen the same situation in the pre-filtering experiment too, both cases were affected by the unbalanced distribution of book sentiments (see Figure 4).

A major decision to consider with sentiment was how it would tailor a user’s experience. Referring to the views of ”Tkalčič et al. [52], Alharthi et al. [5], and Mar et al. [34], there are various approaches an RS model can take to handle the influence that an emotive state has on a user’s decision making. Two main directions that Mar et al. [34] had suggested was mood maintenance or Mood Management Theory [47]. The former relates to the efforts of assisting users in keeping their feelings the same when they are first recommended a book till the end of reading that book; the latter attempts to promote and maintain positive moods or circumvent negative moods when possible [34], [47].

It is evident from the findings of Mar et al. [34] that there still exists a large complexity in figuring out a defined method to understand the emotional reasoning behind a user’s selection of books. As an example, some users would prefer horror stories to improve their mood while others might prefer positive books like comics to achieve the same effect. Despite these nuances, the Mood Management Theory was still selected for the post-filtering process - constrained to its basic idea and concept. This was mainly due to the fact that it is not trivial to predict the preference of a user when they are in a negative sentiment - this may require a user study and a feedback algorithm for the recommender to methodically learn if a user enjoys reading negative or positive books in such situations.

In light of that, the system would provide books which are of positive sentiment to users who have a positive entry sentiment or a negative sentiment to hopefully maintain or improve that state [34], [47]; Users who are in a neutral state are suggested neutral and positive books to do both [34], [47]. Due to the nature of this filtering process, books of negative sentiment would always be filtered out before the final list of recommendations is seen by the user. Section 5.3 addresses this issue and provides a suggestion for further improvement.

4 Results

This section outlines the qualitative and quantitative results of several evaluations performed on the personality and sentiment aware recommender system (PSA-SVD). All parameter configuration, performance analysis and experiment settings will be shown. It is key to note that the SVD model was kept as the comparison baseline for every test and each test was performed at least two times to ensure correctness and reliability. Further, our model was programmed to automatically assign a random sentiment value to each user for offline testing purposes to combat the issue of requiring live user inputs - this was the best alternative to conducting a user survey. Table 6 shows the environment specifications that were used to conduct the evaluations.

TABLE 6
Evaluation environment specifications

Component	**Title of Component**
Processor	Intel® Core™ i7-10870H CPU @ 2.20GHz × 16
RAM	15.3 GB
OS	Pop!_OS 21.10

4.1 Comparison of APR Methods

As mentioned in Section 3.3.1, our APR algorithm was evaluated on the Essays corpus [43] to assess it’s accuracy classification score, i.e., the fraction of correct predictions. We replicated the method used by Pennebaker and King [43] to obtain the predicted binary trait intensities from personality scores generated by our APR algorithm. This was achieved by performing a median split across each personality trait score. Our APR technique’s, Pure-LIWC, accuracy was then calculated by comparing the set of predicted ($\hat{y}$) and actual (y) values using Equation 3. Moreover, we had also compared the performance of our model against state-of-the-art classification/prediction approaches, namely an SVM from Carvalho and Max [13], and decision tree, nearest-neighbour (NN) and Naive Bayes (NB) from Mairesse et al [33]. Table 7 shows the results of this evaluation.

$$Accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \qquad (3)$$

TABLE 7
Accuracy for each APR model in Essays dataset

	Pure-LIWC	SVM	Decision Tree	NN	NB
OPE	0.51	0.60	0.54	0.53	0.53
CON	0.53	0.55	0.51	0.52	0.56
EXT	0.53	0.55	0.54	0.50	0.54
AGR	0.54	0.55	0.51	0.52	0.54
NEU	0.54	0.57	0.54	0.53	0.60

It is evident that the accuracy of our APR approach held up against the decision tree and nearest-neighbour models, with it scoring slightly higher than both for Conscientiousness, Agreeableness and Neuroticism. Moreover, while NB and SVM are more advanced machine learning models, their accuracy was noticeably higher than Pure-LIWC only in Openness and Neuroticism respectively. Yet, for Conscientiousness, Extraversion and Agreeableness, the difference in values was only by a minimal margin (0.01 - 0.02 better than Pure-LIWC). An interesting observation made was that our model performed the worst on Openness, having an accuracy value that was the lowest across all models and personality traits; we would address this in Section 4.3.

To achieve a better indication of our APR method’s performance, precision and recall were also reported by comparing a user’s predicted and actual genre preference (as mentioned in Section 3.3.1). F1-score was not considered as it was not evaluated in the relevant research [13], [33], thus no comparison could be made. A user’s actual genre preference was obtained by calculating the frequency of each genre from books that they had liked (i.e. books that have a rating of more than the user’s average rating) and picking the top five most frequently seen genres. On the other hand, FFM scores calculated for each user were used to predict their top five most favourite genres (refer to Section 3.6.1). As [33] did not involve precision nor recall in their evaluations, the only comparison that could be made was against the SVM by [13]. Figure 6 shows the results.

