Journal of Engineering Science and Technology
Special Issue on ICIT2022 Vol. 18, No. 6 (2023) 1 - 10

NOVEL RECOMMENDATION SYSTEM BASED ON USER PREFERENCES USING SENTIMENT ANALYSIS

NYIMAS SOPIAH[1,2], TRI BASUKI KURNIAWAN[1,*], DESHINTA ARROVA DEWI[2], MOHD ZAKI ZAKARIA[3], MUHAMMAD TASHDIQ BASRI[3]

[1]Faculty of Computer Science, University of Bina Darma, Palembang, Indonesia
[2]Faculty of Data Science and Information Technology, INTI International University, Malaysia
[3]Faculty of Computer & Mathematics Sciences, University Technology Mara, Malaysia
*Corresponding Author: tribasukikurniawan@binadarma.ac.id

Abstract

When searching online for something to buy, many consumers will look at an online review or rating for a product. 'Unknown' brands of action cameras, watches, and headphones are receiving thousands of reviews, as they are unverified. There is no evidence that the reviewer bought or used the product. The problem is that some people deliberately want to take advantage of others by making a fake review on the product just to make the product appear to be in the market. This problem also happens to users who want to buy a novel from an online website. Reading each review could take a lot of time, and determining whether the reviews are positive or negative would be difficult because some reviews might be fake. In this research, a prototype of the novel recommendation system based on a review of websites is developed. This research's methodology includes an initial investigation, knowledge acquisition, data collecting, analysis, and categorization. The study focuses on how evaluations are differentiated between favourable and unfavourable texts. Based on user choices, the categorization outcome suggests a good novel. The Decision Tree, Back-propagation Neural Network, Naive Bayes, Support Vector Machine, and Recurrent Neural Network classifiers were all used in this research. For this research, the Recurrent Neural Network obtained the best result for classifying the sentiment, with an average accuracy of 87.85% and an average error rate of 9.1%. The classifier was trained with a learning rate of 0.001 and 100 epochs. The result classification can be increased by training more datasets or tuning the hyperparameter value. This statement shows that the future of this research does not end here but can improve from time to time.

Keywords: Consumer behaviour, Consumption, Naïve Bayes, Novel recommendation system, Recurrent neural network, Support vector machine, User references.

1. Introduction

Reading is an occasion at which verses or pulls from books are read to a group of people. Reading books allude to a reader for individuals figuring out how to read, to enable them to get acquainted with taking a gander at and understanding composed words. Reading is important in our life, especially because the computer has started replacing the book in children's hands [1]. High-quality reading is a meta-cognitive thinking element [2].

The analysis research of Malaysians reading 2014 by the National Library of Malaysia (PNM) found that Malaysians read 15 books yearly [3]. The study likewise demonstrated that Malaysians lean toward reading materials such as magazines 62.8 percent; newspapers (61.2 percent); scientific books (54.8 percent); novels and fiction (47 percent); online materials (46.4 percent); true to-life books (29.9 percent); comics (25.3 percent) and diaries (19.9 percent) [3].

The u-libraries administration being imaginative in the library administration opens a roomy space for individuals to read books anywhere without the need to go to the library. Expanded utilization of the u-library entry incorporates new individuals who register, the utilization of online assets, and the quantity of obtained digital book contents are relied upon to build every year. According to [3], 11 million clients utilized PNM benefits last year, while 2 million advances were recorded in physical and virtual books. Therefore, these statistics can generate a million reviews of books. This review lets us know how good the book's quality is, either good or bad.

Big data in social media and websites is one of the platforms for the reader to get information and suggestion about the type of novel, date of novel release, and price of the novel [4, 5]. Most book website reviews and online purchase use technology and the internet to present their product and list information about the novel with interactive text, pictures, and many reviews to attract customer websites [6]. However, a flood of information that includes unnecessary information can cause fatigue in finding a suitable novel for readers.

According to Rebora and Pianzola [7], the comment that comes from the reviewers could give a positive or negative perspective on the book's content. Besides, people annotate books for many reasons [8], and sharing them can be seen as irrelevant in some social reading applications [9]. It has always been a difficult task [10] because some subtasks which as feature extraction, polarity classification, and assessment measures, need to be done to gain unbiased opinions, mainly under the assumption of unrealistic grammar errors.

