Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 6, No. 4, November 2021, Pp. 269-276

Opinion spam classification on steam review using support vector machine with lexicon-based features

Rafif Taqiuddin*[1], Fitra A. Bachtiar[2], Welly Purnomo[3]
Universitas Brawijaya, Malang, Indonesia[1,2,3]

Article Info

Keywords:
Opinion Spam, Text Mining, Support Vector Machine, Lexicon-Based, Steam Review

Article history:
Received: August 17, 2021
Accepted: September 21, 2021
Published: November 30, 2021

Cite:
Taqiuddin, R., Fitra A. Bachtiar, & Welly Purnomo. (2021). Opinion Spam Classification on Steam Review using Support Vector Machine with Lexicon-Based Features. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 6(4). https://doi.org/10.22219/kinetik.v6i4.1323

*Corresponding author
Rafif Taqiuddin
E-mail address:
rafiftq@gmail.com

Abstract

Steam is a video game digital distribution platform developed by Valve Software. Steam provides a user review feature, where users can write about criticism or comments on games that can contain positive or negative sentiments. Based on the questionnaire that the author conducted to Steam users from all over Indonesia, the user review feature provided by Steam was not sufficient. This is because there are fake reviews that allow biased opinions from certain parties so that a phenomenon called review bombing often occurs where users review only to drop or raise the image of a product, not to review it sincerely. From these problems, a solution design is needed that can classify fake reviews on the Steam service. The Support Vector Machine (SVM) classification method was chosen as the model in combination with lexicon-based feature retrieval and Term Frequency – Inverse Document Frequency (*TF-IDF*) weighting. Of the 236 classification test data conducted by SVM, it produced 105 reviews which were categorized as Valid Reviews. Meanwhile, those categorized as Opinion Spam by SVM are 131 reviews. The accuracy level of the data classification model using Support Vector Machine method is of 81% by dividing training data by 70% and test data by 30% with a random state level of 109. A dashboard in the form of a web application has also been made that contains the classification model to be used for buying reference for Steam user.

1. Introduction

Distribution of a product has many ways and methods. This method is not always done physically in today's digital era. Electronic media can mainly be sent from the publisher to the consumer in an instant via the internet. To gain trust from consumers for the products they offer, a comment or review column is usually provided which serves as a testimony that the product is as advertised. However, lately the review column is often misused by irresponsible parties in conveying things that are not related to the product for various purposes [1].

In the game industries, one of the digital distribution service providers that is often used is called Steam [2] . Steam provides a user review feature where users can write about criticism or comments. The criticism or comments on games may contain positive or negative sentiments [3]. On Steam, it is known that posted review may allow biased opinions so that the review itself become untruthful [4]. This phenomenon is called review bombing. Review bombing often occurs where users review only to drop or raise the image of a product instead of reviewing the product sincerely [4]. A game that has been review bombed has high frequency of extreme sentiment (eg, “Very Good” / “Very Bad”) and can cause customer to be manipulated on the perception of the product [5]. It is however isolated to one case of game and other method of research conducted on overall Steam review, detected many abnormal review so that the interpretation of sentiment is not clear [4]. Moreover, study in spam detection review in Steam is not yet fully explored.

Previous research has conducted an effort to detect false opinions on e-Commerce review. The e-Commerce review data has been classified using the Support Vector Machine classification method [6]. In the next study, four features are extracted and used for the classification of fake reviews. Those four features are Sentiment Feature, Content Feature, Metadata Feature, and Profile Feature. The recognition rate obtained in this study is 74.46% [7]. In other study, the level of performance is also influenced by the source of the lexicon dictionary obtained from SentiWordNet to provide a score for three sentiments, namely positive, negative, and objective. This activity score is used as the value of the Sentiment Feature in the classification as a binary representation of the level of polarity and objectivity [8]. However, using term weighting only would not obtain high predictive accuracy. According to another research of sentiment analysis with the Support Vector Machine method and lexicon based features [9][10], word weighting with Term Frequency-Inverse Document Frequency (TF-IDF) can increase accuracy from 60% to 84.6%. It can be concluded that combining weighting and lexicon features can cover the weaknesses of each feature to produce better accuracy.

Based on the limitation of the previous study, authors decided to use the Support Vector Machine (SVM) method to classify fake reviews assisted by lexicon-based features and TF-IDF word weighting. The combination of lexicon-based feature and TF-IDF word weighting is to provide the classification with better accuracy. SVM is an abstract machine learning algorithm that learns from training data sets and tries to generalize and make precise predictions based on new data [11]. Determination of the features of the classification process is done by lexicon-based, namely the process of selecting words in a document based on an existing lexicon/dictionary [12]. There are two dictionaries used, namely a dictionary with a collection of positive sentiment words and a dictionary with negative sentiment words [13]. Word weighting is a step that sorts documents using a vector space model representation of a data set [14]. Documents in the vector space model are represented in a matrix containing the word weights in the document [14]. The weight expresses the importance of the word to a document and a collection of documents [14]. The importance of this word can be seen from the frequency of occurrence of the word in the document [14]. By using the SVM method, the lexicon that has been determined can be trained with word weight values and then made into a model that can be used for the classification of fake reviews made [15].