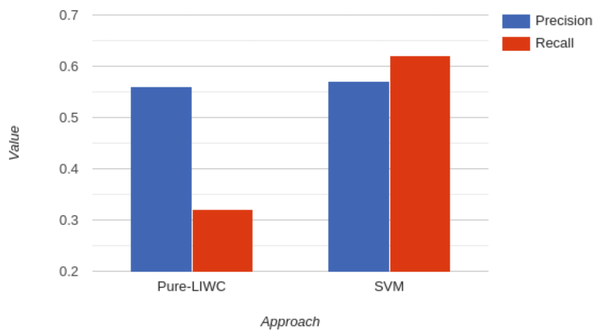


Fig. 6. Precision and recall of APR approaches

TABLE 8
Example user FFM score

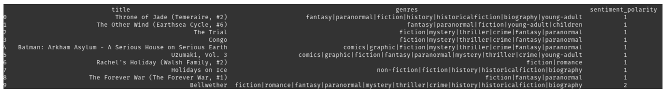
OPE	CON	EXT	AGR	NEU
3.53	3.24	1.62	2.05	2.90

As can be seen, the precision of our method is almost the same as the SVM while recall value was noticeably lower by 0.3. Section 5.2 will further touch on these findings. Nonetheless, both evaluation results compellingly show that the novel APR method used in this research, though simple, was still able to perform on par with a handful of popular prediction and classification models. Likewise, this provided more confidence to us that the inferred users' FFM scores did reflect legitimate user data which, in turn, would provide more accurate and reliable contextual post-filtering in the next steps of our model's algorithm.

4.2 Validating Personalised Recommendations

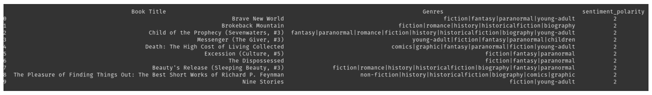
This evaluation is a form of system testing on our PSA-SVD model to ensure that the underlying SVD model and the contextual post-filtering module (for both personality and sentiment) were working properly together. This was the best alternative to carrying out an actual user study considering the timeframe of the project. To conduct this evaluation, recommendations for multiple users (before and after context filtering) were generated and reviewed individually to determine if they were sensible given the user's personality and sentiment.

As an example, a random user with 312 ratings/reviews was selected to illustrate the evaluation process. This user has the personality trait score shown in Table 8 and expressed negative sentiment upon using our system. Figures 7 and 8 will show the initial candidate recommendations generated by SVD and the final recommendations after post-filtering respectively for this user.



```
                                                      title                                                                                         genres  sentiment_polarity
0                            Throne of Jade (Temeraire, #2)                    fantasy|paranormal|fiction|history|historicalfiction|biography|young-adult                   1
1                       The Other Wind (Earthsea Cycle, #6)                                                fantasy|paranormal|fiction|young-adult|children                   1
2                                                 The Trial                                              fiction|mystery|thriller|crime|fantasy|paranormal                   1
3                                                     Congo                                              fiction|mystery|thriller|crime|fantasy|paranormal                   1
4  Batman: Arkham Asylum - A Serious House on Serious Earth                               comics|graphic|fiction|mystery|thriller|crime|fantasy|paranormal                   1
5                                           Uzumaki, Vol. 3                  comics|graphic|fiction|fantasy|paranormal|mystery|thriller|crime|young-adult                   1
6                       Rachel's Holiday (Walsh Family, #2)                                                                                fiction|romance                   1
7                                           Holidays on Ice                                        non-fiction|fiction|history|historicalfiction|biography                   1
8                     The Forever War (The Forever War, #1)                                                                     fiction|fantasy|paranormal                   1
9                                                Bellwether  fiction|romance|fantasy|paranormal|mystery|thriller|crime|history|historicalfiction|biography                   2
```

Fig. 7. Underlying SVD top 10 recommendations



```
                                                                       Book Title                                                                              Genres  sentiment_polarity
0                                                                 Brave New World                                              fiction|fantasy|paranormal|young-adult                   2
1                                                              Brokeback Mountain                                 fiction|romance|history|historicalfiction|biography                   2
2                                         Child of the Prophecy (Sevenwaters, #3)  fantasy|paranormal|romance|fiction|history|historicalfiction|biography|young-adult                   2
3                                                       Messenger (The Giver, #3)                                     young-adult|fiction|fantasy|paranormal|children                   2
4                                        Death: The High Cost of Living Collected                               comics|graphic|fantasy|paranormal|fiction|young-adult                   2
5                                                         Excession (Culture, #5)                                                          fiction|fantasy|paranormal                   2
6                                                                The Dispossessed                                                          fiction|fantasy|paranormal                   2
7                                          Beauty's Release (Sleeping Beauty, #3)              fiction|romance|history|historicalfiction|biography|fantasy|paranormal                   2
8  The Pleasure of Finding Things Out: The Best Short Works of Richard P. Feynman                      non-fiction|history|historicalfiction|biography|comics|graphic                   2
9                                                                    Nine Stories                                                                 fiction|young-adult                   2
```

Fig. 8. Personality and sentiment-aware RS top 10 recommendations

The initial difference between the two recommendations is that the books in Figure 8 are all of positive sentiment (i.e., polarity 2). This is expected due to the user's entry sentiment. Based on the user's personality, the model had predicted that his order of genre preference was *Fantasy, Comics, Crime, Fiction, Poetry, Paranormal, Romance, Thriller, Mystery* and *Non-fiction*. This is evidently reflected in Figure 8 where none of the top-10 recommendations contained less-preferred genres such as Thriller and Mystery, unlike in Figure 7. Likewise, we can see that these recommendations still retained books that are of preferred genres such as Fantasy, Comics, and Fiction.

On the other hand, an interesting observation to point out is that our 50:50 personality incorporation approach did not bias toward the frequency of genres as a book having more genres was not always highly preferred. Nevertheless, Through repeated assessment, it was found that the PSA-SVD functioned as intended in providing personalised book recommendations that considered a user's personality and entry sentiment. Thus, the correctness of our model had been validated to the best of our ability.

