Pertanika J. Sci. & Technol. 33 (1): 333 - 354 (2025)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A Supervised Genre-based Recommendation Model for Game Review

Jun Wang and Keng Hoon Gan*

School of Computer Sciences, Universiti Sains Malaysia, 11800 Gelugor, Pulau Pinang, Malaysia

ABSTRACT

The gaming industry is becoming more and more popular as the number of players increases. Game recommendations allow players to quickly decide if it is worth playing. This article explores how reviews of players who have played the game can be used to decide whether the game should be recommended. Nevertheless, genre-oriented models have not been incorporated in the recommendation. Since different genres have different characteristics that attract different groups of players, generalized recommendation models may not be effective enough in dealing with specific genres of games. This article proposes a genre-based recommendation model using a supervised machine learning model. All game datasets will be divided into six genres (Action, RPG, Adventure, FPS, Horror, and Strategy). Each game genre is trained separately with three feature selection methods (Bag of Words, N-Gram and Part of Speech) and three classification algorithms (Naive Bayes, Support Vector Machine and Decision Tree). The experiment results found that genre-based models mostly outperform the general model (without differentiating by genre). The best feature selection and combination of classification algorithms is also obtained for each genre, with Bag of Words and Naive Bayes topping most genres. For example, the accuracy achieved by the FPS model is 0.854 compared to the general (all genres) model with an accuracy of 0.828.

Keywords: Feature selection, game review, machine learning, recommendation

INTRODUCTION

In recent years, the need for recommendations has arisen in domains like online selling, academia and entertainment. This need stems from information overload, which causes users to be overwhelmed by choices when buying products, choosing courses and picking movies. Providers must assist users in making better decisions.

ARTICLE INFO

Article history:
Received: 19 December 2023
Accepted: 3 September 2024
Published: 27 January 2025

DOI: https://doi.org/10.47836/pjst.33.1.15

E-mail addresses:
wangjun@student.usm.my (Jun Wang)
khgan@usm.my (Keng Hoon Gan)
*Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

In gaming, game quality is vital for player experience. Game review content offers hints about the game's quality. Analyzing the contents helps players grasp game aspects and playability, aiding the decision of whether a game is worth playing. This task is treated as classification. The models will predict whether a game will be recommended based on the supervised learning of previous players' reviews.

Steam, a major digital distribution platform, allows players to buy, discuss, and share games and software. With millions of users and 47 million active players daily, Steam has been developed for 20 years to improve the player experience. Players often review games and express whether they would recommend them. The percentage of recommendations affects others' perceptions and buying decisions. Figure 1 shows reviews posted on the Steam gaming platform.

Figure 1. Game review on the Steam platform (https://store.steampowered.com/)

Nevertheless, not all game discussion forums have a recommendation feature, making it difficult for players to judge a game's playability. For example, discussions about games on Twitter do not indicate whether the game is recommended (Figure 2).

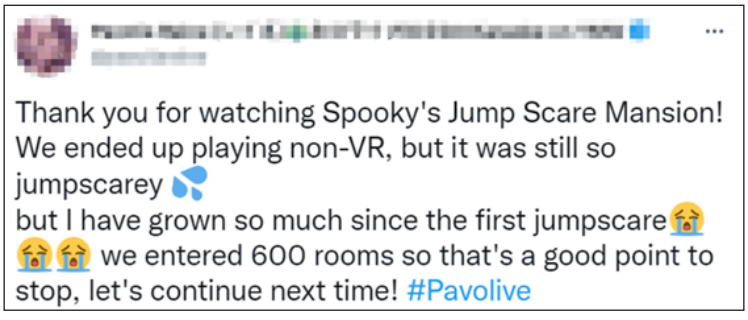

Figure 2. Game review on Twitter

In this context, players must rely on reviews to assess the games. The game recommendation model would be useful for automating the process of processing the reviews and suggesting a recommendation. Recommendations benefit players by allowing

them to swiftly take others' feedback as their reference. Game developers can also learn from the non-recommended reviews to understand complaints, overcome shortcomings, identify needs, and create better work.

In the domain of gaming, each game will have different characteristics and different ways of playing, so they will have their own genres. Some games may also span multiple genres, incorporating multiple play styles and elements. For example, The Legend of Zelda: Breath of the Wild, a game that contains both Action and Adventure genres, focuses on exploring an open world and upgrading characters by completing quests. Therefore, some genre-based keywords, such as combat and backstory, may appear in the review. In other games from other genres, there will be different feature words; for example, horror games will appear as scary, frightening, and so on.

A classifier is constructed to understand players' written reviews to create a game recommendation model. These reviews are categorized as recommended or not. Zuo (2018) presented a similar classifier for Steam reviews, categorizing them as positive (recommended) or negative (not recommended). Naïve Bayes and Decision Tree (with N-Gram features) were used. Decision Trees outperformed Gaussian Naive Bayes with about 75% accuracy on the Steam Review Dataset. However, there's room for improvement due to limited feature selection methods and algorithms.

Player-written descriptions play a pivotal role in training a recommendation model. For instance, "scared with creepy voices" indicates recommendation, while "repetitive without excitement" indicates not recommended. Previous research explored enhancing recommendation models using machine learning or rule-based methods. Viggiato et al. (2021) used Stanford CoreNLP, Sentistrength, and self-trained Naïve Bayes to analyze sentiment, with modest and poorly performing results on game reviews. Accuracy ranged from 0.37 (Stanford CoreNLP) to 0.61 (NLTK). Zuo (2018) reported higher accuracy (0.7495) using Decision Tree and N-Gram features. Therefore, exploring different textual features like Bag of Words, N-Gram, and Part of Speech (POS) could be valuable.

Viggiato et al. (2021) identified the root causes of poor classification, including negative terminology causing misinterpretation of positive words. They proposed a stratified training process based on the game genre for their sentiment analysis classifier. The results show a positive outcome.

As far as we are concerned, no related work combines game genres with various feature selection methods and classification algorithms applied to sentiment classification. Therefore, Therefore, we propose a genre-based model for game recommendation. Our proposed model will be compared with a general recommendation that will be trained based on reviews from all games (without separating by genre). Three feature extraction methods (BOW, N-Gram and POS) and three classification algorithms (NB, SVM and DT) that have been shortlisted from the literature will be used in the model training.

LITERATURE REVIEW

The literature related to game recommendations is mostly related to the sentiment expressed in the review text. This discussion covers three main areas of work: game recommendation, genre classification and sentiment classification.