By doing this research, it is hoped that the results of the support vector machine classification with the influence of lexicon-based features in the classification of fake game reviews will be known. It is hoped that the results of this study can be used as a reference for Steam users in determining the truth of a review.

2. Research Method

This study uses the following stages: 1) data collection; 2) text preprocessing; 3) term weighting; 4) features extraction with lexicon and directly from dataset; 5) classification model creation with SVM to determine fake reviews and 6) dashboard setup.

2.1 Data Collection

In this study, the data collection technique was to utilize the Steamworks API which had been provided by Steam itself to obtain review data [4]. The review data taken is sourced from the Fallout 76 game because it is assessed that the number of positive and negative sentiments is balanced with the overall category of mixed reviews [16]. The data needed in the classification process later is the review text, the reviewer's playing time when giving a review, the length of the review text, and the number of likes from other users for the reviews made [3]. As well as for the parameters of the data collected is English language review data with the most helpful category. The data finally obtained amounted to the last 1000 reviews per October 2020.

At the dataset design stage, manual review classification is carried out based on predetermined classes, namely True and False [1]. This classification is done by 3 people who are familiar with Steam reviews, with the final labeling results for each review taken from the majority of the classes determined by the examiner [17]. To be able to label each review text based on a predetermined class, each class has an explanation and description which can be seen in Table 1.

Table 1. Review Labels Description

Labels	Description
True	This class contains product-relevant reviews, benefiting readers by providing reasons for the sentiments given.
False	This class contains reviews that have little or no relevance to the product, do not benefit the reader because they do not give reasons for the sentiments given.

2.2. Text Preprocessing

It is necessary to do a text preprocessing process or text pre-processing before processing the review dataset in order to make it easier for the algorithm to classify the text [18][19]. Here are some steps that go into text processing [20]: 1) Tokenization, the first step involves selecting the unit of text to be analyzed and separating the text based on the unit of analysis. This unit can be a word, but in some cases, it may be a group of words or phrases. In its implementation will use the tokenize library from NLTK to separate words in review documents; 2) Cleansing, a process to standardize text by converting all words into lower case and then cleaning the text from elements that are not needed or have no influence in text processing such as the use of symbols and emoticons in the text; 3) Stop Words Removal, stop words can be omitted at this stage because they do not signify any importance in understanding the meaning of the word; 4) Stemming and Lemmatization, the final step in the pre-process is transforming the word into its basic form. Stemming removes suffixes to reduce vocabulary size. Meanwhile, lemmatization is similar to stemming but involves information related to the word class of a term. Both methods combine words that contain the same root word into one token because words that have the same root word also have similar meanings. Each stage will remove unnecessary information from the text to further standardize the data. Thereby, reducing the number of dimensions in the text dataset [19].

2.3 Term Weighting

Term Weighting is a step that ranks documents using a vector space model representation of a data set [18]. Documents in the vector space model are represented in a matrix containing the word weights in the document [18]. The weight expresses the importance of the word to a document and a collection of documents [18]. The importance of this word can be seen from the frequency of occurrence of the word in the document [18]. One of the word weighting methods that is often used is the Term-Frequency Inverse Document Frequency (*TF-IDF*) which is the value of the occurrence of words in the document offset by the repetition of the word throughout the document [21]. The value of word occurences is calculated by Equation 1 and the repetition is calculated in Equation 2. Both equations are then calculated in Equation 3 to determine the value of *TF-IDF*.

$$TF(d,t) = f(d,t) \quad (1)$$

$$IDF(t) = log(\frac{Nd}{df(t)}) \quad (2)$$

$$TF.IDF(d,t) = TF(d,t)\ x\ IDF(t) \quad (3)$$

For example, it can be seen in Table 2 using the available dataset.

Table 2. Term Weighting Example

Term	Doc. 1	Doc. 2	Doc. 3	Doc. 4
issue	0.6	0	0	0
fun	0.6	0	0	0
friend	0.3	0	0	0.3
sale	0.3	0	0	0.3
update	0	0.6	0	0

2.4 Features Extraction

2.4.1 Lexicon Based Features

To carry out the Sentiment Feature (SF) examination, a dictionary / word lexicon that has been defined previously is used to determine the polarity and subjectivity weights of each review text document [7]. This lexicon itself is taken by utilizing the TextBlob library from NLTK, in which the library has provided a function to predict the polarity and subjectivity values of a text [22] [23]. The rating range itself is worth -1 to 1 for polarity where -1 is the most negative sentiment and 1 is the most positive sentiment, while for subjectivity it is 0 to 1 where the higher the value, the higher the level of subjectivity [24]. Example result of this feature extraction can be seen in Table 3.