4.3 Model Performance evaluation

Evaluations are crucial for recommendation systems to ensure predictions produced satisfy users' expectations of relevant content and add more value and purpose for them to remain using the system [48]. One of the most commonly used measures in the research field is utility, and it can be evaluated by rating and classification accuracy based metrics [48]. In spite of that, rating accuracy could only be evaluated for the underlying 2D recommenders and not the final PSA-SVD model due to the manner of our post-filtering approach. To elucidate, while "filtering" by personality does modify the rating, it becomes a distance score where a smaller value represents a larger interest. Therefore, it cannot be used as a one-to-one comparison with the actual rating given by the user. Additionally, filtering by sentiment does not change the scores/ratings at all.

Therefore, the PSA-SVD model was only evaluated on classification accuracy for its utility. Moreover, diversity@K and personalisation@K were also evaluated to further investigate the impact of involving personality and sentiment. The variable K is used to indicate the size of a recommendation list. It should be noted that these evaluations were completed for three different configurations of the PSA-SVD model: using both personality and sentiment (PSA-SVD), using only personality (PA-SVD) and using only sentiment (SA-SVD) to provide better visibility on the roles of each context played in changing the quality and accuracy of a recommendation list. The baseline used for these evaluations was the non-context-aware SVD recommender.

4.3.1 Baseline Model Rating Accuracy

Error metrics are commonly used to evaluate prediction rating accuracy [48] as they aim to calculate the error value between an RS model's predicted rating and the user's actual rating [26]. The three most common metrics: Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Average Error (MAE) [26], [48] had been adopted to select the best performing 2D recommender for the Goodreads Dataset. This is a more in-depth look at the results aforementioned in Section 3.5. MAE is a staple for calculating the difference between a user's actual rating $r(i)$ and the predicted rating $p(i)$ [26]. Correspondingly, MSE and RMSE perform similar calculations, yet they penalise and add more penalties for inaccurate predictions. This is especially for RMSE as it penalises larger errors the most which better highlights any poor performance in the model. Accordingly, the use of these three metrics would provide a rigorous evaluation to ensure a justified model selection for our research. The respective formulae are shown below from Equation 4 to 6.

$$RMSE = \sqrt{\frac{\sum_{i \epsilon R_u} (p(i) - r(i))^2}{|R_u|}} \tag{4}$$

$$MSE = \frac{\sum_{i \epsilon R_u} (p(i) - r(i))^2}{|R_u|} \tag{5}$$

$$MAE = \frac{\sum_{i \epsilon R_u} |p(i) - r(i)|}{|R_u|} \tag{6}$$

Figure 9 displays the rating accuracy of four 2D recommenders that were considered in Section 3.5 where the two SVD models outperformed both NCF and CBF. It was also surprising that SVD and SVD++ produced the same results for all three metrics, triggering further analysis to be carried out. According to [24], the SVD++ model considers the act of rating an item (regardless of value) to be the implicit rating. As a result, this meant that the explicit rating of the user (i.e. the actual rating value given by the user) and implicit rating were not too dissimilar as choices of learning information. The results show clearly that this was a poor choice of implicit rating, at least for this dataset, causing its performance to be no better than its counterpart that relies solely on explicit ratings. Together with its long training times, it justified the option of SVD to be the primary choice of this study. On the other hand, content-based filtering with TF-IDF (CBF) showed to be the worst performing model, followed by Neural Collaborative Filtering (NCF). The former was expected as CBF tends to suffer from over-specialisation as only the ratings and preferences (in this case book genres) of the active user are considered when predicting ratings [4]. The lack of shared knowledge of this model causes it to not provide accurate ratings. Likewise, NCF only performed a little better in terms of RMSE and MAE than CBF. An assumption for this finding would be that the size of the dataset was not sufficient for such an advanced model to train on as evident by the fact that the three accuracy metrics stabilised just after three to four training epochs.

4.3.2 Classification Accuracy

Classification accuracy is a method of calculating the ratio of users to determine if they had liked the items that were recommended to them [26]. Precision and recall are two of the most common classification accuracy metrics [26], [48]. The former relates to the proportion of items that the user had liked from the given recommendation list in comparison to all the items in the same list; the latter, on the other hand, looks at the ratio of recommended items that the user has liked against their previously liked items in the whole system [26]. Equations 7 and 8 visualises the respective metric calculation where C_u is all the items liked by a user u while R_u is the user's recommendation list.

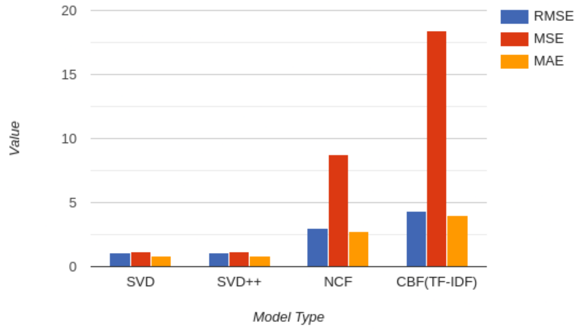


Fig. 9. Bar chart showing 2D recommender model accuracy

$$P(R_u) = \frac{|C_u \cap R_u|}{|R_u|} \tag{7}$$

$$R(R_u) = \frac{|C_u \cap R_u|}{|C_u|} \tag{8}$$

Both precision and recall were calculated for all top K recommendations, where K ϵ [5, 10, 15], for each user. The averaged precision and recall scores @K are displayed in Table 9. Moreover, to ensure correctness and consistency, the same list of baseline recommendations was used on every configuration of the context-aware model for each user.

TABLE 9
Precision and Recall @K of models. Underlined values are the highest for that column.