Game Recommendation

Bais et al. (2017) conducted a sentiment analysis study using Steam reviews as a dataset to detect sarcasm in reviews. Methods including Lexicon Score Aggregation, Multinomial Naive Bayes, Modified Turney's Algorithm, Logistic Regression and Linear SVM were evaluated in the experiment. According to the reported results, linear SVM performed best with an accuracy of 0.935. Britto and Pac´ıfico (2020) used the Portuguese language game data set and BOW as the feature extraction method for the training of three classifiers, which are Random Forest, Support Vector Machines and Logistic Regression, respectively. The best-performing model is the SVM model, with an accurate rate of 82.54%.

Genre Classification

Another class of work related to the game domain would be genre-related research. Jiang and Zheng (2023) utilized the cover image, description text, title text and genre information of 50,000 video games as a dataset and divided them into 21 game genres. The authors thoroughly evaluated image—and text-based state-of-the-art models in classification and developed an efficient image—and text-based multimodal framework.

Puppala et al. (2021) classified music into ten different genres by extracting the Mel Frequency Cepstral constant (MFCC) value in music as a feature and using a convolutional neural network to train the model. The final model training accuracy rate is as high as 97%.

Besides the game domain, genre-based research has also been conducted in domains like books and movies. Biradar et al. (2019) used book covers and titles as features for genre classification. The authors used a logistic regression model for training and prediction. In the end, the accuracy of using both the cover and the title as features was 87.2%.

Chu and Guo (2017) propose to implement movie genre classification based only on movie poster images. A deep neural network is constructed to jointly describe visual appearance and object information and classify a given movie poster image into genres. Simoes et al. (2016) proposed a novel deep neural architecture based on Convolutional Neural Networks (ConvNets) to perform multi-label movie trailer genre classification. It encapsulates an ultra-deep ConvNet with residual connections and utilizes special convolutional layers to extract temporal information from image-based features before performing movie trailer-to-genre mapping.

Sentiment Classification

Since the problem of game recommendation resembles the problem of sentiment classification, techniques used in sentiment classification are important baselines and references, especially in the identification of algorithms and feature selection methods. Hadwan et al. (2022) proposed an improved sentiment classification method that used machine learning methods with feature engineering and SMOTE techniques to measure user satisfaction with government service mobile applications. The authors used a total of six feature engineering methods and five machine learning algorithms. Among them, the NB classifier outperformed the other models (74.25% accuracy) when using the original Arb-Apps Review dataset and extracting features by BOW.

Chakraborty et al. (2021) researched the Bangla dataset, predicted text polarities (positive, neutral, negative) using N-Gram as feature extraction, and two ML algorithms (SVM, Random Forest). The results showed that SVM outperformed, especially with unigram, achieving a 68% F1 score.

Shahzad et al. (2022) proposed a framework for analyzing global perceptions and attitudes towards the COVID-19 vaccines AstraZeneca, Pfizer, Sinovac, Moderna, and Sinopharm, respectively. Different machine learning classifiers, such as Random Forest (RF), Parsimonious Bayes (NB), Decision Trees (DT), Logistic Regression (LR) and Support Vector Machines (SVM), were used. DT performed the best results on five different datasets.

A review on exploring the impact of embedding models on text analysis was published (Asudani et al., 2023). It introduces a number of word representation approaches, including traditional methods such as the bag-of-words model, n-gram, TF-IDF, and word embedding models Word2Vec, GloVe, and so on. It also shows the performance of these word representation models in different tasks. They explored the performance of word embeddings in deep learning models in text analysis tasks and concluded that combining domain-specific word embeddings and LSTM models can improve the performance of text analysis tasks.

Based on the literature review, many achievements have been captured in the field of genre classification (Puppala et al., 2021). However, genre-based research is lacking, especially genre-based recommender systems. Genre is important in differentiating the characteristics or types of the domain's objects. Hence, the genre-based model is often used to capture specific characteristics. Therefore, we explore whether recommendation models in the gaming domain can benefit from a genre-based setting.

Since the feature selection method can also affect the model's accuracy, this article chooses three common feature selection approaches in the field of text classification: BOW, N-Gram and POS. BOW explores the effect of individual words on the model, N-Gram explores the effect of sequences on the model, and POS explores the effect of actions or sentiments on the model.

As for machine learning algorithms, previous studies provide mixed results. According to Hadwan et al. (2022), NB performed best when BOW was the feature selection method, so NB was also chosen. SVM was chosen because it performed better compared to LR and RF in the game dataset (Britto & Pac´ıfico, 2020). DT was chosen because Zuo (2018) used both NB and DT in his method and performed better. In similar tasks, NB, SVM and DT algorithms have achieved better results than other machine learning algorithms, such as random forest and logistic regression. Therefore, we have chosen NB, SVM and DT as the algorithms for machine learning.

In contrast to previous work, our study will systematically compare multiple feature selection methods and classification algorithms to construct recommendation models in the gaming domain. Figure 3 summarizes the related literature, which shows the gaps in research and highlights the methodology that is the focus of this paper. This approach addresses the identified gaps and provides a robust analysis of genre-based recommendations, thereby advancing the field.

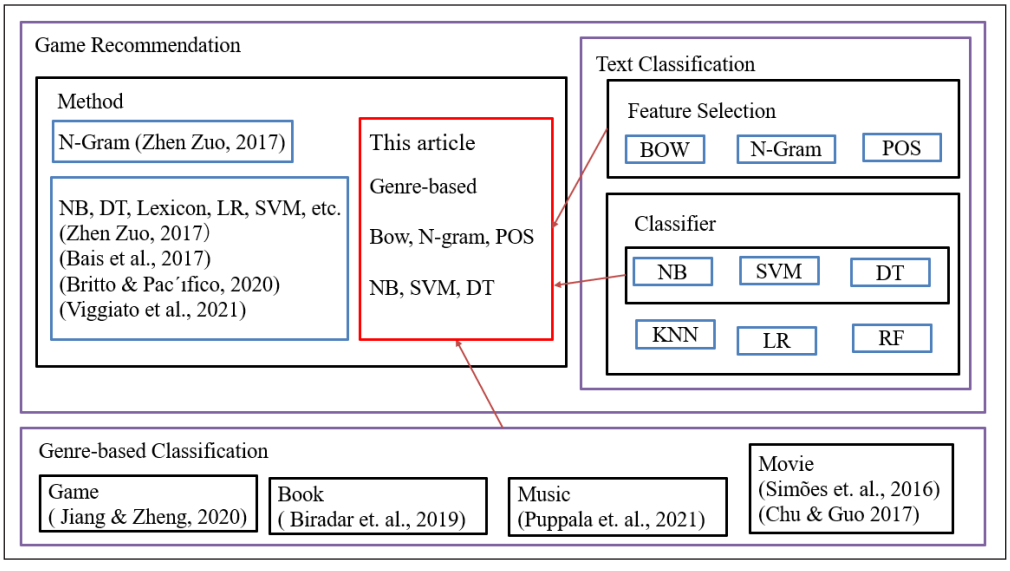

Figure 3. Literature gap in game recommendation

METHODS

A classification model is created to achieve the classification goal on the game review dataset. Firstly, games are classified by genre, and the dataset is split accordingly. Various text preprocessing steps are applied based on different feature extraction methods: Bag of Words (Punctuation removal, lowercase, lemmatization, and stop word removal), N-Gram (Punctuation removal, lowercase, and lemmatization), Part of Speech (Punctuation removal and lowercase).

