International Journal of Research (IJR)

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 9 Issue 02
February 2022

Classifying Amazon Product Reviews into Positive and Negative Classes Using TF-IDF Features

Ankit More & Mr.Girish Gogate

Department, of Computer Science,
Laxmi Narayan College of Technology (RIT),

Indore MP India

ankitmore17@gmail.com

Abstract – Text mining and its approaches are useful in a variety of real-world applications, including medical information retrieval, text analysis for fraud detection, and customer management in corporate intelligence applications. The data mining and natural language processing processes aid in the processing of data and the identification of required patterns in these applications. Text mining algorithms are utilized to classify Amazon product reviews in this project. This analysis may be useful for purchasers in making purchasing decisions, as well as for offering product feedback to the product producer. The Amazon product review dataset from Kaggle is used in this context. Preprocessing of the data set removes stop words and special characters from the text data. Following that, there will be two different steps. The Part of Speech Tagging (POS) tagging technique based on a Natural Language Processing (NLP) parser was used for feature extraction, and then the Term Frequency-Inverted Document Frequency (TF-IDF) based feature selection technique was utilized to identify probable keywords from the reviews. Both characteristics are then integrated, and the combined data features are utilized to create training and testing datasets. The training dataset is then used to train the Support vector machine (SVM), and the test dataset created after that is used to validate the performance. Experiments are conducted out with samples of varying sizes, increasing in size. Performance in terms of precision, recall, F1-score, time, and memory were also assessed. the effectiveness of

The model has a higher level of accuracy and uses fewer resources. Finally, some future extensions of the work are offered based on the experimental investigation.

Keywords*:* Amazon product review, natural language processing, supervised learning, feature selection, and text classification are some of the terms used.

INTRODUCTION

The text mining is an essential domain of data mining. In this domain the different algorithms has applied on data to recover the patterns from the text documents or sentences. The utilization of text mining techniques reduces the human efforts and improves the productivity of system. Therefore these techniques are frequently utilized in various real world applications. These techniques are helpful in extraction of information, classification and categorization of data. Therefore medical, engineering, business and finance are various domains where these techniques are applied. In this presented work the text mining technique is applied for classifying the reviews of the Amazon product reviews based on their positive and negative review orientation.

Basically in the ecommerce platforms the number of products are available buyers are looking for best products in low price. In this context a number of buyers are usages product reviews to make decisions. Therefore the product reviews are essential factor for influencing a product buyer's decision. Therefore to maintain quality of service and product credibility the review analysis is a useful tool to understand the buyer's requirement and improvement in their services and product. In this context the proposed work is intended to make an automatic review classification system using supervised learning techniques based on the previous buyer's sentiments.

PROPOED WORK

The main aim is to implement and simulate the data mining technique for classifying the Amazon product review according to their emotional orientations. Therefore a hybrid feature selection technique is proposed. This section explains how the entire process has been taken place.

A. System Overview

The buyers are reading the reviews for getting best and suitable product and services from the online market places such a

s Amazon. Basically a review consists of buyer sentiments about the particular product or service. Using the analysis of reviews user can get best suitable products as well as product owner can know about the consumers' feedback about the products. In this context review is beneficial for both the product owner and consumer. In such kind of problem of data analysis the data mining and machine learning has been used with natural language processing (NLP) for their emotion classification. Based on emotion classification the buyer's sentiments are discovered towards the particular product or service.

In this presented work the Amazon product reviews are analyzed for getting the buyers sentiments about products i.e. positive or negative. In this context first the Amazon product reviews are collected from the online data sources (i.e. Kaggle). In next process the data is preprocessed for refining the content and reducing noise from the data. After preprocessing the features are extracted. The features are basically the vector representation of the text content therefore we utilized the TF-IDF and Part of Speech (POS) tagging. Further the support vector machine (SVM) has been used for classifying the extracted features from both the feature extraction techniques. Thus a combined feature is developed for training and classification of the both kinds of features. The advantage of this feature is that it contains the data from their lexical structure as well as syntactical structure. In this section an overview of the required model is presented the next section explain the entire processes involved in the proposed Amazon review classification system.

B. Methodology

The proposed model for Amazon product review classification has demonstrated in figure 2.1. The figure consists of different functional components and the flow of data.

Amazon Product review dataset: the Amazon product review dataset is collected from a well known data repository named as Kaggle. This dataset contains the product reviews in a row of dataset which also contains the labels for each row of data. The label_2 indicates the positive review and the label_1 shows the negative reviews in the dataset.

Data preprocessing: the data preprocessing is an essential step of the data mining and learning. The noisy data can impact on the learning performance of the algorithms. Therefore the preprocessing techniques are used to clean the data. In this work, two steps of data preprocessing has been implemented. First the special characters from the review text are removed and then the stop words are removed. The clean data is further being used with two different feature selection techniques to transform the data.