Table 3. Example of Lexicon Based Features Extraction Result

Review Text	Polarity	Subjectivity
The game has its issues, but overall is fun. especially, with friends. However, I would wait for a sale.	-0.02	0.4

2.4.2 Non-Lexicon Features

The feature extraction process carried out is for three other features that are used to identify fake reviews, namely Content Feature (CF), Metadata Feature (MF), and Profile Feature (PF) [7]. For CF the data taken is the number of likes from other users on the reviews [17]. Then, for MF, it is obtained from the length of the text in each review [17]. Meanwhile PF feature is taken from the value of the length of time played by users who make reviews [17]. The three values of the features obtained will then be normalized using the min-max method to facilitate the next classification process. Example result of this feature extraction can be seen in Table 4.

Table 4. Example of Non-Lexicon Features Extraction Result Normalized

Review Text	CF	MF	PF
The game has its issues, but overall is fun. especially, with friends. However, I would wait for a sale.	0.09	0.01	0.06

2.5 Classification Model using SVM

Support Vector Machine or abbreviated as SVM is an abstract machine learning that learns from training data sets or training data and tries to generalize and make correct predictions on new data [11]. SVM is used to find the best hyperplane by maximizing the distance between classes. Hyperplane itself is a function that can be used to separate between classes. For functions that are in two dimensions of classification, the separator that is sought is whereas line, which is a dividing line between the two classification classes. Illustration of SVM can be seen in Figure 1.

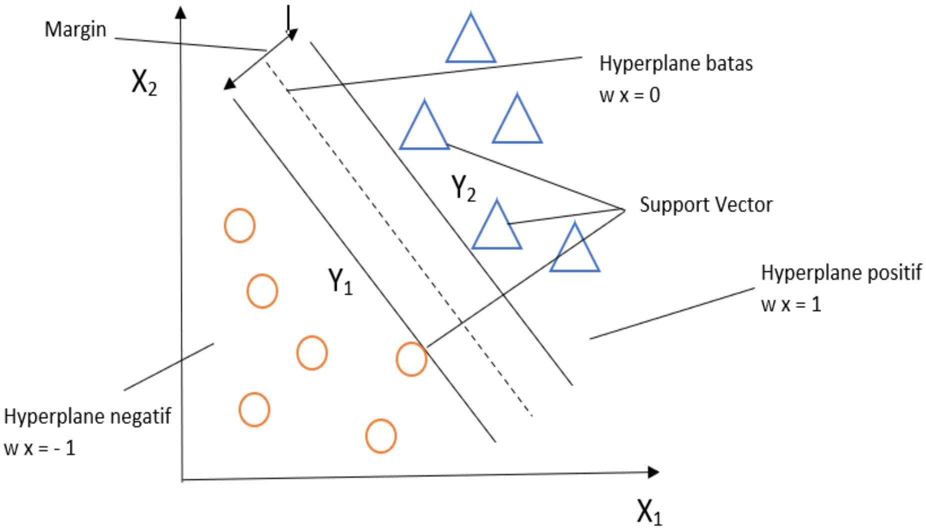


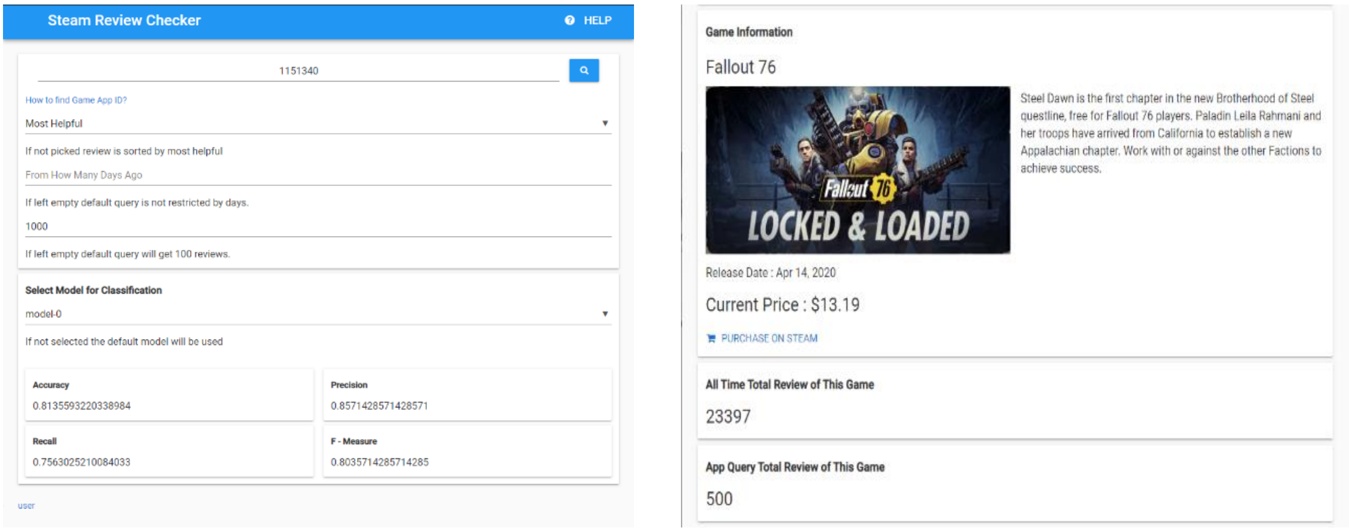
Figure 1. Hyperplane Illustration of SVM

