香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

DDA4210
Advanced Machine Learning

BorderBrews: Location-based Beer Recommendation System

Group 9

Group Member:	*Student Number:*
Hitomi Tee Jin Ling	121010011
William Alexander Tanex	120040014
Ding Yeen Yi	121040040
Kee Cheng	121040041

~ May 25, 2024 ~

I. Introduction

A. Significance

Beer consumption is a cultural and traditional activity in many countries, such as Germany, Belgium, and the United States. It transcends mere recreation and socialization, representing a generational heritage. Beer, the most consumed alcoholic beverage globally, also provides nutrients, including vitamin B, minerals, and fiber derived from its ingredients—barley, hops, and yeast (Neves et al., 2011).

Over the past decade, the brewery market has continuously flourished, particularly with the proliferation of online evaluation forums. This presents beer drinkers with a vast array of options. However, this abundance can leave consumers feeling overwhelmed by the sheer number of choices or undereducated about the available options to try something new (Allen & Wetherbee, n.d.). Consequently, when consumers find a beer they enjoy, they often struggle to discover other similar beers that might suit their tastes. This is where our Beer Recommendation System comes into play. Whether you're searching for a similar beer while traveling or curious about new beers that better match your taste, our system assists users in navigating this expansive market to find beers aligned with their preferences. This not only enhances the consumer experience but also fosters appreciation for the rich diversity within the beer world.

B. Novelty

We introduced BorderBrews, a location-based beer recommendation system where it addresses three main gaps in the existing available beer recommendation systems.

Firstly, traditional beer recommendation systems have been trained on the readily available BeerAdvocate or RateBeer dataset, which includes subjective ratings and text reviews. However, these systems often overlook important factors such as the brewery location of the beers, which significantly influence beer preferences (Offutt, 2013). Studies have demonstrated a strong correlation between brewery location and consumer preferences for certain beer styles (Byeon et al., 2021; McCluskey & Shreay, 2011). Understanding these regional and cultural associations is essential for accurate recommendations. Our system addresses this by incorporating brewery location filters into its algorithm. Whether you prefer German lagers, Belgian ales, or American craft beers, our system considers these factors to deliver recommendations aligned with your tastes. This enhancement improves accuracy and enriches the user experience by recognizing the importance of geographical and cultural factors in beer preferences.

Secondly, traditional recommendation systems lack the ability to let users specify their preferences for particular beer features, such as aroma or taste, when making recommendations (Roy & Dutta, 2022). Many consumers evaluate beers based on these specific features, such as the floral notes in a hoppy IPA, the rich maltiness in a stout, or the fruity esters in a Belgian ale. These sensory characteristics are often key determinants in a consumer's enjoyment of a beer (Maria Isabel Betancur et al., 2020; Habschied et al., 2022). Therefore, our recommendation system accommodates these preferences to provide more personalized and satisfying suggestions.

Thirdly, traditional recommenders often rely on a single similarity metric, such as the Pearson correlation coefficient or cosine similarity, to generate recommendations (*BEER DATASET ANALYSIS*, 2020; Alex Yuan Li, 2017; HsiangHung, 2017; robin26091991, 2020). While Pearson correlation identifies linear relationships between users' ratings and is useful for identifying similar rating patterns, it's sensitive to rating scales and data availability (Benesty et al., 2009; Armstrong, 2019). On the other hand, cosine similarity measures the orientation between vectors, suitable for high-dimensional spaces but overlooks rating magnitudes and struggles with sparse data (Li & Han, 2013; Xia et al., 2015). Combining both metrics is beneficial as each captures different aspects of similarity. By leveraging Pearson's strength in identifying linear relationships and cosine's scale independence, our recommendation system offers more nuanced and effective recommendations, enhancing accuracy and reliability.

II. Data Collection & Preprocessing

Our project utilized the online BeerAdvocate database, housing over 1.5 million beer reviews. These reviews not only provided subjective ratings—appearance, aroma, taste, palate, and overall impression— but also detailed textual descriptions of consumers' experiences. See Appendix A for the link to raw data and Appendix B for more statistical details. In the preprocessing stage, reviews lacking complete features i.e. missing one of the five numerical ratings or the textual review, were removed, ensuring data completeness and reliability. Each beer and brewery received a unique, randomly hashed ID for simplified data manipulation and integrity maintenance. Additionally, a minimum threshold of 10 reviews per beer was set to focus on well-reviewed beers, minimizing data sparsity and removing less relevant entries. These efforts resulted in a structured dataset with each review represented as a row containing seven features, supporting advanced hybrid recommendation functionalities blending collaborative and content-based filtering for personalized beer suggestions. See Table 1 for the statistics.