After text preprocessing, the respective feature extraction processes are carried out. Extracted features are then fed into various classifier models, including Naive Bayes, Support Vector Machines, and Decision Trees.

Subsequently, each model's performance is evaluated in terms of accuracy, precision, recall, and F1 score. The models for each game genre are compared, and the best-performing model is chosen. Figure 3 is the framework of the model.

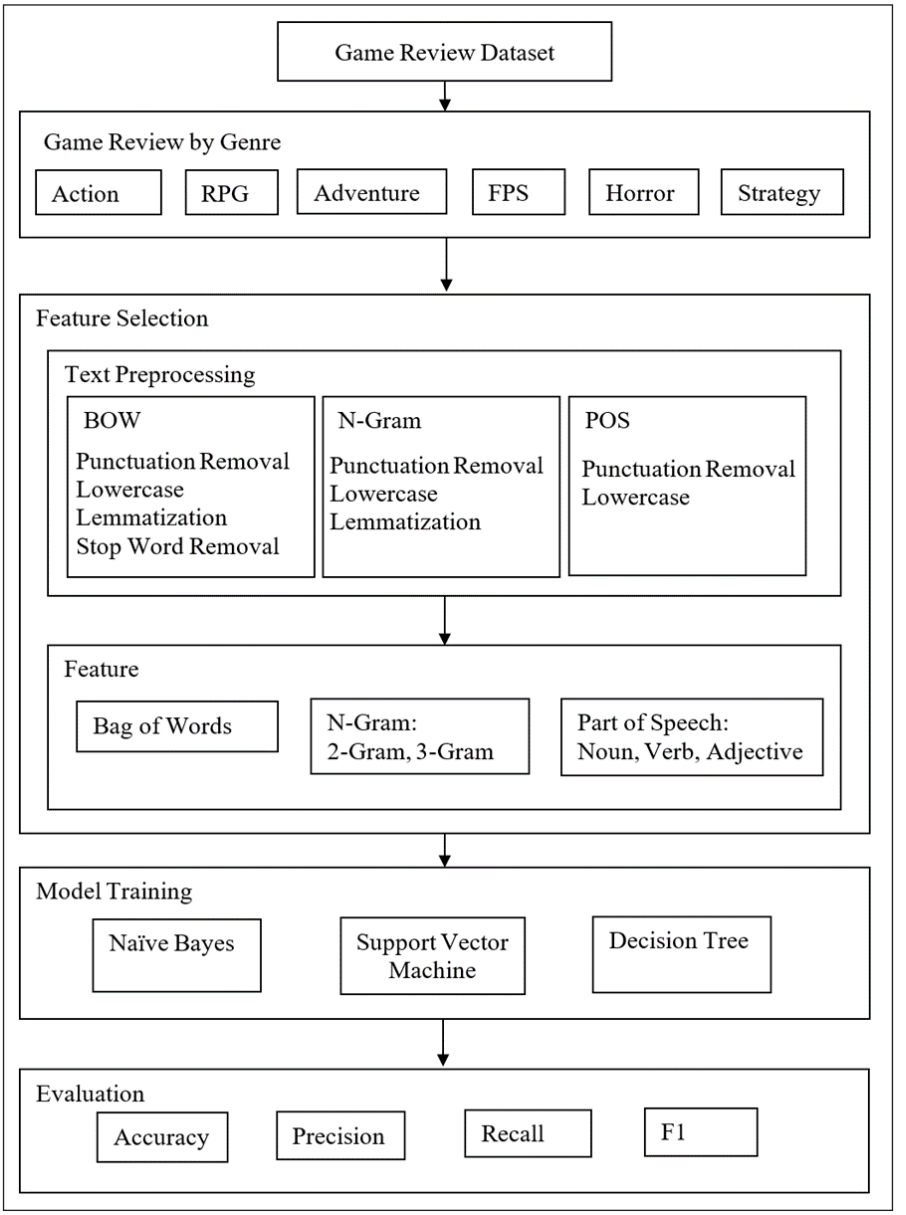

Figure 4. The framework of the genre-based game recommendation model

Dataset

This dataset is from Kaggle. The link is https://www.kaggle.com/datasets/arashnic/game-review-dataset, and reviews originate from the Steam gaming platform. The dataset includes 44 games, encompassing various genres, resulting in 157 different game genres represented.

For the sake of practicality in conducting experiments, six specific game genres were selected for analysis. These genres were chosen because all 44 games included at least one of these six genres, with many games falling under two of these categories. The six selected game genres are as follows: Action (Act), Role-playing game (RPG), Adventure (Adv), First-person shooting game (FPS), Horror (Hor) and Strategy (Str).

Upon categorizing the games into these genres, the dataset provides information for all genres collectively and for each genre (Table 1). It's worth noting that since many games belong to two genres simultaneously, the total number of data sets for these six game genres will exceed the original number of data sets in the dataset.

Table 1
Data set information of all genres and each genre

Genre	Number of games	Number of reviews	Recommended	Not recommended
All	44	17436	9933 (57.0%)	7503 (43.0%)
Act	28	10852	6871 (63.3%)	3981 (36.7%)
RPG	15	4681	2802 (59.9%)	1879 (40.1%)
Adv	11	4916	2435 (49.5%)	2481 (50.5%)
FPS	12	5330	3051 (57.2%)	2279 (42.8%)
Hor	5	1747	845 (48.4%)	902 (51.6%)
Str	19	6496	3324 (51.2%)	3172 (48.8%)

Text Preprocessing

Text preprocessing is a series of processing of source data to extract more useful features. Here is how the text was preprocessed for this article.

Punctuation removal : Strip punctuation from reviews so that a space precedes each word.

Stop word removal : Stop words are words that appear frequently in the text and are of little value in indicating the content of the text. Eliminating some meaningless and ineffective words can avoid the interference caused by these words.