2.Methodology

This methodology framework consists of 8 phases: preliminary study and knowledge acquisition to complete the research objective, which are system design, development, and system testing and evaluation, as shown in Fig. 1.

The preliminary study is one of the phases to complete the first objective, which is to identify the suitable novel to choose based on the reader's preference, as shown in Fig. 1. In this phase, research and understanding the importance of readers and novels that readers commonly read are done by reading and understanding the literature review that consists of journal, article, and paper.

Fig. 1. In phases on methodology framework.

Knowledge acquisition is a process of collecting knowledge and information about the project. Activities involving these phases are gathering knowledge from the literature review and online searching from well-known book review websites such as BookLikes, Goodreads, and New York Time Book Blog. A literature review is a gain from an online database such as Scopus, IEEE, ScienceDirect, Emerald Insight, and ResearchGate.

Data collection is the phase where a suitable approach is used to collect and retrieve data from a well-known and famous online selling website, Amazon.com. The data can easily be extracted manually; however, with over 20 thousand reviews and comments for each book, it is an intensive activity. Hence, the data collection phase requires scraping websites to gather the data, such as the reviews comment for each book based on the type of book categories. The data is collected using the Web Scraper extension on a web browser, as shown in Fig. 2.

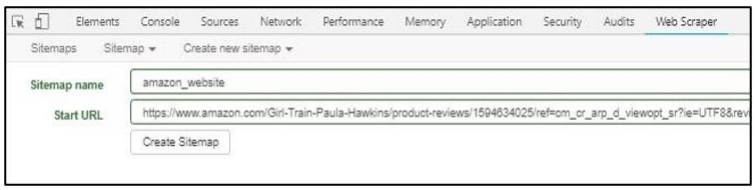
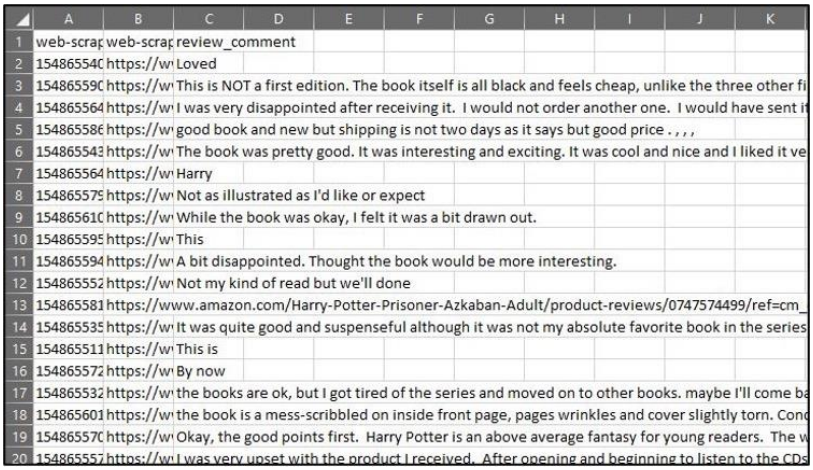


Fig. 2. Create a sitemap for scraping.

Figure 2 shows that a new sitemap must be created to extract data from websites; this sitemap will simultaneously contain all the scraping operations. Before it runs the process, it requires a website URL to know the location of the required data, as shown in Figs. 3 and 4.

Fig. 3. Select the reviews comment.



	A	B	C	D	E	F	G	H	I	J	K
1	web-scrap	web-scrap	review_comment								
2	154865540	https://w	Loved								
3	154865590	https://w	This is NOT a first edition. The book itself is all black and feels cheap, unlike the three other fi								
4	154865564	https://w	I was very disappointed after receiving it. I would not order another one. I would have sent it								
5	154865586	https://w	good book and new but shipping is not two days as it says but good price . , , ,								
6	154865543	https://w	The book was pretty good. It was interesting and exciting. It was cool and nice and I liked it ve								
7	154865564	https://w	Harry								
8	154865575	https://w	Not as illustrated as I'd like or expect								
9	154865610	https://w	While the book was okay, I felt it was a bit drawn out.								
10	154865595	https://w	This								
11	154865594	https://w	A bit disappointed. Thought the book would be more interesting.								
12	154865552	https://w	Not my kind of read but we'll done								
13	154865581	https://www.amazon.com/Harry-Potter-Prisoner-Azkaban-Adult/product-reviews/0747574499/ref=cm_									
14	154865535	https://w	It was quite good and suspenseful although it was not my absolute favorite book in the series								
15	154865511	https://w	This is								
16	154865572	https://w	By now								
17	154865532	https://w	the books are ok, but I got tired of the series and moved on to other books. maybe I'll come ba								
18	154865601	https://w	the book is a mess-scribbled on inside front page, pages wrinkles and cover slightly torn. Cond								
19	154865570	https://w	Okay, the good points first. Harry Potter is an above average fantasy for young readers. The w								
20	154865557	https://w	I was very upset with the product I received. After opening and beginning to listen to the CDs								

Fig. 4. Result of scraping on excel file.