To perform the classification, the entire dataset will be divided into two parts [15]. The first part is training data which is review data that has been manually labeled and the second part is test data which will be classified by class label by the Support Vector Machine classifier. Specifically, the SVM Kernel used in this study is Linear with assistance of LinearSVC function in Python scikit.learn library and the parameters used for the training purpose are C = 1.0, maximum iteration of 1000, multi-class strategy is one versus rest and random state of 109 which are all default parameters for LinearSVC. The distribution of data is 70% for training data and 30% for test data. It is shown in empirical analysis that the best result for testing is 20%-30% of the actual data, because the value is close to the v-fold cross validation which often gives the best results [25].

2.6 Dashboard Setup

After the classification model has been obtained, a dashboard will be made to deploy the model and display the classification result information that can later be used by Steam users as supporting information in their purchases. The dashboard is made using Anvil, a python web framework that can easily configured to work with data science notebook.

For user dashboard pages as seen in Figure 2 and Figure 3, start with the search field at the top of the page. In Figure 2 (a), users can enter the ID of the game they want to search. This search can be based on several selectable query sorting parameters. Users can choose the model to be used for classification. These models are models that have been trained furthermore by the admin after a lot more of classification data is generated by user and it is based on the initial model that was the result of this research. Users can also view the accuracy, precision, recall, and f-measure metrics for each model. After the user presses the search button, the first part that will be displayed is the general information of the game being searched for. In Figure 2 (b), this section contains the title of the game, release date, brief description, current price, and a link to purchase it directly on Steam. In addition, there is also information about the overall number of reviews on the game and the number of reviews that users query. Next is the section that contains information about how many reviews are categorized as valid and spam which is displayed with a bar chart so that users can easily see a visualization of the comparison as shown in Figure 3 (a). The classification map section contains information about the category of review recommendations in the classification class as shown in Figure 3 (b). This is useful for users who want to see how many positive and negative recommendations are in each class. The next section is a review table sorted by the number of upvotes and playtime.



Steam Review Checker

1151340

How to find Game App ID?

Most Helpful

If not picked review is sorted by most helpful

From How Many Days Ago

If left empty default query is not restricted by days.

1000

If left empty default query will get 100 reviews.

Select Model for Classification

model-0

If not selected the default model will be used

Accuracy 0.8135593220338984	Precision 0.8571428571428571
Recall 0.7563025210084033	F - Measure 0.8035714285714285

Game Information

Fallout 76

Steel Dawn is the first chapter in the new Brotherhood of Steel questline, free for Fallout 76 players. Paladin Leila Rahmani and her troops have arrived from California to establish a new Appalachian chapter. Work with or against the other Factions to achieve success.