Lowercase : Convert all letters to lowercase, aiming to improve feature extraction so that words with different cases are not treated as different.

Lemmatization : Lemmatization is the reduction of a language vocabulary of any form to its general form. It will remove the word's affixes and extract the word's main part.

Feature Selection

Feature selection is a critical process aimed at choosing the most relevant features from the original set of features to reduce data dimensionality and enhance the performance of machine learning algorithms. For the task of this article, the feature selection methods of word embedding, such as Word2Vec and GloVe, are not used. It requires a large amount of data and computational resources to train and generate vectors dependent on pre-trained data. Meanwhile, traditional feature extraction approaches are simpler and computationally efficient for tasks with small data. Therefore, the following three feature selection methods are used in this article. They are Bag of Words, N-Gram and Part-of-Speech, respectively.

Bag-of-words is a basic and well-known text representation technique (Deniz et al., 2021). Each word of the bag-of-words model occurs independently in the text. The frequency of all words is counted, and a part of the words with the highest frequency is selected as the word bag. BOW can investigate how individual words impact the model. BOW is widely used for tasks such as text analysis and sentiment classification because it captures the word frequency information in the text very well and is easy to compute. It can be used as input to various machine learning algorithms.

N-Gram is an algorithm based on statistical language models (Cavnar & Trenkle, 1994). N-Gram is enough to capture more semantic information and relationships between words and can delve into how word sequences affect the model. In this article, 2-gram and 3-gram will be chosen.

Part-of-speech tagging involves classifying and tagging words in a sentence, assigning part-of-speech tags to each word based on its grammatical role. After tagging, this article will extract verb phrases, nouns, and adjective phrases from the text as features. POS tagging can capture actions through verb phrases and sentiments through adjective phrases. These features related to actions and sentiments can be leveraged to explore their impact on the model. It can also reduce noise since words with specific lexical properties may be more important than others. Verb phrases and adjective phrases have specific extraction rules, as outlined in Table 2.

Table 2
Extraction rules with POS

Phrases	**Extraction rules**
Verb phrases	verb + preposition
	verb + adverb
	verb + adverb + preposition
Adjective phrases	adjective + coordinating conjunction + adjective
	adjective + adverb
	adverb + adjective

After extracting the corresponding phrases, the frequency of their occurrence was calculated, and the part with the highest frequency was used as a feature. Comparing BOW, N-Gram and POS, machine learning models can capture textual features at different dimensions, thus exploring the ability of the models to understand and process textual data when using different feature selection methods.

Classification Algorithm

Three classification algorithms are used in this article, namely Naive Bayes (Raschka, 2014), Support Vector Machine (Gaspar et al., 2012) and Decision Tree (Patel & Prajapati, 2018). NB is based on Bayes' theorem and assumes that features are independent of each other. It is computationally simple, fast in training and prediction, and suitable for large-scale text data. SVM has strong generalization ability, can effectively avoid overfitting, and performs well in dealing with binary classification problems. DT is good at capturing the nonlinear relationship between features and categories and is suitable for complex sentiment classification tasks. Combining the advantages of these three classifiers and their excellent performance in related work, they were considered suitable for this study and were finally selected as the classification algorithm for this article.

Experiment Setup

The dataset for each game genre will be split into a training set and a test set, with a ratio of 8 to 2. Following this division, the parameters for feature extraction will be explored. Given the substantial size of the dataset, only the most frequently occurring features will be extracted. Different parameter values will be tested for various feature selection methods. In this context, the parameter refers to the number of features with the highest frequency ranking.

These parameters will be established using all data sets for the various game genres. The experimental results from this comprehensive analysis will serve as a reference for standardizing the parameters for the models of each game genre. Their parameters will be set to be identical to ensure a fair comparison between classification algorithms.

Ultimately, the parameters for each feature selection method will be determined separately by selecting the values that yield the highest average accuracy across the three models. The experiments will lead to the selection of the most suitable parameters, which are summarized in Table 3.

After confirming feature extraction parameters, a baseline was specified. In this article, the results of the model trained with all game genres as the data set are selected as the baseline. Not only was the model trained on all game genres on all datasets tested, but the datasets for each genre were also tested separately. This result was used to compare with the model of each game genre. The model evaluation metrics are accuracy (A), precision (P), recall (R), and F1 (F) score.

Table 3
Parameters for each feature selection method

Feature selection method	Parameter
Bag of Words	1500
2-Gram	3000
3-Gram	4000
Part of Speech	2000

RESULTS

Statistical Significance Tests

Before the experiment, we did statistical significance tests to confirm that the three machine learning algorithms are different. Here, all genre datasets are used. The method used is a t-test, and since a t-test can only validate two models at a time, we did three t-test validations. We set the alpha value to 0.05 for the experiment and used 5×2 cross-validation to validate. That is, we repeated the 2-fold cross-validation five times. The results of the experiment are shown in Table 4.

Table 4
Results of statistical significance tests

Models for comparison	NB/SVM	NB/DT	SVM/DT
p-value	0.004	<0.001	<0.001
Result	Reject	Reject	Reject

Based on each of the results above, the null hypothesis was rejected. Therefore, the three models show variability, proving that the experiment is feasible.

Results for Bag of Words

When BOW is used as a feature selection method, the accuracy results are shown in Table 5.

Table 5 shows that using BOW as the feature selection method generally yields higher accuracy with the NB classification algorithm. Additionally, when training NB and DT models based on game genres, most accuracy rates surpass those of models trained on all genres. Furthermore, performance is suboptimal when applying the all-genres model to individual genre datasets. Precision, recall and F1 results are detailed in Table 6.

Table 6 shows that when BOW is used as the feature selection method, the precision, recall and F1 of NB as the classification algorithm are generally higher. The Horror genre NB model's precision value is much higher than that of all the other genres, reaching 0.883. The highest recall is in the Adventure genre, reaching 0.844. The highest F1 scores were observed for NB in the Adventure and FPS genres, both at 0.831.