Table 1. *Post-processed Dataset Statistics*

Dataset Statistics Summary	
Number of reviews	86,530
Number of users	9,854
Number of beers (>10 reviews)	8,653

Note. See Appendix C for the dataframe.

To enhance the dataset's utility, we scraped brewery geographical details from BeerAdvocate forums for location filtering in our recommendation system, capturing brewery names, states, and countries. This location dataset consists of 162 countries with 38,105 unique beers. See Appendix D for more details.

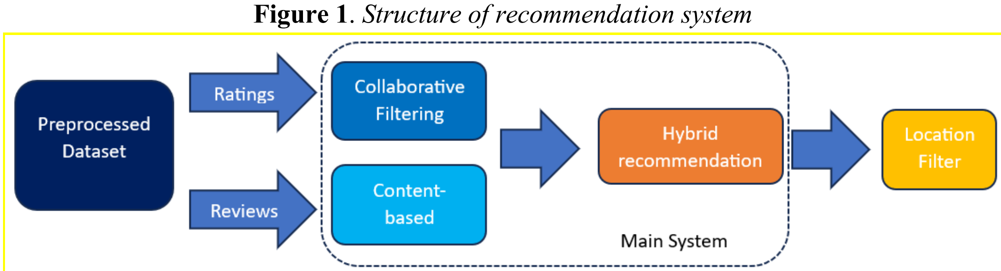
III. Methodology

We implement a hybrid beer recommendation system combining Singular Value Decomposition (SVD) and Term Frequency-Inverse Document Frequency (TF-IDF) models to overcome the limitations of using Collaborative Filtering (CF) and Content-Based (CB) models individually. CF struggles with the cold start problem and scalability issues, while CB relies on high-quality textual data and tends to over-specialize. By merging CF and CB, the hybrid model leverages their strengths and mitigates their weaknesses, resulting in more accurate, diverse, and personalized beer recommendations, enhancing user satisfaction.

The motivation for not using Large Language Models (LLMs) in our recommendation system lies in the efficiency and effectiveness of conventional hybrid models, which combine metrics like Pearson correlation and cosine similarity. These models are less resource-intensive, simpler, and more interpretable, offering faster inference and better scalability. Also, we aim to explore the performance of the fundamental hybrid recommendation systems, thoroughly assess their capabilities and advantages in recommendation tasks, gaining insights into their strengths and limitations relative to more advanced approaches like LLMs.

Our hybrid system takes four inputs: beer name, number of recommendations, user-preferred features (users can choose any feature rating from overall, appearance, aroma, palate, and taste; the default is overall), and location. It then outputs a ranked list of similar beer recommendations.

Figure 1. *Structure of recommendation system*



A. Collaborative Filtering (CF)

For CF, we implemented truncated Singular Value Decomposition (SVD) to identify latent factors within user-beer interaction matrices and utilized a matrix R of size 9,000 (beers) ×10,000 (users), where each entry R_{ij} represents the rating given by user j to beer i. Given the computational challenges posed by large datasets, we employed a truncated SVD model with 500 components, balancing efficiency and representation quality, capturing approximately 60% of the dataset's variance (see Appendix F).

To form the basis for recommendations, we constructed a beer similarity matrix S of size m×m using the Pearson correlation. Here, S_{ij} represents the similarity between beers i and j:

$$S_{ij} = \frac{\sum_{u \in U}(R_{iu} - \bar{R}_i)(R_{ju} - \bar{R}_j)}{\sqrt{\sum_{u \in U}(R_{iu} - \bar{R}_i)^2}\sqrt{\sum_{u \in U}(R_{ju} - \bar{R}_j)^2}}$$

where R_{iu} and R_{ju} are the ratings of beers i and j by users u, and $\underline{R_i}$ and $\underline{R_j}$ are the average ratings.

B. Content-Based Method (CB)

For CB, we employed the TF-IDF model to analyze textual data from user reviews. This method enhances beer recommendations by leveraging the descriptive content of user experiences.