Model	P@5	P@10	P@15	R@5	R@10	R@15
SVD	0.909	0.893	0.884	0.389	0.597	0.705
PSA-SVD	0.364	0.464	0.510	0.242	0.320	0.354
PA-SVD	0.236	0.381	0.474	0.105	0.206	0.267
SA-SVD	0.870	0.859	0.855	0.644	0.773	0.823

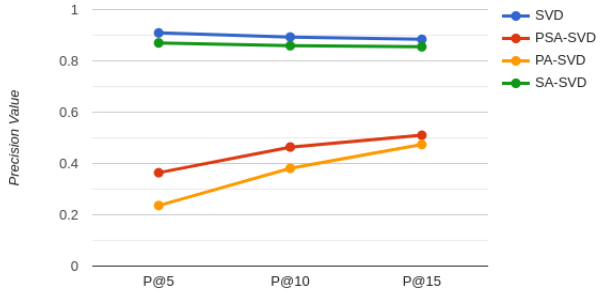


Fig. 10. Line chart displaying trend of precision@K of models

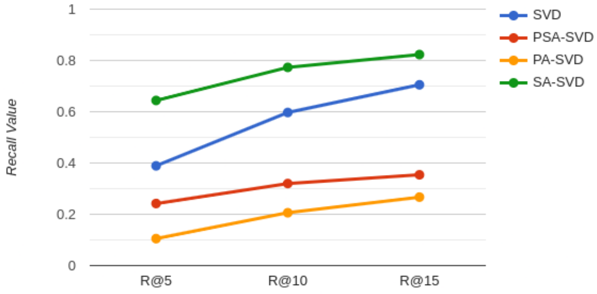


Fig. 11. Line chart displaying trend of recall@K of models

Overall, it is clear that the baseline SVD outperformed all three context-aware models in terms of precision, with SA-SVD producing the closest results with a negligible discrepancy of 0.03 - 0.04. The reason for SA-SVD's close performance stems from the fact that it only filters the recommendations and does not re-rank them in any manner. Similarly, SA-SVD is also the best performing model in terms of recall, which in turn makes it the model that classified and ranked book recommendations the most accurately. One key observation is that while the personality-based models performed poorly, their precision values were increasing as the value of K grew in contrast to SVD and SA-SVD which experienced a decline in the same situation. Hence, one could likely assume that, given a high K value, PSA-SVD or PA-SVD would show similar or even better performance than the other two models. All things considered, findings from this evaluation propose the likelihood that sentiment has the biggest positive impact on producing relevant and accurate recommendations, and that adding personality to larger recommendations can prove to be beneficial.

4.3.3 Diversity

Another metric used was diversity. i.e. how varied are the items in a recommendation list. [4] has found that users tend to have more interests in recommendations that have high diversity and it also helps to reduce model prediction error rates by attempting to cover a larger set of user interests [4], [48]. As one of our models is personality aware, using this metric would provide good insight into how it's impacted by this human-based context. We have used diversity@K which was derived from the inverse of the average intra-list similarity (ILS) of all user recommendations M, which calculates the "distance" between items i and j (using cosine similarity or cosSim) in a given recommendation list of K items. [48].

In this case, genres had been chosen as the defining feature of a book for the item-to-item comparison process, being in line with its utility in contextual generalisation. To accomplish this, every book in a recommendation list was replaced with its corresponding genre vector, generated using a one-hot encoding approach. In contrast, the two other prominent features, i.e. title and description were not suitable as they would be distinct for each book, thus skewing the results of the evaluation. Just like with classification accuracy, the same list of baseline recommendations was used on every configuration of the context-aware model for each user.

$$SLS(R_u) = \sum_{i \epsilon R_u} \sum_{j \epsilon R_u : i \neq j} cosSim(i, j) \tag{9}$$

$$ILS(M) = avg(\sum_{R_u \epsilon M} SLS(R_u)) \tag{10}$$

$$Diversity(M) = 1 - ILS_M \tag{11}$$

The diversity score was calculated for all top K recommendations, where K ϵ [5, 10, 15, 30], for each user. The averaged diversity scores @K are displayed in Table 10. Only larger K values have been considered for this metric as a recommendation list of substantial size was required to properly visualise and calculate the heterogeneity of items within a list. It would not be trivial to identify diversity in a list of one or two items.

The results show that all three context-aware RS outperformed the baseline only when K = 5, with the personality-and-sentiment aware SVD producing the most diverse recommendations for each user. The only other similar occurrence is when K = 15 where PSA-SVD and SA-SVD performed better than the regular SVD, both producing about the same high score. Otherwise, the three models had under-performed more times (7 out of 12 cases) against the baseline, especially for PA-SVD which had only beaten the baseline once. Hence, these findings suggest that either the use of personality and sentiment or the post-filtering algorithm itself had lowered the diversity of recommendations. By only comparing the context-aware models, however, it can be observed that one of the personality-aware models always performs better/similar to the purely sentiment-aware model at each K. This further suggests that recommendations made based on personality tend to be more diverse.

TABLE 10
Diversity@K of models. Highest value at each K are underlined.

Model	D@5	D@10	D@15	D@30
SVD	0.455	0.555	0.542	0.545
PSA-SVD	0.508	0.542	0.551	0.523
PA-SVD	0.490	0.550	0.516	0.537
SA-SVD	0.489	0.547	0.555	0.527

4.3.4 Personalisation

The level of personalisation of a recommendation is slightly different to its diversity; the latter measures the dissimilarity of items within a recommendation list while the former measures the dissimilarity of recommendation lists generated for each user. A simple example is that the personalisation between $[A, B, C, D]$ and $[A, X, C, D]$ would be 0.25 while it would be 1.0 between two completely different lists like $[A, B, C, D]$ and $[J, K, L, M]$.

$$Personalisation(M) = \sum_{R_u \epsilon M} \sum_{R_v \epsilon M : R_v \neq R_u} cosSim(R_u, R_v) \tag{12}$$

Equation 12 shows the calculation formula where M is a list of all user recommendations, and R_u and R_v are the individual recommendations for a user u and v. This was a useful metric for the research as it provided another way of giving visibility to the potential of personality and/or sentiment in improving the recommender's understanding of a user's preference, in turn giving more relevant recommendations. To date, there have been no known studies that have used this specific metric.