Release Date : Apr 14, 2020

Current Price : $13.19

All Time Total Review of This Game

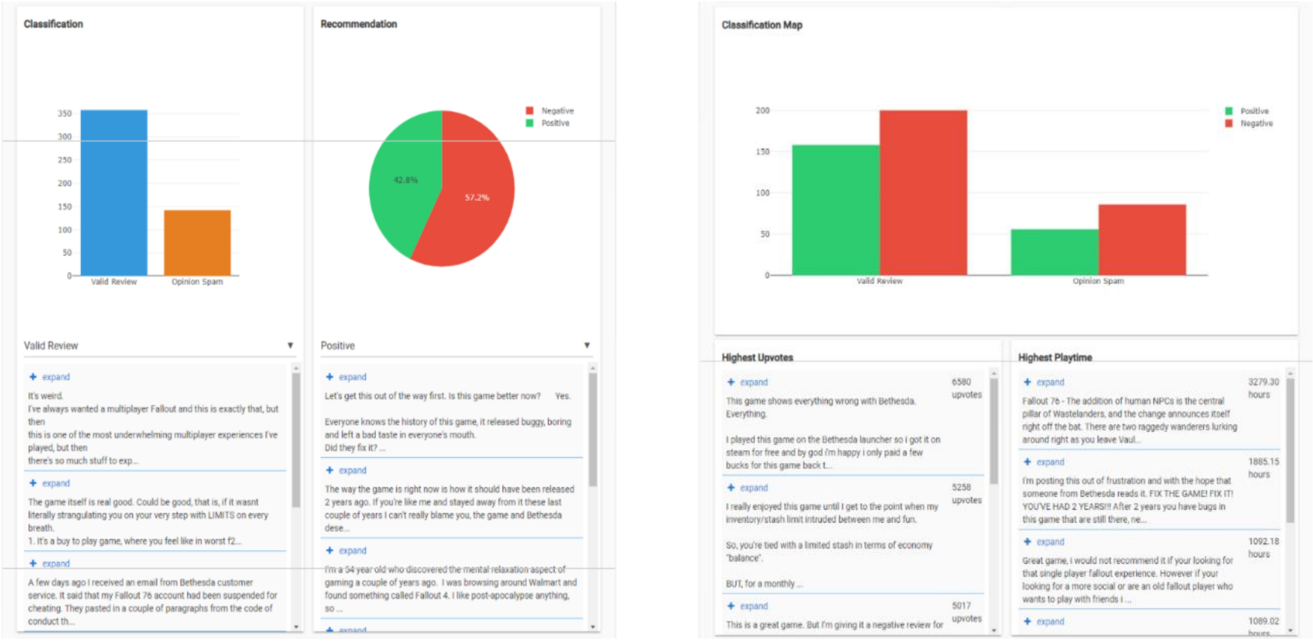
23397

App Query Total Review of This Game

500

(a) (b)

Figure 2. Search Bar and Search Result of User Dashboard



Valid Review

+ expand

It's weird.
I've always wanted a multiplayer Fallout and this is exactly that, but then
this is one of the most underwhelming multiplayer experiences I've played, but then
there's so much stuff to exp...

+ expand

The game itself is real good. Could be good, that is, if it wasnt literally strangulating you on your very step with LIMITS on every breath.
1. It's a buy to play game, where you feel like in worst f2...

+ expand

A few days ago I received an email from Bethesda customer service. It said that my Fallout 76 account had been suspended for cheating. They pasted in a couple of paragraphs from the code of conduct th...

Positive

+ expand

Let's get this out of the way first. Is this game better now? Yes.

Everyone knows the history of this game, it released buggy, boring and left a bad taste in everyone's mouth.
Did they fix it? ...

+ expand

The way the game is right now is how it should have been released 2 years ago. If you're like me and stayed away from it these last couple of years I can't really blame you, the game and Bethesda dese...

+ expand

I'm a 54 year old who discovered the mental relaxation aspect of gaming a couple of years ago. I was browsing around Walmart and found something called Fallout 4. I like post-apocalypse anything, so ...

+ expand

Highest Upvotes

+ expand 6580 upvotes

This game shows everything wrong with Bethesda. Everything.

I played this game on the Bethesda launcher so i got it on steam for free and by god i'm happy i only paid a few bucks for this game back t...

+ expand 5258 upvotes

I really enjoyed this game until I get to the point when my inventory/stash limit intruded between me and fun.

So, you're tied with a limited stash in terms of economy "balance".

BUT, for a monthly ...

+ expand 5017 upvotes

This is a great game. But I'm giving it a negative review for

Highest Playtime

+ expand 3279.30 hours

Fallout 76 - The addition of human NPCs is the central pillar of Wastelanders, and the change announces itself right off the bat. There are two raggedy wanderers lurking around right as you leave Vaul...

+ expand 1885.15 hours

I'm posting this out of frustration and with the hope that someone from Bethesda reads it. FIX THE GAME! FIX IT! YOU'VE HAD 2 YEARS!!! After 2 years you have bugs in this game that are still there, ne...

+ expand 1092.18 hours

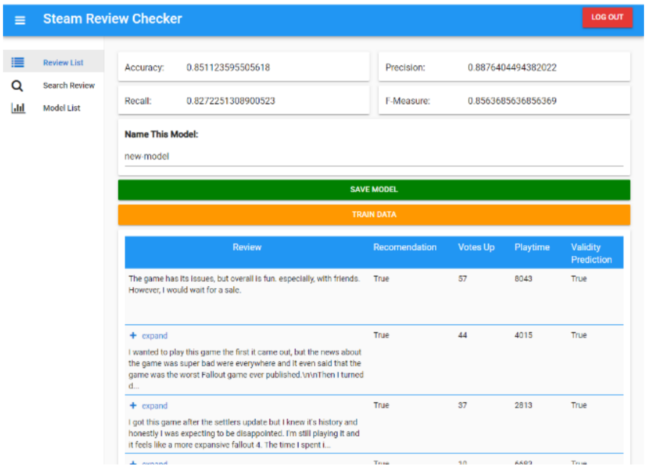
Great game, i would not recommend it if your looking for that single player fallout experience. However if your looking for a more social or are an old fallout player who wants to play with friends i ...