Table 5
Accuracy of all and genre-based models with BOW

Target & Model \ Classifier		NB	SVM	DT
All	All	0.828	0.822	0.688
Act	Act	0.851	0.802	0.719
	All	0.564	0.523	0.514
RPG	RPG	0.812	0.732	0.693
	All	0.488	0.522	0.509
Adv	Adv	0.832	0.804	0.669
	All	0.543	0.567	0.508
FPS	FPS	0.854	0.776	0.698
	All	0.519	0.520	0.468
Hor	Hor	0.797	0.697	0.720
	All	0.506	0.529	0.471
Str	Str	0.832	0.766	0.676
	All	0.535	0.520	0.528

Table 6
Precision, recall and F1 of all and genre-based models with BOW

Target & Model \ Classifier		NB			SVM			DT		
		P	R	F	P	R	F	P	R	F
All	All	0.822	0.786	0.804	0.783	0.797	0.790	0.627	0.637	0.632
Act	Act	0.820	0.793	0.806	0.736	0.739	0.737	0.630	0.628	0.629
	All	0.639	0.446	0.526	0.585	0.409	0.481	0.551	0.397	0.461
RPG	RPG	0.742	0.751	0.746	0.676	0.631	0.653	0.625	0.581	0.602
	All	0.585	0.378	0.459	0.421	0.374	0.396	0.533	0.385	0.447
Adv	Adv	0.818	0.844	0.831	0.788	0.816	0.802	0.646	0.679	0.663
	All	0.572	0.543	0.557	0.620	0.563	0.590	0.588	0.510	0.546
FPS	FPS	0.833	0.829	0.831	0.738	0.742	0.740	0.664	0.647	0.655
	All	0.538	0.452	0.492	0.592	0.457	0.516	0.538	0.411	0.467
Hor	Hor	0.883	0.748	0.810	0.620	0.721	0.667	0.690	0.724	0.707
	All	0.895	0.497	0.639	0.725	0.512	0.600	0.772	0.475	0.588
Str	Str	0.797	0.842	0.819	0.720	0.775	0.747	0.653	0.664	0.659
	All	0.248	0.531	0.338	0.178	0.496	0.262	0.569	0.506	0.536

Results for 2-Gram

When 2-gram is used as a feature selection method, the accuracy results are shown in Table 7.

Table 7
Accuracy of all and genre-based models with 2-gram

Target & Model \ Classifier		NB	SVM	DT
All	All	0.816	0.764	0.663
Act	Act	0.825	0.756	0.684
	All	0.513	0.550	0.491
RPG	RPG	0.819	0.724	0.655
	All	0.505	0.472	0.508
Adv	Adv	0.826	0.725	0.691
	All	0.535	0.504	0.510
FPS	FPS	0.839	0.721	0.662
	All	0.479	0.500	0.492
Hor	Hor	0.757	0.694	0.697
	All	0.503	0.469	0.551
Stra	Stra	0.817	0.758	0.664
	All	0.514	0.520	0.505

Table 8
Precision, recall and F1 of all and genre-based models with 2-gram

Target & Model \ Classifier		NB			SVM			DT		
		P	R	F	P	R	F	P	R	F
All	All	0.816	0.768	0.791	0.733	0.721	0.727	0.601	0.608	0.605
Act	Act	0.786	0.760	0.772	0.670	0.680	0.675	0.549	0.589	0.568
	All	0.563	0.398	0.466	0.393	0.403	0.398	0.683	0.399	0.504
RPG	RPG	0.785	0.743	0.763	0.605	0.636	0.620	0.590	0.534	0.561
	All	0.542	0.383	0.449	0.734	0.389	0.508	0.665	0.403	0.502
Adv	Adv	0.861	0.807	0.833	0.679	0.750	0.713	0.687	0.695	0.691
	All	0.527	0.538	0.533	0.485	0.504	0.496	0.572	0.512	0.540
FPS	FPS	0.779	0.837	0.807	0.644	0.691	0.667	0.592	0.613	0.603
	All	0.445	0.407	0.425	0.371	0.413	0.391	0.521	0.429	0.470
Hor	Hor	0.819	0.722	0.767	0.760	0.663	0.708	0.725	0.678	0.701
	All	0.485	0.491	0.488	0.474	0.458	0.466	0.620	0.535	0.575
Str	Str	0.783	0.825	0.804	0.717	0.762	0.739	0.658	0.646	0.652
	All	0.465	0.491	0.478	0.318	0.497	0.388	0.579	0.485	0.528

From Table 7, when 2-Gram is employed as the feature selection method, the accuracy of SVM and DT models for each game genre fluctuates, sometimes surpassing and other

times falling below the accuracy of the all-genres model. Conversely, most NB genre-specific models achieve higher accuracy than the all-genres models. The highest accuracy value was for the FPS genre NB model, which exceeded the value for all genres of models, reaching 0.839. Table 8 shows the precision, recall and F1 results.

It can be seen from Table 8 that when 2-Gram is used as the feature selection method, the precision, recall, and F1 of each game genre model of NB, SVM, and DT fluctuates up and down the value of all game genre models. Among them, the precision of the NB model of Adventure reached 0.861. The recall of the NB model of FPS reached 0.837, which is the highest. The F1 of the NB model of Adventure reached 0.833, which is the highest.

Results for 3-Gram

When 3-gram is used as a feature selection method, the accuracy results are shown in Table 9.

Table 9
Accuracy of all and genre-based models with 3-gram

Target & Model \ Classifier		NB	SVM	DT
All	All	0.779	0.696	0.659
Act	Act	0.792	0.704	0.656
	All	0.514	0.532	0.481
RPG	RPG	0.761	0.670	0.670
	All	0.540	0.489	0.475
Adv	Adv	0.762	0.646	0.618
	All	0.552	0.558	0.494
FPS	FPS	0.771	0.682	0.639
	All	0.516	0.508	0.482
Hor	Hor	0.723	0.637	0.597
	All	0.489	0.500	0.449
Str	Str	0.768	0.664	0.633
	All	0.502	0.497	0.477

It can be seen from Table 9 that when 3-Gram is used as the feature selection method, the accuracy of the models for each game genre for NB, SVM and DT fluctuates around the value of the models for all game genres. The highest accuracy was achieved by the Action Genre NB model with 0.792. In addition, when the models trained by all game genres are put into the data sets of each genre, the models of each genre do not perform well. Next, precision, recall and F1 results are in Table 10.