The web scraper must choose the comment part depicted in Fig. 3 to select the review comment. This comment is the only part where the web scraper scrapes the data. After the scraping process was completed, all data were transferred into an excel file. Figure 4 shows the result of scraping on an excel file.

3. Data Analysis

Data analysis is the phase where the data gathered from the previous phases is processed and analysed to produce useful information. The process includes cleaning the dataset, which removes the non-letter word. The following procedure is transforming all characters in a record to lowercase. After that, tokenization was performed.

After removing the stop words, the remaining word was sent into the WordNet library. This process will remove the redundant word that may happen before moving to the next stage. This process is shown in Fig. 5. The process data were saved into an excel file before proceeding to the analysis.

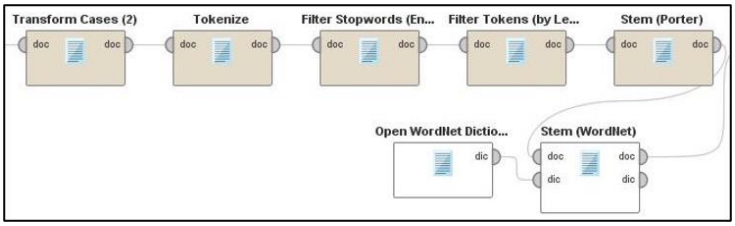


Fig. 5. Transforming, cleaning, and tokenization process of the data.

Pre-processing is the phase where it defines the target from the dataset. In the RapidMiner, the attribute used as the target is 'sentiment,' and the target role is 'label,' which is used for learning. This phase also defines either positive or negative class. The project employed nominal to binominal operative to accomplish this. To change the value of nominal features to a binominal type, use the Nominal to Binominal operation.

After the nominal to binominal process is completed, the result transfers to the Remap Binominals operator. The Remap Binominals operative changes the inward mapping of binominal highlights associated with the predefined positive and negative values. The positive and negative values are identified by the positive value and negative value parameters, respectively.

3.1. Filter, sample, and split data

The filter process is essential before transferring the dataset into the training model. At this point, it determines which review data are kept and which are removed.

In RapidMiner, the Filter Examples are used to do the filtering process. These operative returns those reviews data that match the given condition. The condition set was 'no missing labels' where only those review data are coordinated and do not have a missing incentive in the characteristic with the label role.

In the parameters set, the operator used an absolute sample, which means the example is made of a precisely indicated number of the data. The required number of review data is indicated in the example size parameter. The partitions are the most important parameter of this operative. It shows the number of allotments and the overall proportions of each partition. The ratios used for training are 0.6 and 0.4 for testing. The sum of the ratio is 1. The filter, sample, and slit data process can be seen in Fig. 6.

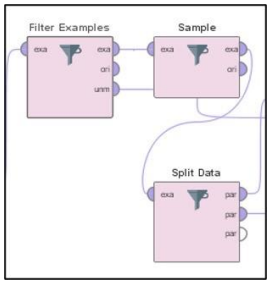


Fig. 6. Filter, Sample and Split data process.

3.2. Classification

Classification is the phase where the review data is classified based on positive, neutral, and negative words. Text that reviews data is tokenized using Text Vectorization's operation before it can be classified.

The operative delivers a pre-processing model, which can relate to the new data sets to perform the same processing on this data. It is necessary to transform scoring data sets in the same way as training data sets. This step can be seen in Fig. 7.

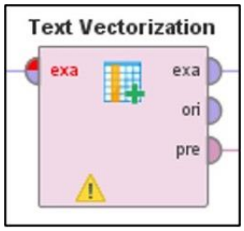


Fig. 7. Text vectorization process.

Figure 7 shows this operative performs a fully automated feature engineering process covering feature selection and generation. The function of this operator is to ensure consistent behavior of the operator between classification and regression tasks and avoid maximizing "negative" error rates, which is often confusing, and sometimes the result is ridiculous.