The personalisation score was calculated for all top K recommendations, where K ϵ [1, 2, 3, 4, 5, 10, 15], for each user. The averaged personalisation scores @K are displayed in Table 11. Moreover, just like classification accuracy, the same list of baseline recommendations was used on every configuration of the context-aware model for each user.

It is evident that incorporating personality and/or sentiment in our model has significantly improved the level of penalisation for each user's recommendations as all scores for PSA-SVD, PA-SVD and SA-SVD are at least eight times higher than the baseline. Looking more closely, the baseline SVD score remains stagnant after a 0.01 drop from personalisation@1 to personalisation@2 while the other three models show signs of a trend. More specifically, recommendations by the personality-and-sentiment aware SVD start to appear more personalised when there are 10 or more items in comparison to the sentiment-aware model where recommendations decrease in personalisation as the length of list increases. Likewise, the sentiment-aware SVD had consistently been providing the most personalised recommendations until K = 5, where the personality-and-sentiment SVD outperforms all the board's scores. The results shown suggest that sentiment played a larger role in improving recommendation personalisation while the additional incorporation of personality complements this well, especially at higher values of K.

TABLE 11
Personalisation@K of models. Highest value at each K are underlined.

Model	Per@1	Per@2	Per@3	Per@4	Per@5	Per@10	Per@15
SVD	0.020	0.012	0.011	0.010	0.010	0.009	0.010
PSA-SVD	0.157	0.224	0.253	0.361	0.437	0.455	0.461
PA-SVD	0.262	0.241	0.209	0.309	0.323	0.205	0.255
SA-SVD	0.451	0.445	0.445	0.446	0.453	0.399	0.361

5 Evaluation

In this section, we review the achievements, limitations, and possible expansions of the proposed personality and sentiment-aware RS implementation. Additionally, the objectives and research questions will also be reflected upon to assess the overall project's success.

5.1 Solution Achievement and Reflection

It should be first highlighted that our model had successfully been implemented and was able to generate recommendations that suit a user's personality and entry sentiment. This was clear from the evaluation performed in Section 4.2 as the recommendations generated followed the concept of mood management and prioritised books with genres more closely related to the user's Big Five Personality scores. Achieving so has also verified the correctness of our 50:50 personality incorporation approach as our model's reranking behaviour matches that of the authors in [38].

Based on the results and findings of our project, we can confidently say that personality and sentiment were only able to improve the quality of an RS from a personalisation standpoint but failed to improve its accuracy and diversity. The former part is clear from the fact that the level of personalisation for PSA-SVD was at least eight times higher than the baseline, reflecting the significant improvement. However, part of the latter point can be challenged as personality and sentiment did provide insight into possible improvements of precision and recall at larger size recommendations. Diversity, on the other hand, was not improved by the incorporation of personality and sentiment; only showing signs of slight increase from the baseline on some occasions. Given this, it is safe to say that the research question (RQ) for this project has been properly addressed.

Despite some unfavourable results, there were still some interesting insights that suggest greater potential for our system. The fact that these findings were bounded by a single approach and domain leaves room for more experimentation. Thus, we acknowledge that the PSA-SVD would serve as a stepping stone for further research on similar recommender systems that use combinations of human psychological factors too. This is because there have been no known implementations achieving a similar task. Another gap filled would be the context-aware book recommender space where minimal studies had been conducted to improve the personalisation of book recommendations (based on our literature survey in Section 2).

While not stated in the RQ, an interesting find to point out is that the sentiment-aware model, on average, is the best performing model out of the three context configurations. This mainly holds for smaller sized recommendations as personality does seem to have the potential to optimise recommendations further when used for higher amounts of recommendations. This was noted especially in classification accuracy where personality-based recommenders surprisingly show signs of precision growth as the number of recommendations increased in contrast to the baseline and sentiment-aware model. Similarly, PSA-SVD started producing more personalised recommendations than SA-SVD when recommending 10 or more books. Based on this assessment, these two contexts can be seen to complement each other well for recommending (larger amounts of) books.

Based on the objectives set out, the project was a success, especially in terms of analysing the impact and utility of personality and sentiment in a book recommender system. To be more specific, all basic and intermediate objectives from Section 1.2 had been met. On the other hand, only one of the advanced objectives: *Validate and test the chosen personality and sentiment acquisition methods against more advanced methods* had been partially accomplished as our APR method was evaluated against models from two well-cited research; the sentiment analysis model, however, was not formally evaluated due to time constraints.

5.2 Solution Limitations

A primary limitation faced at the early stage of implementation was the lack of access to advanced APR models. As evidence, the Personality Recogniser [33] and IBM Watson Personality Traits Service (see Section 2.4.1) were two publicly available model-based APR software that are currently no longer maintained or has been deprecated. With the additional time constraint of the project, a novel approach (Pure-LIWC), as detailed in Section 3.3.1, was implemented to address this issue.

Based on its evaluation results, our exploration to find an alternative personality recognition method was deemed fruitful. We had managed to implement a non-model-based technique, Pure-LIWC, that performed as well as half of the more advanced models in terms of accuracy score and precision. While it did underperform in its recall, the result should be treated more as a guideline as the method used to evaluate this specific metric was different for our approach and the approach of the cited research [13]. Nonetheless, the evaluation findings still do provide insight into the potential of the Pure-LIWC approach.