The process begins with the creation of a vectorizer, which transforms the text data into feature vectors while removing stopwords to enhance the relevance of the features and improve content-based filtering clarity. The text data is then transformed into a TF-IDF matrix, with each cell corresponding to a weight (TF-IDF scores) of a word for a beer, quantifying the importance of each term within the reviews.

Figure 2. *TF-IDF Matrix*

66455 important words (terms)

8653 beer data (documents)

Beer Name	Term 1	Term 2	Term 3	...
Beer A				
Beer B				
Beer C				
...				

The TF-IDF scores are computed as TF-IDF(t,d) = TF(t,d) × IDF(t) where TF(t,d) is the term frequency of term t in document d and IDF(t) is the inverse document frequency of term t. To identify the top similar beers, we calculate the cosine similarity between the TF-IDF vectors. Cosine similarity, $c \in [0,1]$, measures the cosine of the angle between two vectors, providing a metric for textual similarity:

$$C_{ij} = \frac{M_i \cdot M_j}{||M_i|| \, ||M_j||}$$ where M_i and M_j are TF-IDF vectors for two beers.

To align these similarity scores with the Pearson correlation coefficient used in the collaborative filtering approach, we rescale the cosine similarity scores from the range [0,1] to [-1,1]: $c' = 2c - 1$, where c is the original cosine similarity score and c' is the rescaled score.

C. Hybrid Model with Location Filtering

To calculate the final recommendation score for each beer, we compute the mean of the rescaled cosine similarity score from the content-based method and the Pearson correlation coefficient from the collaborative filtering method. This averaging approach ensures that both CF and CB similarities are equally weighted in the recommendation process. The hybrid model then generates a list of beer recommendations based on their recommendation score. Beers with higher recommendation scores, indicating greater similarity to the user's preferences, are prioritized in the list. $Recommendation\ score = \frac{S_{ij} + C'_{ij}}{2}$, where S_{ij} and C'_{ij} are the similarity scores from CF and CB methods, respectively.

Lastly, we incorporate a location filter. The final output is a ranked list of similar beers, including their brewery names, filtered based on the user's location. Location filtering is the final, optional step in our system. If users opt out of location-based recommendations, the list will include a broader range of beers, offering very similar options without considering geographical constraints. This ensures that users can find the most similar beers regardless of location, enhancing the completeness of the recommendation list.

IV. Experimental Results & Discussion

To test the models, we introduced two evaluation metrics for the recommendation system:

1. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

2. Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left| \hat{y}_i - y_i \right|$$

y_i: $average\ rating\ of\ inputted\ beer, \widehat{y_i}$: $average\ ratings\ of\ recommended\ beers, n$: $number\ of\ users$

Due to the lack of real user feedback on recommendations for ground truth, y_i and $\widehat{y_i}$ are unknown; therefore, we take the mean of all users' ratings to replace them. Specifically, y_i is the vector that contains the average of all subjective ratings—appearance, aroma, taste, palate, and overall—from all users of the particular inputted beer. Meanwhile, $\widehat{y_i}$ contains as many average rating vectors as the expected number of recommendations. For example, if four similar beers of "Devil Dog" beer are required, and there are 56 user ratings for "Devil Dog" beer, y_i is a 1×5 dimensional vector representing the average ratings of its 56 users, and $\widehat{y_i}$ is a 56×4 matrix where each column contains a 1×5 vector representing the average ratings for each of the four recommended beers.

Table 2. *Models Comparison based on RMSE and MAE*

Model		SVD	TF-IDF	Hybrid Model
Before location filtering	RMSE	0.24401664272282197	0.2689495876280654	0.22799919653984968
	MAE	0.1935	0.2230	0.1848
After location filtering	RMSE	0.6167937317218032	0.549641050192565	0.45565190020552215
	MAE	0.4636	0.4048	0.348

Note. See Appendix G and Appendix H for more details.

As illustrated, the RMSE and MAE of our hybrid model are approximately 0.228 and 0.1848, respectively, the lowest among the SVD and TF-IDF models individually. Thus, we can conclude that our hybrid model outperforms both SVD and TF-IDF by leveraging their combined strengths.