The following process in this stage is training the model with the classifier. The classifier used in this project were Deep Learning, Support Vector Machine (SVM), Naïve Bayes, and Decision Tree, which are shown in Fig. 8. Each of the classifiers can make a prediction. It is chiefly used to gauge how precisely a model performs during learning.

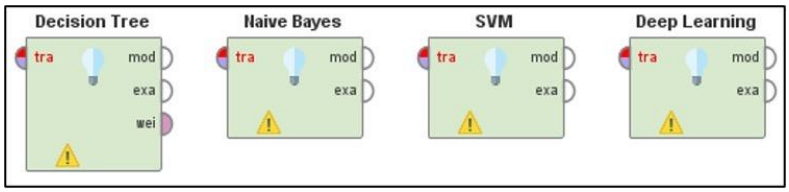


Fig. 8. Type of classifier used in the model for training and testing.

Figure 8 shows the classifier used for training and testing was first connected to the model simulator operator. This operative gives a simple, ongoing technique to change the contributions to a model and view the output. It allows users to see the predictions, confidence, and explanation for every output produced. The significant advantage of using this operator is that clients can utilize the built-in optimization technique to discover ideal information settings to accomplish a perfect result.

Calculating classifier performance was the last stage before delivering the output to the user. Before the testing data is sent to the Performance operator, the data are sent to the Generate Batch operator to create an index for each data and generate another batch column that partitions the data into a predetermined number of batches. These batches are doled out by utilizing the mod function on the row number, as shown in Fig. 9.

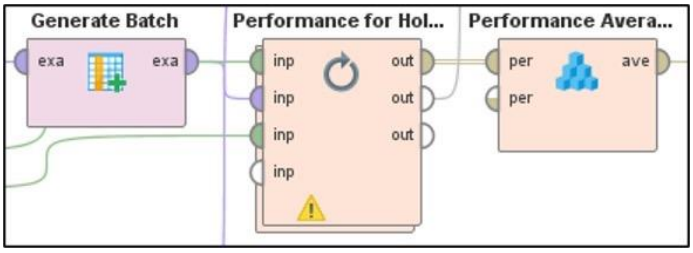


Fig. 9. Generate index for every data and calculate classifier performance.

Figure 9 shows inside the Performance operator; the testing data was processed here to calculate the confidence, accuracy, and performance. Based on the training data that the model simulator had learned; the Apply Model operator used this data to predict testing data. Then the output data from this operator were sent to the

Performance Binominal Classification operator. This operator is used to identify the value of true positives, false positives, false negatives, and true negatives. These values are then used to calculate the accuracy, error rate, class precision, and recall for positive and negative, as shown in Fig. 10.

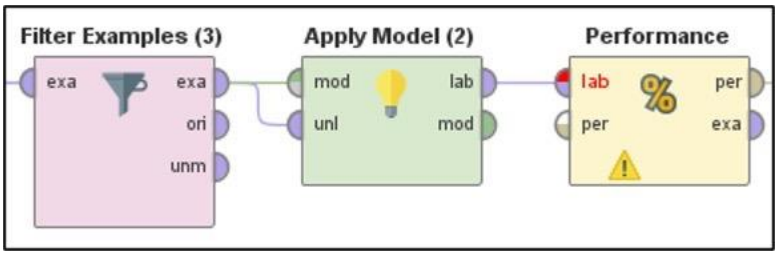


Fig. 10. Apply the learning algorithm to the testing model.

4. Result and Discussion

The accuracy average is calculated for each classifier to find the best accurate classifier for the model. The formulae to calculate accuracy and error rate can obtain from the confusion matrix value, which is True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Equation (1) is used to calculate the accuracy, while Equation (2) is used to calculate the error rate.

$$Accuracy = \frac{TP + TN}{TP + TN + FN}$$

$$Error\ Rate = \frac{FP + FN}{TP + FN + FN}$$

All results from different classifiers used for the training and testing were recorded and tabulated to compare each classifier with other reviews dataset. This result can be seen in Table 1 and Fig. 11.

Table 1. Test model specifications and test conditions.