Furthermore, another methodology limitation stemmed from the sentiment analysis performed on book titles and descriptions. As evident from the disproportionate number of neutral-sounding books predicted in the dataset, Stanza was not able to accurately classify sentiment values for each book. While it was trained on a significant amount of data, including user reviews, it would seem it still lacked contextual awareness when analysing sentences. Through manual evaluation, it was discovered that the model defaulted to a neutral sentiment when the title and description contained both negative and positive words. This limitation could be considered a bias and potentially one of the reasons why our results have shown sentiment to have the greatest impact on recommendation quality. A possible approach to resolve said bias is mentioned in the next subsection.

Due to the nature of our context post-filtering approach, the diversity of all three configurations of the PSA-SVD model was not improved from the baseline (see Table 10). The primary issue was that filtering eliminates a certain subset of books to tailor to a user's profile [2], [4]. Mood management with sentiment filtering, coupled with the unbalanced distribution of book sentiments, meant that a vast majority of books were not considered unless a user had a neutral sentiment. Additionally, not considering negative sentiment books lowers the choices of recommendation even further by 15%. Moreover, re-ranking by personality was assumed to decrease diversity as the model would rank books that contained a user's predicted preferred genre higher. Since diversity is based on the heterogeneity of genres, having books of similar genres clumped together as the top K recommendations would not benefit this recommendation metric.

As our PSA-SVD model is a novel approach, especially in the book domain, there is a lack of state-of-the-art models or past experimental results that this research's model can be compared and contrasted against. The only comparison that can be made against the present study's model is the precision from a personality-and-emotion-aware travel destination recommender by Ishanka and Yukawa [25]. The results showed that their model had 61.18% precision which was better than the baseline, a non-context CF approach [25]. Our PSA-SVD model under-performs against this at a smaller recommendation size, K, conversely, it has the potential to do as well or even better where $K > 15$. This is not surprising as our sentiment-aware model in Section 4.3.2 had shown similar performance to that of the baseline as it only applies a simple filter, similar to Ishanka and Yukawa [25] who opted for context pre-filtering. While its precision may be higher, no evaluations on recall, personalisation and diversity were performed on the travel recommender. Hence, this comparison should not be wholly relied on in terms of evaluating all aspects of our system as more varied experimentation needs to be performed.

5.3 Future Improvements

Based on the findings and limitations faced, future work can consider the following. While our APR method has performed well against a few model-based approaches, more evaluations will have to be conducted to verify its correctness and robustness. Moreover, it would also be beneficial to look into the possibility of implementing an ML-based technique too that also takes into account the MRC Psycholinguistic dataset [60] as it has shown to produce better results [13]. Similarly, with more time, a domain-and-context-aware sentiment analysis model could be built too to facilitate more accurate and precise natural language processing of user review texts for books. This would potentially mitigate the frequency disparity issue of the book sentiments as highlighted by a previous subsection.

A primary suggestion would be to carry out online evaluations, i.e. user studies to obtain live feedback on the model's performance as this would help complement the findings of the offline experiments [48]. Moreover, understanding how a user behaves while interacting with the system is crucial, especially for public consumer-facing recommender systems such as this paper's model [4]. Adding to that, it would also be interesting to collect sentiment at the end of a user session too to further improve the model's performance, and it can be used as a user's entry sentiment too [52], [63] when they log on to the system again within a reasonably short amount of time. With user studies, more information about a user, such as their gender, could be collected to aid filtering recommendations by sentiment. This is because Raymond et al. [34] and Reinecke [47] discovered that different demographic groups have varying preferences in how they would like to manage their moods/sentiment. This would potentially resolve the current issue of negative sentiment books not being recommended at all in our system.

Furthermore, as results have shown that personality acts more as a complement to sentiment to improve personalisation and classification accuracy of recommender systems, the possibility of only considering a subset of personality traits could be explored. According to previous research conducted by one of my peers [31], it was found that models trained on personality configurations that involve openness and conscientiousness produced the best results in terms of rating and classification accuracy. Examples of these configurations include: OCEN, OCN, O, etc [31]. Hence, referencing these specific configurations to expand the research would provide great insight into the possibilities of improving our model's performance, especially, in terms of precision and recall.

Lastly, as the evaluation of 2D recommenders only covered model-based Collaborative Filtering (CF), testing should also be conducted for a memory-based CF model, e.g., user-based or item-based, to investigate its performance against the SVD. In addition, a new SVD++ model should be implemented from scratch that takes a more unique feature, such as if a user had added a book to their library, as implicit data, unlike the model provided by Surprise [24]. Personality data can also be incorporated into the Collaborative Filtering process by finding users who share similar FFM scores as the active user to potentially achieve better predictive accuracy.

6 Conclusion

In summary, the project aim was successfully achieved in terms of implementing a personality and sentiment-aware model that recommended physical books using a context post-filtering scheme. This involved the adoption of the 50:50 technique [38] on the Big Five Personality traits and mood management approach [47] for three stages of sentiment. The former relates to using 50% of a user's genre preference scores (based on their personality score) and 50% of the predicted rating from our underlying SVD model, while the latter aimed to maintain or promote positive sentiment in a user. Being a novel implementation, the methods used and analysis performed would contribute to the research area of similar recommender systems and domains surrounding books. Furthermore, the unique non-model-based approach of generating FFM scores using LIWC is a side contribution of this research which acts as a launchpad for further research into simpler yet accurate personality extraction methods.