Testing before location filtering showed that TF-IDF alone had the lowest accuracy, highlighting its limitation in context capture. The TF-IDF model fails to capture the semantics of phrases since each word in user reviews is treated separately without considering its context (Neural Ninja, 2023). Furthermore, the "bag of words" approach used in TF-IDF ignores word order, potentially leading to misunderstandings of the text. Another factor contributing to this result is the biased "ground truth." Since the ground truth is the mean of all users' ratings, the evaluation metrics may favor collaborative filtering models that recommend beers based on similar user ratings rather than content-based methods that suggest beers based on textual reviews.

Testing after location filtering showed that SVD alone had the lowest accuracy. The RMSE for all models was around 0.5, indicating suboptimal recommendation quality. This issue arises because location filtering significantly reduces the dataset size (see Appendix E), and SVD requires a relatively dense matrix for effective decomposition. In this case, the user-item matrices were too sparse, affecting the quality of the recommendations.

VI. Conclusion

In conclusion, we present BorderBrews, a novel beer recommendation system distinguished by its innovative incorporation of brewery location filters, user-specific feature preferences, and the combination of Pearson correlation and cosine similarity metrics to deliver personalized and culturally relevant suggestions. Our experimental results demonstrate that the hybrid model outperforms both SVD and TF-IDF models individually, achieving the lowest RMSE and MAE scores. However, there are several limitations to this system. Firstly, evaluating the performance of hybrid models can be challenging without a solid ground truth value, making it difficult to assess their effectiveness objectively. Therefore, we recommend future research focusing on feature expansion and model stability to address these limitations. On top of that, efforts should be made to identify suitable evaluation metrics that can provide meaningful insights into the performance of hybrid models. Moreover, as factors such as price, sourness, and sweetness also play crucial roles in determining beer consumption preferences, future research could involve expanding the dataset to incorporate these attributes, further enhancing the recommendation system's accuracy and relevance. Future work can also implement improved NLP algorithms to refine content-based recommendations, contributing to more precise and personalized suggestions and more accurate context capture. Lastly, a large dataset is always preferable for all SVD, TF-IDF and hybrid models.

References

Alex Yuan Li. (2017). *NINKASI: Beer Recommender System*. Data Science Blog. https://nycdatascience.com/blog/student-works/ninkasi-beer-recommender-system/

Allen, A., & Wetherbee, R. (n.d.). BeerMe: A Beer Recommendation System. https://www.cs.uml.edu/ecg/uploads/AIfall14/allen_wetherbee_beerme_recommendation.pdf

Armstrong, R. A. (2019). Should Pearson's correlation coefficient be avoided? *Ophthalmic and Physiological Optics/Ophthalmic & Physiological Optics, 39*(5), 316–327. https://doi.org/10.1111/opo.12636

BEER DATASET ANALYSIS Final Report B8IT110 Higher Diploma in Science in Data Analytics. (2020). https://esource.dbs.ie/server/api/core/bitstreams/2a7378d8-1011-4a3d-8100-5f809384bced/content

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. *Springer Topics in Signal Processing*, 1–4. https://doi.org/10.1007/978-3-642-00296-0_5

Byeon, Y. S., Lim, S. T., Kim, H. J., Kwak, H. S., & Kim, S. S. (2021). Quality characteristics of wheat malts with different country of origin and their effect on beer brewing. *Journal of Food Quality*, 1-12. https://doi.org/10.1155/2021/2146620

Habschied, K., Vinko Krstanović, & Krešimir Mastanjević. (2022). Beer Quality Evaluation—A Sensory Aspect. *Beverages, 8*(1), 15–15. https://doi.org/10.3390/beverages8010015

HsiangHung. (2017). *GitHub - HsiangHung/Beer-Recommender*. GitHub. https://github.com/HsiangHung/Beer-Recommender

Li, B., & Han, L. (2013). Distance Weighted Cosine Similarity Measure for Text Classification. *Lecture Notes in Computer Science*, 611–618. https://doi.org/10.1007/978-3-642-41278-3_74

Maria Isabel Betancur, Kosuke Motoki, Spence, C., & Velasco, C. (2020). Factors influencing the choice of beer: A review. *Food Research International, 137*, 109367–109367. https://doi.org/10.1016/j.foodres.2020.109367

McCluskey, J. J., & Shreay, S. (2011). Culture and beer preferences. *The economics of beer*, 161-170. https://10.1093/acprof:oso/9780199693801.003.0009