+ expand 1089.02 hours

(a) (b)

Figure 3. Classification Result of User Dashboard

In the admin dashboard, there are several more sub-sections that are separated based on their respective functionality. The first is the Review List page shown in Figure 4, which is the section that displays the results of review queries from users that have been entered into the database. The reviews in this database can be retrained using a classifier into a new model, so that it is expected to further increase the level of prediction accuracy. The next section on the admin dashboard is the Search Review as seen in Figure 5 (a), whose features are similar to the user dashboard. The query results from the Search Review page can be directly trained by the classifier to get a new model as well as on the Review List page. However, the dataset being trained is the review data on this page, not mixed with user query data in the database. In Figure 5 (b) is the last section on the admin dashboard is the Model List page. This page displays all the models that have been saved by the admin. Admin can compare which model performs best and decide to use that model as the default classifier on the user dashboard.

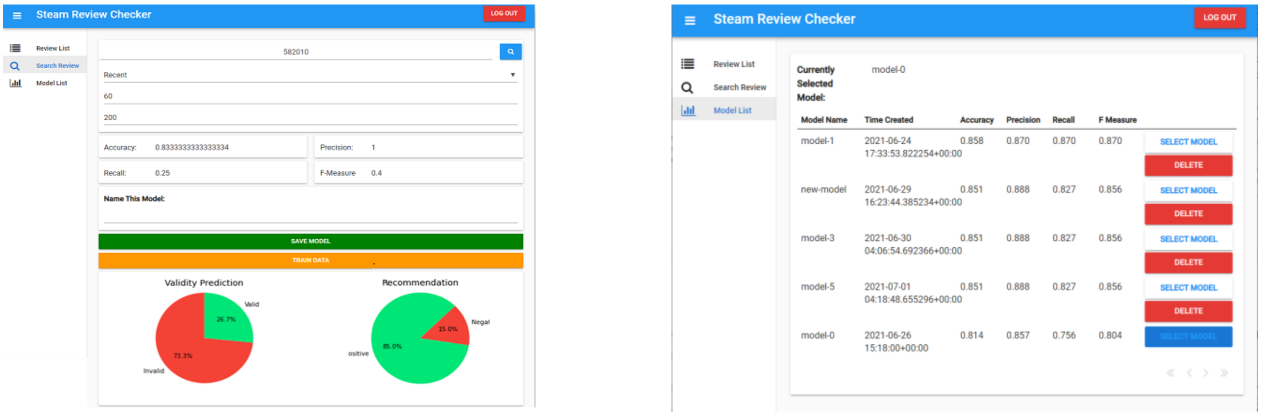


Accuracy: 0.851123595505618
Precision: 0.8876404494382022
Recall: 0.8272251308900523
F-Measure: 0.8563685636856369

Name This Model: new-model

Review	Recomendation	Votes Up	Playtime	Validity Prediction
The game has its issues, but overall is fun. especially, with friends. However, I would wait for a sale.	True	57	8043	True
+ expand I wanted to play this game the first it came out, but the news about the game was super bad were everywhere and it even said that the game was the worst Fallout game ever published.\n\nThen I turned d...	True	44	4015	True
+ expand I got this game after the settlers update but I knew it's history and honestly I was expecting to be disappointed. I'm still playing it and it feels like a more expansive fallout 4. The time I spent i...	True	37	2813	True

Figure 4. List of Review on Admin Dashboard



582010
Recent
60
200

Accuracy: 0.8333333333333334
Precision: 1
Recall: 0.25
F-Measure 0.4

Name This Model:

(a)

Currently Selected Model: model-0

Model Name	Time Created	Accuracy	Precision	Recall	F Measure
model-1	2021-06-24 17:33:53.822254+00:00	0.858	0.870	0.870	0.870
new-model	2021-06-29 16:23:44.385234+00:00	0.851	0.888	0.827	0.856
model-3	2021-06-30 04:06:54.692366+00:00	0.851	0.888	0.827	0.856
model-5	2021-07-01 04:18:48.655296+00:00	0.851	0.888	0.827	0.856
model-0	2021-06-26 15:18:00+00:00	0.814	0.857	0.756	0.804

(b)

Figure 5. Review Search Bar and Classification Model List on Admin Dashboard

3. Results and Discussion

To ensure that the classification carried out by the method can predict correctly, it is necessary to evaluate it with an assessment metric. There are several quality measurement metrics that are usually used for binary classification problems, including accuracy, precision, recall, and f-measure [26]. Accuracy itself is the ratio of the true class compared to the total amount of data which can be calculated using Equation 4. Precision is the ratio of positive correct predictions to the overall positive prediction results shown in Equation 5. Recall is the ratio of positive true predictions compared to the overall positive value of the actual data as seen in Equation 6. F1-score, which is a weighted comparison of the average precision and recall which shown in Equation 7.