The results have shown that the incorporation of personality and sentiment do notably impact a recommender system's recommendation accuracy and quality. More specifically, incorporating personality and sentiment shows signs of improving RS accuracy for larger recommendations while level of personalisation is significantly improved at the expense of diversity. Aside from the main research question, there were also a number of interesting findings that came out of our evaluations.

The sentiment-aware configuration had compellingly scored the highest in terms of recall against all compared models while was slightly behind on precision against the baseline SVD. Nevertheless, it still outperformed the personality-aware and personality and sentiment-aware models, thus showing its dominant role in affecting the recommendation output. Conversely, it should be highlighted that the two personality-aware models did present trends of increasing precision for larger recommendation sizes, suggesting a benefit in using personality to enhance recommendation accuracy in those circumstances. Moreover, the levels of personalisation from the context-aware models were significantly higher than the baseline while their diversity scores were no better than the baseline.

In the pipeline, our research could easily be expanded to support recommendations for E-books or audiobooks which have been growing in popularity in the last decade. Aside from sharing similar features to physical books (e.g., genres and descriptions), these newer domains have more features, such as listening time, which could be useful to further improve the system's performance.

References

[1] Erik , Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. *A Practical Guide to Sentiment Analysis*. Springer Publishing Company, Incorporated, 1st edition, 2017.

[2] Gediminas Adomavicius, Bamshad Mobasher, Francesco Ricci, and Alexander Tuzhilin. Context-aware recommender systems. *AI Magazine*, 32(3):67–80, 10 2011.

[3] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Syst.*, 23(1):103–145, jan 2005.

[4] Gediminas Adomavicius and Alexander Tuzhilin. *Context-Aware Recommender Systems*, pages 217–253. Springer US, Boston, MA, 2011.

[5] Haifa Alharthi, Diana Inkpen, and Stan Szpakowicz. A survey of book recommender systems. *J. Intell. Inf. Syst.*, 51(1):139–160, aug 2018.

[6] Nana Yaw Asabere, Amevi Acakpovi, and Mathias Bennet Michael. Improving socially-aware recommendation accuracy through personality. *IEEE Transactions on Affective Computing*, 9(3):351–361, 2018.

[7] American Psychological Association. APA Dictionary of Psychology. Available at: https://dictionary.apa.org/affect (Accessed on: 06 Dec 2021).

[8] 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.

[9] Vimala Balakrishnan and Hossein Arabi. Hyperm: A hybrid personality-aware recommender for movie. *Malaysian Journal of Computer Science*, 31:48–62, 01 2018.

[10] Prof. Ahmed Banafa. What is affective computing?, Aug 2018. Available at: https://www.bbvaopenmind.com/en/technology/digital-world/what-is-affective-computing/ (Accessed on: 01 Dec 2021).

[11] Shivani Bhosale, Pranjal Nimse, Siddhi Wadgaonkar, and Aishwarya Yeole. Suggestabook: A book recommender engine with personality based mapping. *International Journal of Computer Applications*, 159:1–4, 2017.

[12] Iván Cantador, I Fernandez-Tobias, Alejandro Bellogín, Michal Kosinski, and David Stillwell. Relating personality types with user preferences multiple entertainment domains. volume 997, 01 2013.

[13] Maira B. Carvalho and Max Louwerse. Grammar-based and lexicon-based techniques to extract personality traits from text. In Glenn Gunzelmann, Andrew Howes, Thora Tenbrink, and Eddy Davelaar, editors, *Proceedings of the 39th annual conference of the Cognitive Science Society*, pages 1727–1732, 2017. cogsci 2017 ; Conference date: 26-07-2017 Through 29-07-2017.

[14] Fabio Celli. Adaptive personality recognition from text. 2013.

[15] K. M. Kavin Chendhur, V. Priya, R. Mohana Priya, and S. Lavanya Lakshmi. Book recommender system using improved collaborative filtering. *International Journal of Research in Engineering, Science and Management*, 4(4):51–56, Apr. 2021.

[16] Oleg Chertov, Armelle Brun, Anne Boyer, and Marharyta Aleksandrova. Comparative analysis of neighborhood-based approach and matrix factorization in recommender systems. *Eastern-European Journal of Enterprise Technologies*, 3:4, 06 2015.

[17] Sahraoui Dhelim, Nyothiri Aung, Mohammed Amine Bouras, Huansheng Ning, and Erik Cambria. A survey on personality-aware recommendation systems, 2021.

[18] Mehdi Elahi, Matthias Braunhofer, Francesco Ricci, and Marko Tkalcic. Personality-based active learning for collaborative filtering recommender systems. volume 8249, 12 2013.

[19] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 149–156, 2011.

[20] Gustavo Gonzalez, Josep Lluis de la Rosa, Miquel Montaner, and Sonia Delfin. Embedding emotional context in recommender systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*, pages 845–852, 2007.

[21] Samuel D Gosling, Peter J Rentfrow, and William B Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

[22] 'Adli Ihsan Hariadi and Dade Nurjanah. Hybrid attribute and personality based recommender system for book recommendation. In *2017 International Conference on Data and Software Engineering (ICoDSE)*, pages 1–5, 2017.

[23] Khalid Haruna, Maizatul Akmar Ismail, Damiasih Damiasih, Haruna Chiroma, Tutut Herawan, and My. A comprehensive survey on comparisons across contextual pre-filtering, contextual post-filtering and contextual modelling approaches. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 15:1865 – 1874, 12 2017.

[24] Nicolas Hug. Surprise: A python library for recommender systems. *Journal of Open Source Software*, 5(52):2174, 2020.

[25] UA Piumi Ishanka and Takashi Yukawa. User emotion and personality in context-aware travel destination recommendation. In *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pages 13–18, 2018.