No.	Name of Novel	Decision Tree	Deep Learning	Naïve Bayes	SVM	RNN
1	Dark Tower	85.71	85.49	65.11	88.06	90.70
2	Fellowship	93.66	92.73	67.57	93.73	94.44
3	Game of Throne	89.57	88.14	53.77	89.50	90.60
4	Harry Potter and Prisoner of Azkaban	95.24	94.60	56.99	95.27	95.40
5	Hobbit	92.46	89.59	63.77	92.30	92.70
6	Edge of Eternity	73.04	56.74	41.34	73.09	70.10
7	Fifty Shade	71.14	66.40	39.35	71.13	71
8	Me Before You	91.77	85.41	50.48	91.77	91.77
9	The Girl on The Train	73.88	66.51	38.16	73.91	73.30
10	The Nightingale	95.35	91.72	42.97	95.41	95.41
11	Catching Fire	86.73	82.64	45.84	86.70	86.40
12	Hunger Game	89.76	90.33	53.08	89.74	91.80
13	Mockingjay	89.76	90.33	53.08	89.74	92.80
14	The Martian	92.48	92.48	56.74	92.44	93.70
15	The Running Man	85.62	85.62	58.98	87.19	87.60
Accuracy average		87.08	83.92	52.49	87.34	87.85

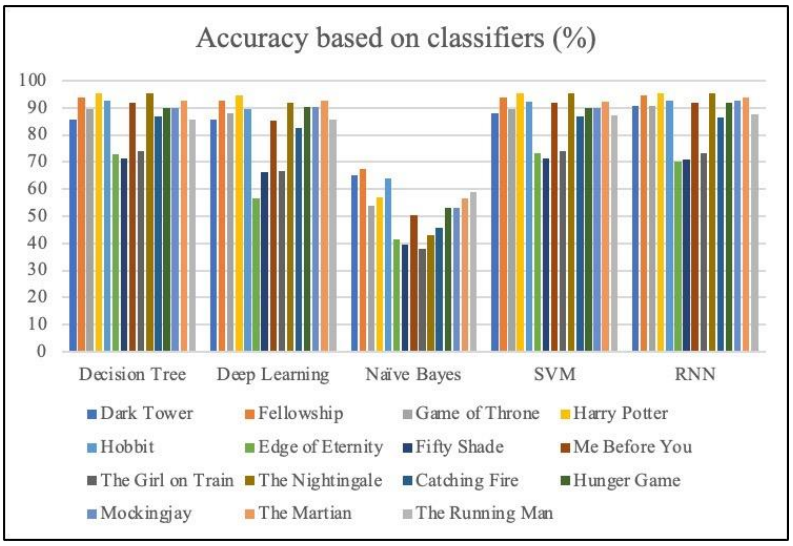


Fig. 11. Accuracy for each classifier.

Table 1 and Fig. 11 show the best average accuracy obtained by the RNN classifier, 87.85, and the worst result is Naïve Bayes at 52.92. Other classifiers, SVM, Decision Tree, and Deep Learning, get the result almost like RNN at 87.34, 87.08, and 83.92, respectively.

Each classifier's error rate has also been recorded to analyse which classifier produces less error rate. The best classifier for learning not just has a higher accuracy but also has a lower error rate. The error rate means how often the model predicts wrong. Suppose the model has higher accuracy, but a higher error rate means it might have overfitting or underfitting issues where it cannot make an accurate prediction. The result of the error rate for all review datasets can be seen in Table 2 and Fig. 12.

Table 2. Comparison of average classification error rate.

No.	Name of Novel	Decision Tree	Deep Learning	Naïve Bayes	SVM	RNN
1	Dark Tower	14.29	14.51	34.89	11.94	8
2	Fellowship	6.34	7.27	32.43	6.27	5.3
3	Game of Throne	10.43	11.86	46.23	10.5	8.5
4	Harry Potter and Prisoner of Azkaban	4.76	5.4	43.01	4.73	4.2
5	Hobbit	7.54	10.41	36.23	7.7	6.9
6	Edge of Eternity	26.96	43.26	58.66	26.91	17.4
7	Fifty Shade	28.86	33.6	60.65	28.87	20
8	Me Before You	8.23	14.59	49.52	8.23	5.7
9	The Girl on The Train	26.12	33.49	61.84	26.09	17.7
10	The Nightingale	4.65	8.28	57.03	4.59	3.8
11	Catching Fire	13.27	17.36	54.16	13.3	9.2
12	Hunger Game	10.24	9.67	46.92	10.26	7.2
13	Mockingjay	10.24	9.67	46.92	10.26	6.2
14	The Martian	7.52	8.39	43.26	7.56	5.7
15	The Running Man	14.38	14.05	41.02	12.81	10.6
Accuracy average		12.93	16.13	47.52	12.67	9.1