Neural Ninja. (2023, June 30). *TF-IDF: Weighing Importance in Text - Let's Data Science*. Let's Data Science. https://letsdatascience.com/tf-idf/

Neves, M. F., Trombin, V. G., Lopes, F. F., Kalaki, R., & Milan, P. (2011). World consumption of beverages. In The orange juice business. *Wageningen: Wageningen Academic Publishers*. https://doi.org/10.3920/978-90-8686-739-4_31

Offutt, B. (2013). *HapBeer: A Beer Recommendation Engine CS 229 Fall 2013 Final Project*. https://cs229.stanford.edu/proj2013/Offutt-Hapbeer.pdf

robin26091991. (2020, April 22). *Beer Recommendation System Ashish_Pandey*. Kaggle.com; Kaggle. https://www.kaggle.com/code/robin26091991/beer-recommendation-system-ashish-pandey#Using-Cosine-Similarity

Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data, 9*(1). https://doi.org/10.1186/s40537-022-00592-5

Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information Sciences*, *307*, 39–52. https://doi.org/10.1016/j.ins.2015.02.024

Appendices

A. BeerAdvocate Dataset

https://datarepo.eng.ucsd.edu/mcauley_group/data/beer/beeradvocate.json.gz

The dataset will automatically download once you click on the link.

B. Description of the Original BeerAdvocate Dataset

Table 3. *Original BeerAdvocate Dataset Statistics*

Dataset Statistics Summary	
Number of reviews	1,586,259
Number of users	33,387
Number of beers	66,051
Users with >50 reviews	4,787
Median of the number of words per review	126
Timespan	January 1998 - November 2011

C. Description of the Post-processed Dataset After Preprocessing

Figure 3: *Dataframe of Post-processed Dataset*

	beer_name	beer_id	user_name	user_overall_rating	user_appearance_rating	user_aroma_rating	user_palate_rating	user_taste_rating	user_review
0	Rauch √úr Bock	58046	UCLABrewN84	4.5	3.0	4.5	4.0	4.5	Pours a murky light brown with a 1 inch fizzy ...
1	Rauch √úr Bock	58046	zaphodchak	4.0	4.0	4.0	3.0	4.0	Faint sudsy head with some with some dissipati...
2	Rauch √úr Bock	58046	Tilley4	4.0	4.0	4.5	3.5	4.0	A new arrival to the West TN area... \t\tPours...
3	Rauch √úr Bock	58046	bashiba	4.5	4.0	5.0	4.0	4.0	Poured a slightly cloudy deep amber/red color ...
4	Rauch √úr Bock	58046	Klym	4.5	3.5	4.5	4.0	4.5	Big thanks to N2168 for knocking this off my w...
...	...	...	...	...	...	...	...	...	...
86525	Little Thumper Ale	33647	Gmann	4.0	3.0	3.5	3.5	4.0	Pours a clear straw color with a small wispy w...
86526	Little Thumper Ale	33647	sleazo	5.0	4.0	3.5	4.0	4.5	A_Slightly hazy. The liquid itself is straw ye...
86527	Little Thumper Ale	33647	RblWthACoz	4.0	4.0	4.0	4.0	4.5	Pours a straw yellow. Smells like dried straw ...
86528	Little Thumper Ale	33647	tgbljb	4.5	4.0	4.0	4.0	4.5	Served directly from the tank at the brewery.\...
86529	Little Thumper Ale	33647	TongoRad	5.0	4.0	4.0	4.5	4.0	Nicely carbonated- especially for a growler fi...

86530 rows × 9 columns

D. Description of the Scrapped Brewery Location Dataset

Table 4. *Brewery Location Dataset Statistics*

Dataset Statistics Summary	
Number of beers	38,105
Number of countries	162
Number of unique locations	213
Locations of beers with >10 reviews	104