Table 10
Precision, recall and F1 of all and genre-based models with 3-gram

Target & Model \ Classifier		NB			SVM			DT		
		P	R	F	P	R	F	P	R	F
All	All	0.770	0.729	0.749	0.649	0.643	0.646	0.606	0.601	0.604
Act	Act	0.724	0.725	0.724	0.618	0.606	0.612	0.619	0.539	0.576
	All	0.471	0.384	0.423	0.424	0.391	0.407	0.486	0.361	0.415
RPG	RPG	0.702	0.671	0.686	0.539	0.560	0.549	0.570	0.556	0.563
	All	0.539	0.410	0.466	0.473	0.358	0.407	0.484	0.351	0.407
Adv	Adv	0.794	0.749	0.771	0.574	0.675	0.620	0.529	0.647	0.582
	All	0.505	0.561	0.531	0.479	0.572	0.521	0.481	0.497	0.489
FPS	FPS	0.664	0.775	0.715	0.564	0.653	0.605	0.566	0.585	0.576
	All	0.449	0.441	0.445	0.456	0.435	0.445	0.527	0.421	0.468
Hor	Hor	0.789	0.689	0.736	0.596	0.638	0.616	0.649	0.578	0.612
	All	0.538	0.479	0.507	0.520	0.489	0.504	0.485	0.441	0.462
Str	Str	0.732	0.771	0.751	0.627	0.655	0.641	0.598	0.621	0.609
	All	0.482	0.479	0.481	0.511	0.476	0.493	0.474	0.455	0.465

Table 10 shows that when using the 3-Gram as the feature selection method, the precision, recall and F1 of the models for each game genre for NB, SVM and DT fluctuate around the value of the models for all game genres. The highest precision was for the Adventure genre NB model, which reached 0.794. The highest recall was for the FPS-type NB model, which reached 0.775. Among them, the highest F1 is the Adventure genre NB model, reaching 0.771.

Results for Part of Speech

The experiment results are shown in Table 11 according to the feature extraction rules.

It can be seen from Table 11 that when POS is used as the feature selection method, the accuracy of the models of each game genre of NB and DT fluctuates up and down the value of the model of all game genres. Among them, the NB model with the highest accuracy is the FPS game genre, reaching 0.827. In addition, when the models trained by all game genres are put into the data sets of each genre, the models of each genre do not perform well. Table 12 shows the precision, recall and F1 results.

From Table 12, when POS is used as the feature selection method, the precision, recall, and F1 of each game genre model of NB, SVM, and DT fluctuate up and down in value. The highest precision is the NB model of the Horror game genre, reaching 0.930. The Adventure genre NB model has the highest recall and F1, 0.808 and 0.821.

Table 11
Accuracy of all and genre-based models with POS

Target & Model \ Classifier		NB	SVM	DT
All	All	0.812	0.775	0.645
Act	Act	0.822	0.767	0.668
	All	0.500	0.569	0.555
RPG	RPG	0.785	0.719	0.642
	All	0.493	0.549	0.485
Adv	Adv	0.817	0.737	0.634
	All	0.503	0.505	0.480
FPS	FPS	0.827	0.751	0.628
	All	0.508	0.518	0.534
Hor	Hor	0.734	0.663	0.614
	All	0.486	0.483	0.494
Str	Str	0.805	0.737	0.639
	All	0.477	0.529	0.498

Table 12
Precision, recall and F1 of all and genre-based models with POS

Target & Model \ Classifier		NB			SVM			DT		
		P	R	F	P	R	F	P	R	F
All	All	0.814	0.762	0.788	0.717	0.746	0.732	0.575	0.587	0.581
Act	Act	0.783	0.755	0.769	0.682	0.696	0.689	0.516	0.567	0.540
	All	0.839	0.419	0.559	0.463	0.434	0.448	0.523	0.428	0.470
RPG	RPG	0.756	0.695	0.724	0.659	0.615	0.636	0.504	0.521	0.512
	All	0.621	0.388	0.477	0.536	0.417	0.469	0.570	0.374	0.452
Adv	Adv	0.834	0.808	0.821	0.705	0.755	0.729	0.624	0.640	0.632
	All	0.705	0.504	0.588	0.675	0.506	0.578	0.471	0.482	0.476
FPS	FPS	0.796	0.803	0.800	0.720	0.709	0.715	0.555	0.571	0.563
	All	0.696	0.455	0.550	0.479	0.446	0.462	0.358	0.451	0.399
Hor	Hor	0.930	0.663	0.774	0.649	0.657	0.653	0.620	0.602	0.611
	All	0.784	0.484	0.598	0.532	0.474	0.501	0.234	0.465	0.311
Str	Str	0.780	0.807	0.793	0.740	0.719	0.729	0.630	0.618	0.624
	All	0.831	0.473	0.603	0.693	0.506	0.585	0.526	0.478	0.501

DISCUSSION

Based on the results presented in this study, it's clear that models trained on datasets encompassing all game genres perform reasonably well when applied to the same genre of dataset. However, their performance significantly lagged behind models trained specifically in individual game genres. The primary reason for this disparity lies in feature extraction. Models trained in all genres extract features that encompass elements from all six game genres, whereas models trained in a specific genre extract feature tailored to that genre alone. Consequently, the performance gap becomes substantial when both models are applied to the same genre-specific dataset.

For instance, when both models are applied to the Action dataset with identical test data, the models trained using all game-genre datasets (using BOW and NB) achieve an accuracy, precision, recall, and F1 score of 0.564, 0.639, 0.446, and 0.526, respectively. In contrast, the models trained using the Action dataset achieve much higher results, with an accuracy of 0.851, precision of 0.820, recall of 0.793, and F1 score of 0.806. This illustrates that models tailored to specific game genres significantly outperform those trained in all genres.

Table 13 summarizes all the above results and provides a concise overview of the highest accuracy, precision, recall, and F1 for different combinations of feature extraction methods and classifiers for each game genre. Figure 4 illustrates the results in a bar chart, providing a more visual view of the gap between each result.

Referring to Table 13 and Figure 5, it's apparent that, for most game genres, the combination of BOW and NB achieved the best performance, followed by 2-Gram and NB. Also, models based on specific game genres generally perform better than those trained in all genres. However, there are still individual game genres where the results of the models are not as good as the models trained on all game genres, and it is possible that this occurs due to an insufficient dataset, which prevents the models from being adequately trained. Notably, the FPS game genre models outperform models of all genres by a significant margin. However, some genres like horror and RPG have relatively lower performance, possibly due to their small dataset sizes, hindering the training of more robust models.

In summary, Table 13 reveals that when employing the combination of BOW and NB, the accuracy of the all-genres model reaches 0.828. However, most genre-based models surpass their accuracy, including FPS, Action, Adventure, and Strategy, which achieve accuracy values of 0.854, 0.851, 0.832, and 0.832, respectively. It implies that genre-specific models consistently outperform the all-genres model in accuracy. This result is an improvement over previous similar tasks in the literature, e.g., (Zuo, 2018) (accuracy: 0.75), (Britto & Pac´ıfico, 2020) (accuracy: 0.825).

The Action game genre models were used as a reference for text features, and Table 14 summarizes the experimental results.