To calculate the value of the four classification accuracy parameters, a confusion matrix must first be created containing the values of TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative). These four values will later be calculated to get the Accuracy, Precision, Recall, and F1-score values. After calling the confusion_matrix function from the scikit learn library, the matrix is obtained as shown in Table 5.

$$accuracy = \frac{TN + TN}{Total\ Data} \tag{4}$$

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$f1 - score = \frac{2(recall * Precision)}{recall + precision} \tag{7}$$

Table 5. Confusion Matrix Result

		Prediction	
		Valid	Fake Review
Actual	Valid	90	29
	Fake Review	15	102

As can be seen in Table 5, the value of True Positive and True Negative is fairly high compared to the False Positive and False Negative. This shows that the prediction accuracy in general looks good, but to prove it, we will see by calculating the parameter values of accuracy, precision, recall, and f1-score. To find these values, scikitlearn library from Python was used which can calculate these values by using the classification_report function. The results of this function call can be seen in Table 6.

Table 6. Accuracy, precision, recall, and f1-score result

Classification	Accuracy	Precision	Recall	F1-Score
Valid	-	0.85	0.75	0.80
Fake Review	-	0.78	0.87	0.82
Average	0.81	0.82	0.81	0.81

The higher precision value is found in the Valid Review class with a value of 0.86, this means that the Valid Review value with the correct prediction has a higher ratio than the Opinion Spam value, which has the correct prediction. Unlike the case with recall, where the Opinion Spam class shows a higher value at 0.87. This shows that the Opinion Spam class has a higher correct prediction when viewed from the overall actual data which is actually False compared to the Valid Review class. The f1-score values for both classes are also good at 0.80 and 0.82 plus good accuracy at 0.81 or 81% of classifying overall reviews indicate that the classifier used has worked well in predicting class classification from the review data. This result is expected within 70% - 80% SVM accuracy from previous research which compare different methods on opinion spam classification [27]. In fact, when comparing accuracy of text-based feature on the previous research, the proposed method of this research is better in term of accuracy of 81% as opposed to the former at 71%.

4. Conclusion

Implementation of the Support Vector Machine method in this study which was carried out in combination with lexicon-based feature retrieval and weighting of Term Frequency – Inverse Document Frequency (*TF-IDF*) can be a solution in classifying fake reviews. From 236 test data classification performed, 90 were classified as true valid review and 102 classified as true opinion spam. The data classification model using the Support Vector Machine produces an accuracy of 81%. This classification model is then deployed into a dashboard in web based application for widespread use by Steam users to help solve the initial problem of identifying fake review on the platform.

Notation

$f(d,t)$: number of occurrences of term t in document d.
$TF(d,t)$: term frequency of term t in document d.
$df(t)$: number of documents containing term t.
Nd : number of documents
$IDF(t)$: inverse document frequency of term t in all of documents.
$TF.IDF(d,t)$: term frequency – inverse document frequency of term t in document d.

References

[1] Jindal, N., & Liu, B. (2008). *Opinion spam and analysis*. ACM Press. https://doi.org/10.1145/1341531.1341560

[2] Lin, D., Bezemer, C. P., & Hassan, A. E. (2017). Studying the urgent updates of popular games on the steam platform. *Empirical Software Engineering*, *22*(4), 2095-2126. https://doi.org/10.1007/s10664-016-9480-2

[3] Bulygin, D. (2020). Game Experience Evaluation. A Study of Game Reviews on the Steam Platform. In *Digital Transformation and Global Society: 5th International Conference, DTGS 2020, St. Petersburg, Russia, June 17-19, 2020, Revised Selected Papers* (Vol. 1242, p. 117). Springer Nature. https://doi.org/10.1007/978-3-030-65218-0_9

[4] Bian, P., Liu, L., & Sweetser Kyburz, P. (2021). Detecting Spam Game Reviews on Steam with a Semi-Supervised Approach. In International Conference on the Foundations of Digital Game. ACM.

[5] Tomaselli, V., Cantone, G. G., & Mazzeo, V. (2021). The Polarising Effect of Review Bomb. arXiv preprint .

[6] Pasaribu, B. E., Herdiani, A., & Astuti, W. (2019). Deteksi fake reviews menggunakan support vector machine. E-Proceeding of Engineering, 6(2), 8788.