Figure 4. *Dataframe of Brewery Location Dataset*

	beer_id	beer_name	brewery_id	brewery_name	state	country
0	351746	The Optimist	617781	Fort George Brewery + Public House	Oregon	United States
1	802233	Fresh IPA	617781	Fort George Brewery + Public House	Oregon	United States
2	779935	Overdub IPA	617781	Fort George Brewery + Public House	Oregon	United States
3	100955	Magnanimous IPA	617781	Fort George Brewery + Public House	Oregon	United States
4	604816	Big Guns	617781	Fort George Brewery + Public House	Oregon	United States
...	...	...	...	...	...	...
47104	736136	Pineapple Pale Ale	499232	Upslope Brewing Company - Lee Hill	Colorado	United States
47105	806302	South African Pale Ale	499232	Upslope Brewing Company - Lee Hill	Colorado	United States
47106	478198	Mary Jane Ale	499232	Upslope Brewing Company - Lee Hill	Colorado	United States
47108	312723	Rye Fish At All?	507497	Boathouse Brewpub & Restaurant	Minnesota	United States
47109	907591	Ely Nevada Pale Ale	507497	Boathouse Brewpub & Restaurant	Minnesota	United States

38105 rows × 6 columns

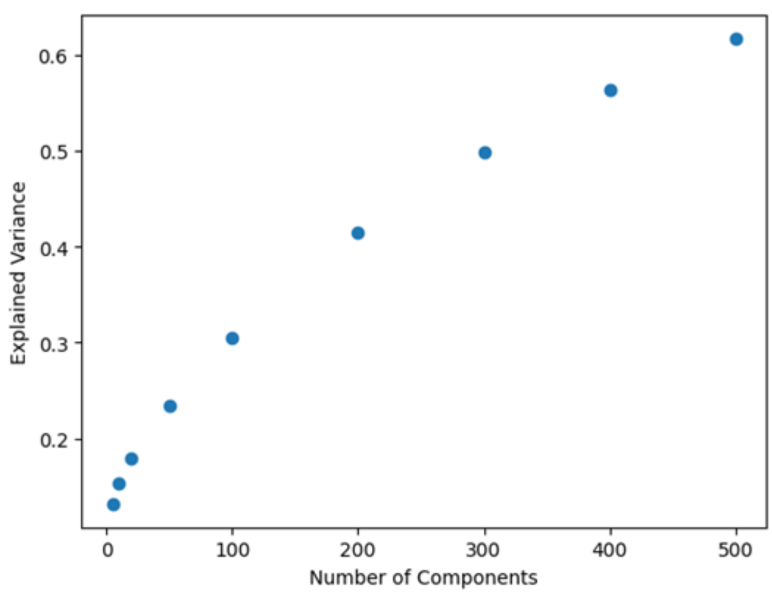
E. Description of the Post-processed Brewery Location Dataset

Table 5. *Brewery Location Dataset Statistics*

Dataset Statistics Summary	
Number of reviews	15,440
Number of users	4,551
Number of beers	1,544
Number of countries	162
Number of unique locations	213