Table 13
Summary of all top-performing models

Genre	Accuracy		Precision		Recall		F1	
	Value	Feature & Classifier	Value	Feature & Classifier	Value	Feature & Classifier	Value	Feature & Classifier
Act	0.851	BOW&NB	0.820	BOW&NB	0.793	BOW&NB	0.806	BOW&NB
RPG	0.819	2-G&NB	0.785	2-G&NB	0.751	BOW&NB	0.763	2-G&NB
Adv	0.832	BOW&NB	0.861	2-G&NB	0.844	BOW&NB	0.833	2-G&NB
FPS	0.854	BOW&NB	0.833	BOW&NB	0.837	2-G&NB	0.831	BOW&NB
Hor	0.797	BOW&NB	0.930	POS&NB	0.748	BOW&NB	0.810	BOW&NB
Str	0.832	BOW&NB	0.797	BOW&NB	0.842	BOW&NB	0.819	BOW&NB
All	0.828	BOW&NB	0.822	BOW&NB	0.797	BOW&SVM	0.804	BOW&NB

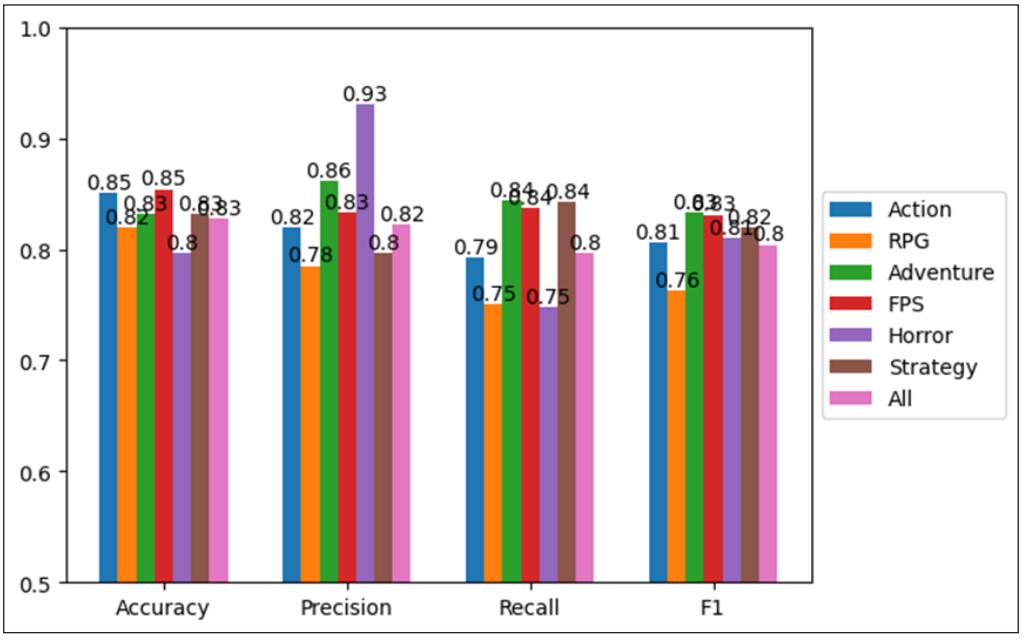

Figure 5. Results for all top-performing models

As indicated in Table 14, within the Action game genre, each classifier delivers the best performance when employing BOW as the feature selection method. They consistently outperform 2-Gram, 3-Gram, and POS across all evaluated metrics. Consequently, it can be inferred that BOW is the most effective among the three feature selection methods.

The subsequent step is to determine the best-performing classification algorithm. Table 15 summarizes the results for the Strategy game genre.

Table 14
Summary of results for the Action game genre

Classifier	Feature	Accuracy	Precision	Recall	F1
NB	BOW	0.851	0.820	0.793	0.806
	2-Gram	0.825	0.786	0.760	0.772
	3-Gram	0.792	0.724	0.725	0.724
	POS	0.822	0.783	0.755	0.769
SVM	BOW	0.802	0.736	0.739	0.737
	2-Gram	0.756	0.670	0.680	0.675
	3-Gram	0.704	0.618	0.606	0.612
	POS	0.767	0.682	0.696	0.689
DT	BOW	0.719	0.630	0.628	0.629
	2-Gram	0.684	0.549	0.589	0.568
	3-Gram	0.656	0.619	0.539	0.576
	POS	0.668	0.516	0.567	0.540

Table 15
Summary of results for the Strategy game genre

Feature	Classifier	Accuracy	Precision	Recall	F1
BOW	NB	0.832	0.797	0.842	0.819
	SVM	0.766	0.720	0.775	0.747
	DT	0.676	0.653	0.664	0.659
2-Gram	NB	0.817	0.783	0.825	0.804
	SVM	0.758	0.717	0.762	0.739
	DT	0.664	0.658	0.646	0.652
3-Gram	NB	0.768	0.732	0.771	0.751
	SVM	0.664	0.627	0.655	0.641
	DT	0.633	0.598	0.621	0.609
POS	NB	0.805	0.780	0.807	0.793
	SVM	0.737	0.740	0.719	0.729
	DT	0.637	0.630	0.618	0.624

As evident from Table 15, among the Strategy game genre models, NB consistently outperforms both SVM and DT in all evaluation metrics when feature selection methods are consistent. This trend is also observed across the other game genre models, with NB emerging as the best-performing classification algorithm. The model performance is notably enhanced when the most effective feature selection method is paired with the top-performing classifier.

The above results may be because BOW can learn valid sentiment tokens from a large amount of data very well, and it does not focus on word order and structural information, making it robust to certain noisy data. In addition, NB assumes conditional independence

between features, an assumption that often provides a valid approximation in sentiment categorization since many sentiment words independently reflect sentiment tendencies. It can handle high-dimensional data well, whereas the features of game reviews are usually high-dimensional. Therefore, the results of combining BOW and NB outperform other combinations.

After the above experiments, we chose the most accurate FPS genre for prediction, and the combination of the models was BOW and NB. We selected 10 game reviews from Twitter without labels. After the prediction was completed and manually verified, 8 out of 10 predicted reviews were correct. The accuracy is relatively high. The model can help users quickly realize the number of positive and negative reviews of a game and thus help them decide whether to play.

The model still exists with some limitations. The current model can only show recommendations and non-recommendations but no justification for the reason for the recommendation. A more specific recommendation can be provided by supporting the decision recommendation with evidence such as phrases related to the recommendation or otherwise. For example, aspect and evidence extraction can be carried out when a game called "Spooky" is recommended. For example, because the music is so scary, it can be extracted as music and scary. Players can determine whether their needs are fulfilled based on the words extracted. The dataset used for training is relatively small and has a limited variety of game genres, with only six considered. There are multiple genres for each game, and it is not possible to consider all of them. It is possible that as games are updated or new games appear, new features or genres will appear, which will require the model to be constantly updated. This article also discusses only three feature extraction methods and classification algorithms. Better results could be obtained by exploring more feature selection approaches or more machine school models.