[7] Li, F., Huang, M., & Zhu, X. (Eds.). (2011). *Learning to identify review spam* (Issue IJCAI International Joint Conference on Artificial Intelligence). https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-414

[8] Cho, H., Kim, S., Lee, J., & Lee, J. S. (2014). Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. *Knowledge-Based Systems*, *71*, 61-71. https://doi.org/10.1016/j.knosys.2014.06.001

[9] Ferlin, J., Bachtiar, F. A., & Rusydi, A. N. (2020). Klasifikasi Customer Intent Untuk Mengetahui Tingkat Kepuasan Pelanggan Menggunakan Metode Support Vector Machine Pada Restoran Bakso President. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 3(7), 9867–9875.

[10] Ruslim, K. I., Adikara, P., & Indriati. (2019). Analisis sentimen pada ulasan aplikasi mobile banking menggunakan metode support vector machine dan lexicon based features. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 3(7), 6694–6702.

[11] Campbell, C., & Ying, Y. (2011). *Learning with Support Vector Machines*. Morgan & Claypool. https://doi.org/10.2200/S00324ED1V01Y201102AIM010

[12] Augustyniak, L., Kajdanowicz, T., Szymański, P., Tuligłowicz, W., Kazienko, P., Alhajj, R., & Szymanski, B. (2014, August). Simpler is better? Lexicon-based ensemble sentiment classification beats supervised methods. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)* (pp. 924-929). IEEE. https://doi.org/10.1109/ASONAM.2014.6921696

[13] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/coli_a_00049

[14] Bafna, P., Pramod, D., & Vaidya, A. (2016, March). Document clustering: TF-IDF approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 61-66). IEEE. https://doi.org/10.1109/ICEEOT.2016.7754750

[15] Singh, M., & Pamula, R. (2018, September). Email spam classification by support vector machine. In *2018 International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 878-882). IEEE. https://doi.org/10.1109/GUCON.2018.8674973

[16] Colhon, M., Vlăduţescu, T., & Negrea, X. (2017). How Objective a Neutral Word Is? A Neutrosophic Approach for the Objectivity Degrees of Neutral Words. *Symmetry*, *9*(11), 280. https://doi.org/10.3390/sym9110280

[17] Fusilier, D. H., Montes-y-Gómez, M., Rosso, P., & Cabrera, R. G. (2015). Detection of Opinion Spam with Character n-grams. *Computational Linguistics and Intelligent Text Processing*, 285–294. https://doi.org/10.1007/978-3-319-18117-2_21

[18] Guillet, F., & Hamilton, H. J. (2010). *Quality Measures in Data Mining (Studies in Computational Intelligence, 43)* (Softcover reprint of hardcover 1st ed. 2007 ed.). Springer. https://doi.org/10.1007/978-3-540-44918-8

[19] Kao, A., & Poteet, S. R. (2010). *Natural Language Processing and Text Mining* (Softcover reprint of hardcover 1st ed. 2007 ed.). Springer. https://doi.org/10.1007/978-1-84628-754-1

[20] Anandarajan, M., Hill, C., & Nolan, T. (2018). *Practical Text Analytics: Maximizing the Value of Text Data (Advances in Analytics and Data Science, 2)* (Softcover reprint of the original 1st ed. 2019 ed.). Springer. https://doi.org/10.1007/978-3-319-95663-3_1

[21] Yun-tao, Z., Ling, G. & Yong-cheng, W. An improved TF-IDF approach for text classification. *J. Zheijang Univ.-Sci. A* **6,** 49–55 (2005). https://doi.org/10.1007/BF02842477

[22] Ahuja, S., & Dubey, G. (2017, August). Clustering and sentiment analysis on Twitter data. In *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)* (pp. 1-5). IEEE. https://doi.org/10.1109/TEL-NET.2017.8343568

[23] Laksono, R. A., Sungkono, K. R., Sarno, R., & Wahyuni, C. S. (2019, July). Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes. In *2019 12th International Conference on Information & Communication Technology and System (ICTS)* (pp. 49-54). IEEE. https://doi.org/10.1109/ICTS.2019.8850982

[24] Gharehchopogh, F. S., & Khalifelu, Z. A. (2011, October). Analysis and evaluation of unstructured data: text mining versus natural language processing. In *2011 5th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICAICT.2011.6111017

[25] Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), 111–147.

[26] Gharehchopogh, F. S., & Khalifelu, Z. A. (2011, October). Analysis and evaluation of unstructured data: text mining versus natural language processing. In *2011 5th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICAICT.2011.6111017

[27] Ahmed H, Traore I, Saad S (2018). Detecting opinion spams and fake news using text classification, Security and Privacy, 2018;1:e9. https://doi.org/10.1001/spy2.9