F. Explained Variance of Truncated SVD

Figure 5. *Explained Variance According to the Number of Components*



G. Experimental Results of Model Comparison (before location filtering)

Figure 6. *Root Mean Square Error of Each Model (Before Location Filtering)*

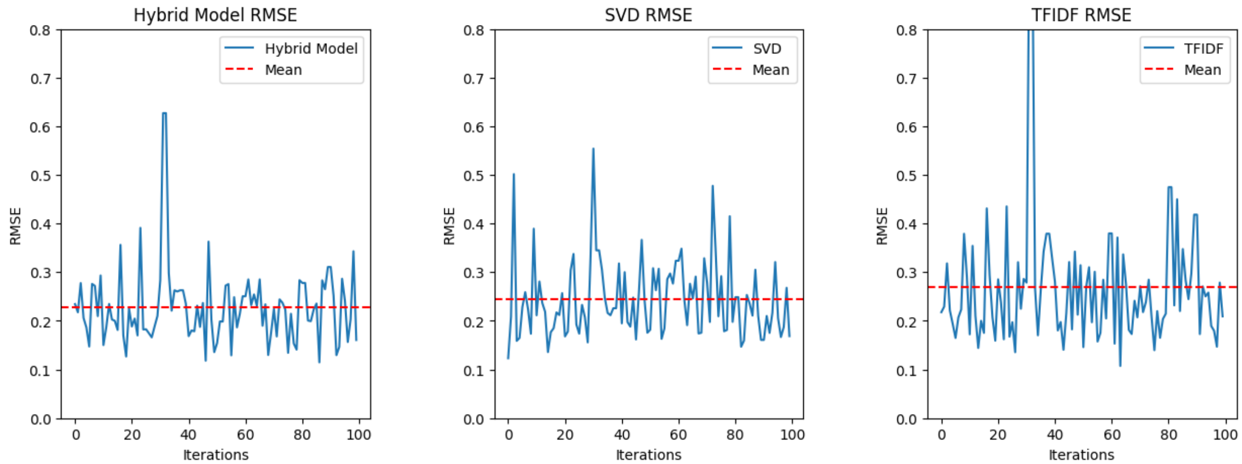
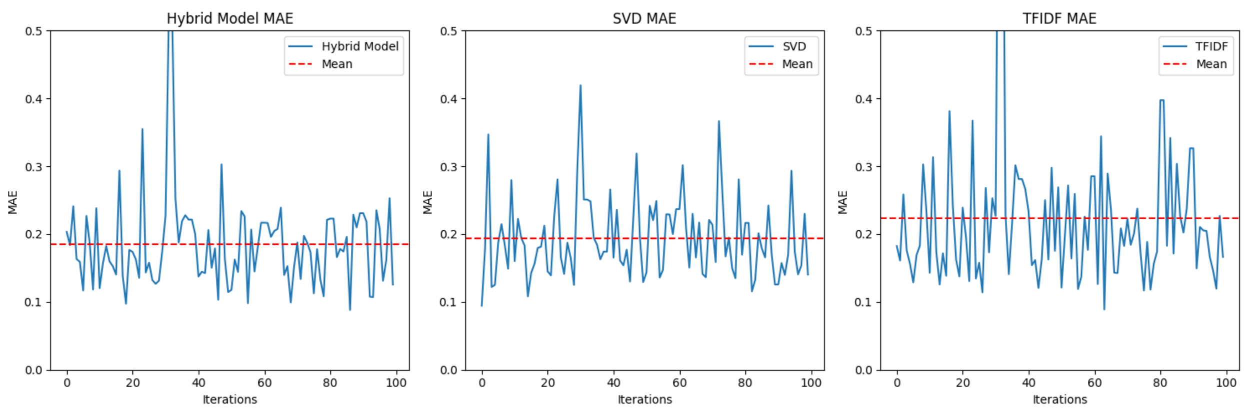


Figure 7. *Mean Absolute Error of Each Model (Before Location Filtering)*



H. Experimental Results of Model Comparison (after location filtering)

Figure 8. *Root Mean Square Error of Each Model (After Location Filtering)*

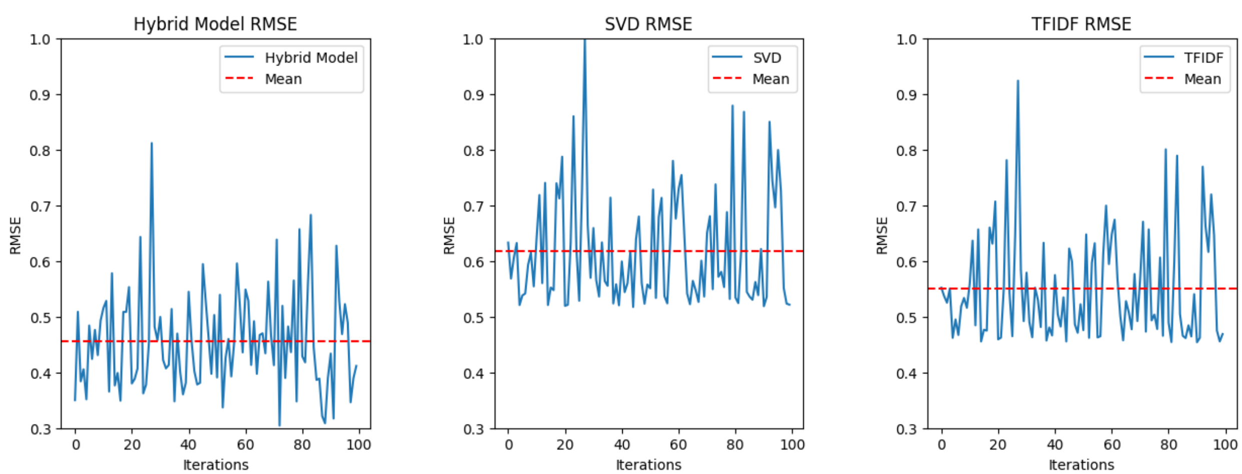


Figure 9. *Mean Absolute Error of Each Model (After Location Filtering)*