[26] Shuhao Jiang and Jinlin Song. Evaluation metrics for personalized recommendation systems. 1920(1):012109, may 2021.

[27] Raghav Karumur, Tien Nguyen, and Joseph Konstan. Personality, user preferences and behavior in recommender systems. *Information Systems Frontiers*, 20, 12 2018.

[28] Amir Kazem Kayhani, Farid Meziane, and Raja Chiky. Movies emotional analysis using textual contents. In Elisabeth Métais, Farid Meziane, Helmut Horacek, and Philipp Cimiano, editors, *Natural Language Processing and Information Systems*, pages 205–212, Cham, 2020. Springer International Publishing.

[29] Yehuda Koren and Robert M. Bell. Advances in collaborative filtering. In *Recommender Systems Handbook*, 2011.

[30] Aldin Kovačević and Zerina Mašetić. Reliable book recommender system: An evaluation and comparison of collaborative filtering algorithms. In Naida Ademović, Edin Mujčić, Zlatan Akšamija, Jasmin Kevrić, Samir Avdaković, and Ismar Volić, editors, *Advanced Technologies, Systems, and Applications VI*, pages 264–280, Cham, 2022. Springer International Publishing.

[31] Alexandra Krajewski. Minimising the storage of personality data in recommender systems without compromising on quality, 2021. MEng Computer Science Final Year Project Report.

[32] John Kalung Leung, Igor Griva, and William G. Kennedy. Text-based emotion aware recommender. *Computer Science Information Technology*, Jul 2020.

[33] Francois Mairesse, Marilyn Walker, Matthias Mehl, and Roger Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res. (JAIR)*, 30:457–500, 09 2007.

[34] Raymond A. Mar, Keith Oatley, Maja Djikic, and Justin Mullin. Emotion and narrative fiction: Interactive influences before, during, and after reading. *Cognition and Emotion*, 25(5):818–833, 2011. PMID: 21824023.

[35] Praveena Mathew, Bincy Kuriakose, and Vinayak Hegde. Book recommendation system through content based and collaborative filtering method. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pages 47–52, 2016.

[36] Jan Mizgajski and Mikolaj Morzy. Affective recommender systems in online news industry: how emotions influence reading choices. *User Modeling and User-Adapted Interaction*, 29:345–379, 2018.

[37] Christian Montag, Eilish Duke, and Alexander Markowetz. Toward psychoinformatics: Computer science meets psychology. *Computational and Mathematical Methods in Medicine*, 2016:1–10, 06 2016.

[38] Orestis Nalmpantis and Christos Tjortjis. The 50/50 recommender: A method incorporating personality into movie recommender systems. 05 2018.

[39] Annalyn Ng, Maarten Bos, Leonid Sigal, and Boyang Li. Predicting personality from book preferences with user-generated content labels. *IEEE Transactions on Affective Computing*, PP, 07 2017.

[40] Tien Nguyen, Franklin Harper, Loren Terveen, and Joseph Konstan. User personality and user satisfaction with recommender systems. *Information Systems Frontiers*, 20:1–17, 12 2018.

[41] Emmanuel Okon, Bartholomew Eke, and Prince Asagba. An improved online book recommender system using collaborative filtering algorithm. 05 2018.

[42] James W. Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. *The development and psychometric properties of LIWC2015*. University of Texas at Austin, 2015.

[43] James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[44] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

[45] 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.

[46] Chhavi Rana and Sanjay Jain. Building a book recommender system using time based content filtering. 11:27–33, 02 2012.

[47] Leonard Reinecke. *Mood Management Theory*, pages 1–13. John Wiley Sons, Ltd, 2016.

[48] Thiago Silveira, Min Zhang, Xiao Lin, Yiqun Liu, and Shaoping Ma. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10:813–831, 2019.

[49] SmartSurvey. Likert scale questions, Mar 2021. Available at: https://bit.ly/3MiIpPw (Accessed: 15 April 2022).

[50] Sanja Stajner and Seren Yenikent. A survey of automatic personality detection from texts. pages 6284–6295, 01 2020.

[51] David Stillwell and Michal Kosinski. Mypersonality.org, May 2018.

[52] Marko Tkalčič, Urban Burnik, Ante Odić, Andrej Košir, and Jurij Tasič. Emotion-aware recommender systems – a framework and a case study. In Smile Markovski and Marjan Gusev, editors, *ICT Innovations 2012*, pages 141–150, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[53] Marko Tkalvcivc, Andrej Kosir, Jurij Tasivc, and Matevž Kunaver. Affective recommender systems: the role of emotions in recommender systems. pages 9–13, 01 2011.

[54] Marko Tkalčič, Matevž Kunaver, Jurij Tasic, and Andrej Kosir. Personality based user similarity measure for a collaborative recommender system. 01 2009.

[55] Meena Toor. Three tips for effectively designing rating scales, 01 2021.

[56] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[57] Mengting Wan and Julian J. McAuley. Item recommendation on monotonic behavior chains. In Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan, editors, *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM, 2018.

[58] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. Fine-grained spoiler detection from large-scale review corpora. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics, 2019.

[59] C.-Y Wang, Y.-C Wang, and S.-C.T. Chou. A context and emotion aware system for personalized music recommendation. *Journal of Internet Technology*, 19:765–779, 01 2018.

[60] Michael Wilson. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20:6–10, 1988.

[61] Hsin-Chang Yang and Zi-Rui Huang. Mining personality traits from social messages for game recommender systems. *Knowledge-Based Systems*, 165:157–168, 2019.

[62] Tal Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373, 2010.

[63] Yong Zheng, Robin Burke, and Bamshad Mobasher. The role of emotions in context-aware recommendation. 10 2013.