CONCLUSION

In this article, we introduce a game genre-based model for assessing the recommendation of reviews across different game genres. Our approach involves identifying the genre of the game and then experimenting with various feature extraction methods and classification algorithms to create an effective model. We compare three feature extraction techniques, Bag of Words (BOW), N-Gram, and Part of Speech (POS), in conjunction with three classification algorithms: Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT), selecting the most suitable combination for each game genre.

Our evaluation results reveal the optimal models for each game genre. BOW emerges as the top-performing text feature selection method, while NB stands out as the best-performing classification algorithm. The most effective combination is BOW and NB. In comparison to the accuracy of the model encompassing all genres (0.828), our proposed

models achieve improved accuracy rates, specifically 0.854 for FPS, 0.851 for Action, and 0.832 for both Adventure and Strategy.

The player can figure out the game's recommendation rate after the model determines whether other players recommend it. If there are 100 player reviews and the model results in 80 recommendations, then the game has an 80 percent positive rating, which can greatly help other players in their decision-making and can also help the game developer monitor whether the game is successful or not.

By demonstrating the effectiveness of specific model genres, we emphasize the importance of tailored approaches in improving recommendation accuracy. Our approach provides new perspectives on the combination of feature extraction and classification algorithms and sets a precedent for future research exploration and improvement.

More datasets will be collected, and more game genres will be involved in model training to overcome the limitations. More feature extraction methods will be evaluated, such as feature extraction by TD-IDF, word embedding, and some optimization algorithms to find the right features, such as the Reptile Search Algorithm (RSA) (Abualigah et al., 2022). After that, deep learning models such as CNN and BiLSTM can be combined. Aspects of sentiment analysis will also be considered, listing the strengths and weaknesses of each game in order to be more helpful to players and game developers. For example, in one review, the action is smooth, but the plot is confusing. Through aspect sentiment analysis, it can be extracted that its strength is the action, and its weakness is the plot.

ACKNOWLEDGEMENT

The authors would like to thank the School of Computer Sciences, Universiti Sains Malaysia, for their support.

REFERENCES

Abualigah, L., Elaziz, M. A., Sumari, P., Geem, Z. W., & Gandomi, A. H. (2022). Reptile Search Algorithm (RSA): A nature-inspired meta-heuristic optimizer. *Expert Systems with Applications, 191*, Article 116158. https://doi.org/10.1016/j.eswa.2021.116158

Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: A review. *Artificial Intelligence Review*, *56*(9), 10345-10425. https://doi.org/10.1007/s10462-023-10419-1

Bais, R., Odek, P., & Ou, S. (2017). *Sentiment Classification on Steam Reviews*. Semantic Sholar.

Biradar, G. R., Raagini, J., Varier, A., & Sudhir, M. (2019, June 29-30). *Classification of book genres using book cover and title*. [Paper presentation]. IEEE International Conference on Intelligent Systems and Green Technology (ICISGT), Visakhapatnam, India. https://doi.org/10.1109/ICISGT44072.2019.00031

Britto, L. F., & Pacifico, L. (2020, November 7-10). *Evaluating video game acceptance in game reviews using sentiment analysis techniques*. [Paper presentation]. Proceedings of SBGames, Recife, Brazil.

Cavnar, W. B., & Trenkle, J. M. (1994, April 11-13). *N-gram-based text categorization*. [Paper presentation]. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, U.S.A.

Chakraborty, P., Nawar, F., & Chowdhury, H. A. (2021). A ternary sentiment classification of Bangla text data using support vector machine and random forest classifier. In J. K. Mandal, P. Hsiung & R. S. Dhar (Eds.), *Topical Drifts in Intelligent Computing* (pp. 69-77). Springer.

Gaspar, P., Carbonell, J., & Oliveira, J. L. (2012). On the parameter optimization of support vector machines for binary classification. *Journal of Integrative Bioinformatics, 9*(3), 33–43. https://doi.org/10.1515/jib-2012-201

Chu, W. T., & Guo, H. J. (2017, October). Movie genre classification based on poster images with deep neural networks. In *proceedings of the workshop on multimodal understanding of social, affective and subjective attributes* (pp. 39-45). ACM Publishing. https://doi.org/10.1145/3132515.3132516

Deniz, A., Angin, M., & Angin, P. (2021). Evolutionary multiobjective feature selection for sentiment analysis. *IEEE Access, 9*, 142982–142996. https://doi.org/10.1109/ACCESS.2021.3118961

Hadwan, M., Al-Sarem, M., Saeed, F., & Al-Hagery, M. A. (2022). An improved sentiment classification approach for measuring user satisfaction toward governmental services' mobile apps using machine learning methods with feature engineering and SMOTE technique. *Applied Sciences*, *12*(11), Article 5547. https://doi.org/10.3390/app12115547

Jiang, Y., & Zheng, L. (2020). Deep learning for video game genre classification. *Multimedia Tools and Applications, 82*(14), 21085-21099. https://doi.org/10.1007/s11042-023-14560-5

Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, *6*(10), 74–78.

Puppala, L. K., Muvva, S. S. R., Chinige, S. R., & Rajendran, P. S. (2021, July 8-10). *A novel music genre classification using convolutional neural network*. [Paper presentation]. 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India. https://doi.org/10.1109/ICCES51350.2021.9489022

Raschka, S. (2014). *Naive bayes and text classification i-introduction and theory.* ArXiv Preprint. https://doi.org/10.48550/arXiv.1410.5329

Shahzad, A., Zafar, B., Ali, N., Jamil, U., Alghadhban, A. J., Assam, M., Ghamry, N. A., & Eldin, E. T. (2022). COVID-19 vaccines related user's response categorization using machine learning techniques. *Computation*, *10*(8), Article 141. https://doi.org/10.3390/computation10080141

Simoes, G. S., Wehrmann, J., Barros, R. C., & Ruiz, D. D. (2016, July 24-29). *Movie genre classification with convolutional neural networks*. [Paper presentation]. International Joint Conference on Neural Networks (IJCNN), Vancouver, Canada. https://doi.org/10.1109/IJCNN.2016.7727207

Viggiato, M., Lin, D., Hindle, A., & Bezemer, C. P. (2021). What causes wrong sentiment classifications of game reviews? *IEEE Transactions on Games*, *14*(3), 350–363. https://doi.org/10.1109/TG.2021.3072545

Zuo, Z. (2018). *Sentiment analysis of steam review datasets using naive bayes and decision tree classifier*. Ideals.