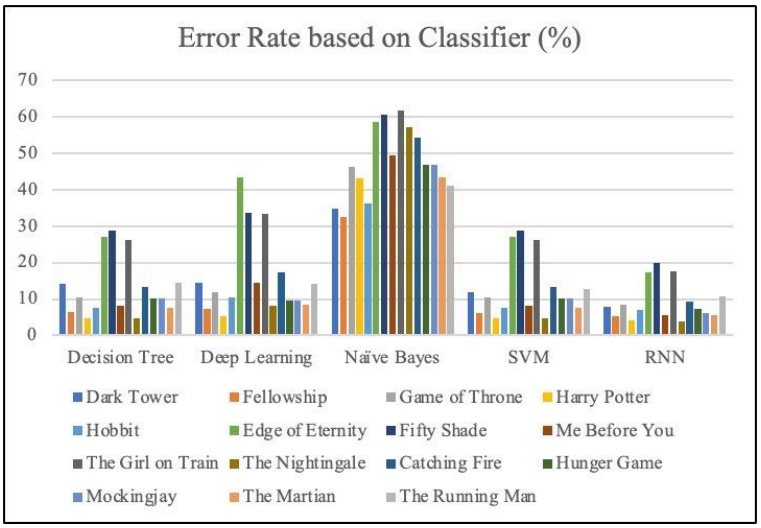


Fig. 12. Error rate for each classifier.

Table 2 and Fig. 12 show the best average error rate obtained by the RNN classifier, 9.1, and the worst result is Naïve Bayes at 47.52. Other classifiers, SVM, Decision Tree, and Deep Learning, obtain similar results like RNN at 12.67, 12.93, and 16.13, respectively.

5. Conclusion

This paper explains the efforts of creating a novel recommendation system based on user preferences using sentiment analysis. Several procedures were performed prior to create and train a model with classifiers from transforming, cleaning, and tokenization data to filtering and text vectorization. The data extracted from the Amazon website has been used to evaluate the proposed system whereby Each book contains over 20,000 reviews and comments, hence web Scraper addon for a web browser is used to gather the data. To select the most accurate classifier for the model, the accuracy average for each classifier is computed. The classification result using RNN has shown higher precision.

References

1. Clark, C.; and Rumbold, K. (2006). *Reading for pleasure: A research overview*. National Literacy Trust.
2. Yakavonitey, D.; and Kilyuvieney, D. (2007). Reading of books and their discussion in class as a precondition to high-quality communication. *Lifelong Learning: Lifelong Learning for Sustainable Development*, 5, 80-84.
3. Laila, N.A. (2009). Pengaruh Pendekatan CTL (Contextual Teaching and Learning) terhadap Hasil Belajar Membaca Pemahaman Bahasa Indonesia Siswa Kelas IV SD. *Jurnal Cakrawala Pendidikan*, 3(3), 238-248.
4. Cate, F.H.; and Mayer-Schönberger, V. (2013). Notice and consent in a world of big data. *International Data Privacy Law*, 3(2), 67-73.

5. Thelwall, M.; and Kousha, K. (2017). Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology*, 68(4), 972-983.
6. Lin, T.M.Y.; Luarn, P.; and Huang, Y.K. (2005). Effect of internet book reviews on purchase intention: a focus group study. *The Journal of Academic Librarianship*, 31(5), 461-468.
7. Rebora, S.; and Pianzola, F. (2018). A new research programme for reading research: analysing comments in the margins on Wattpad. *DigitCult-Scientific Journal on Digital Cultures*, 3(2), 19-36.
8. Wolfe, J.L. (2000). Effects of annotations on student readers and writers. *Proceedings of the fifth ACM conference on Digital libraries*. San Antonio Texas USA, 19-26.
9. Zhu, X.; Chen, B.; Avadhanam, R.M., Shui, H., and Zhang, R.Z. (2020). Reading and connecting using social annotation in online classes. *Information and Learning Sciences*, 121(5/6), 261-271.
10. Akhtar; M.S.; Gupta, D.; Ekbal, A.; and Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. *Knowledge-Based Systems*, 125, 116-135.