POS Tagging: the POS Tagging is also known as Part of Speech Tagging. That is explained in previous chapter in details. The POS tagging produces the part of speech information of given text sentences. Based on the part of speech frequency of attributes the features are constructed. For example in a sentence how many times a noun word occurred, and number of pronoun in a sentences.

TF-IDF based features: the TF-IDF is also known as Term Frequency and Inverted Document Frequency. In this step the preprocessed data is taken into account and TF and IDF of each term in reviews are measured. Further based on term weight W,

$$W = TF * IDF$$

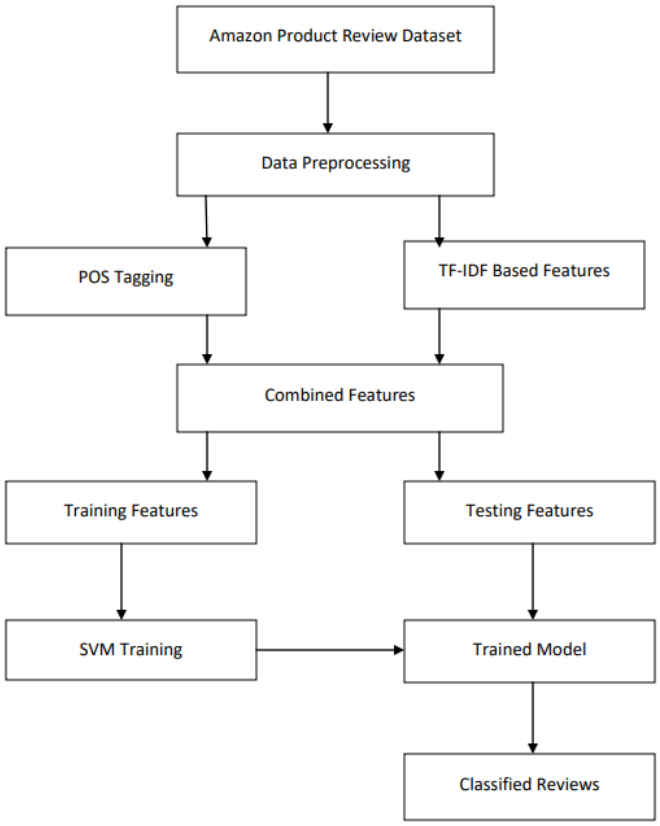

Figure 2.1 proposed system architecture

After weight computation based on the weights of terms the most highly weighted terms are selected. But the length of reviews are different (i.e. the number of tokens in each reviews) are different therefore the total 15 higher weighted

keywords from each review is selected for training and testing dataset preparations.

Combined features: the extracted features obtained from the POS tagging and TF-IDF based selected keywords are combined for each review instances of the data.

Training features: the combined feature vector which consist of POS tags and TF-IDF based selected keywords into an instance is split here in two parts. First part contains 70% of data instances for utilizing in training session of supervised learning classifier. In this experiment the 15 keywords and 7 POS tags as attributes of the dataset.

Testing features: the 30% of entire feature vector is written into a CSV file as the test dataset. The dataset is further used for validation of the trained SVM classifier. The validation results are finally used for performance evaluation task also. Therefore test dataset preparations are an essential task of classification modeling.

SVM Training: the Support Vector Machine (SVM) is a good algorithm for binary classification problems. Additionally our training dataset contains two classes i.e. positive and negative. The training process involves the training features which are separated in previous step. The training of the SVM classifier has been carried out using the combined features which contains the attributes of POS tagging as well as most weighted keywords as the training vector combination. Additionally with the training vector the class labels are also presented in the same training dataset.

Trained model: the output of the training of SVM classifier is given here as the trained model. The trained SVM model accepts the test data features of Amazon product reviews one by one instance. Additionally based on the training sample patterns, the trained model tries to recognize the negative and positive class labels for the reviews, according to their hidden sentiments.

Classified Reviews: this phase returns the classified Amazon reviews with their predicted class labels. Based on predicted class labels and actual class label associated with the data are used for evaluation of the performance in terms of precision, recall and F1-Score.

C. Proposed Algorithm

The proposed data mining technique for classifying the Amazon product reviews are given in previous section. This section provides the steps of working using table 2.1.

Table 2.1 proposed algorithm

Input: Amazon Product review A

Output: classified Review class C

Process:

1. $A_n = readDataset(A)$
2. $for(i = 1; i < n; i++)$
   i. $P_i = preprocessData(A_i)$
   ii. $Tg_i = POS.Tag(P_i)$
   iii. $W_i = TFIDF(P_i)$
   iv. $F_i = Tg_i + W_i$
3. $end\ for$
4. $[D_{train}, D_{test}] = Split(F_i)$
5. $T_{model} = SVM.Train(D_{train})$
6. $for(j = 1; j < D_{test}.size; j++)$
   i. $C_j = T_{model}.predict(D_{test_j})$
7. $end\ for$
8. Return C

I. RESULTS ANALYSIS

The implemented system is evaluated for finding the performance. Therefore this section includes the details about the evaluated performance parameters.

A. Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. These metric measures among a class label, how many actually predicted. High precision relates to the low false positive rate.

$$Precision = \frac{TP}{TP + FP}$$

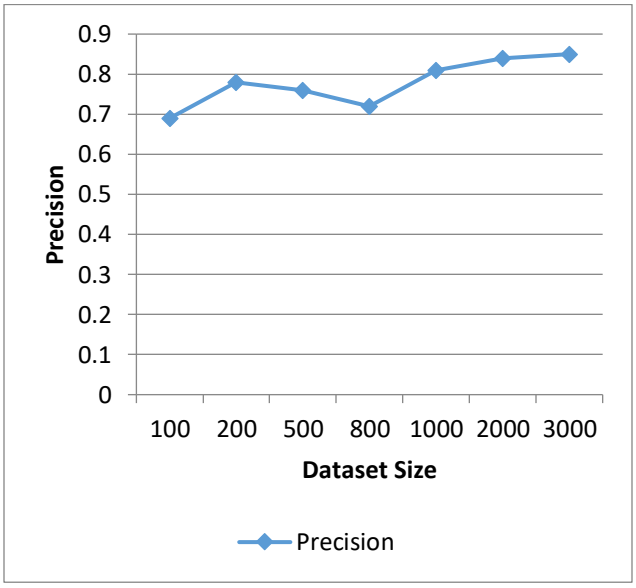

Figure 3.1 Precision

International Journal of Research (IJR)
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 9 Issue 02
February 2022

The precision of the proposed Amazon product review classification has explained in this section. The table 3.1 shows the observations of experiments and figure 3.1 represents the line graph of the precision. According to the line graph X axis contains the size of input data samples for training. Additionally the Y axis shows the precision of the algorithm. The precision of the algorithm has varying with the size of data, because the data may also involve new words and additional noise which requires more processing.

Table 3.1 Precision

Dataset Size	Precision
100	0.69
200	0.78
500	0.76
800	0.72
1000	0.81
2000	0.84
3000	0.85

B. Recall

Recall is also known as Sensitivity. Recall is the ratio of correctly predicted positive observations to the all observations in actual class. The recall answers: Of all the class that truly survived, how many did we label.

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

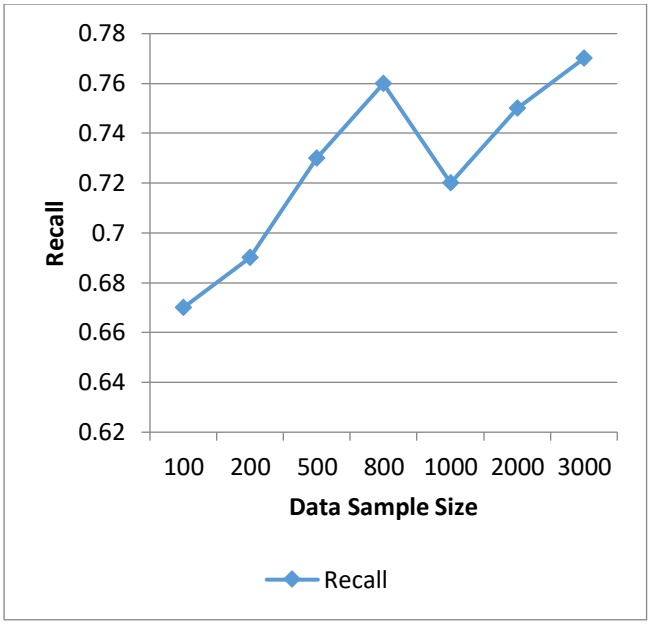

Figure 3.2 Recall

The recall of the algorithm is demonstrated using figure 3.2 and table 3.2. The line graph contains training sample size in X axis and Y axis shows the recall of the algorithm. The recall of the algorithm shows improvement in their recall rate. But due to the noisy attributes the algorithm shows the fluctuations. However the performance is acceptable for review analysis in binary classification problem (positive or negative).

Table 3.2 Recall

Dataset Size	Recall
100	0.67
200	0.69
500	0.73
800	0.76
1000	0.72
2000	0.75
3000	0.77

C. F1-Score

F1 Score is the weighted average of Precision and Recall. This score takes both false positives and false negatives into

account. F1 is usually useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

Table 3.3 F1-Score

Dataset Size	F1-Score
100	0.6798
200	0.7322
500	0.7466
800	0.7394
1000	0.7623
2000	0.7924
3000	0.8080

The F1-score shows the consistency in the classifier's outcome. The performance of the Amazon product review model demonstrates the consistent and increasing performance. The table 3.3 shows the observations of the experiment and figure 3.3 shows the line graph representation of f1-score. In this graph X axis shows the training samples and Y axis shows the F1-score obtained based on precision and recall. According to the obtained performance the results in good trend and with the increasing amount of training samples the improvement we can observe.

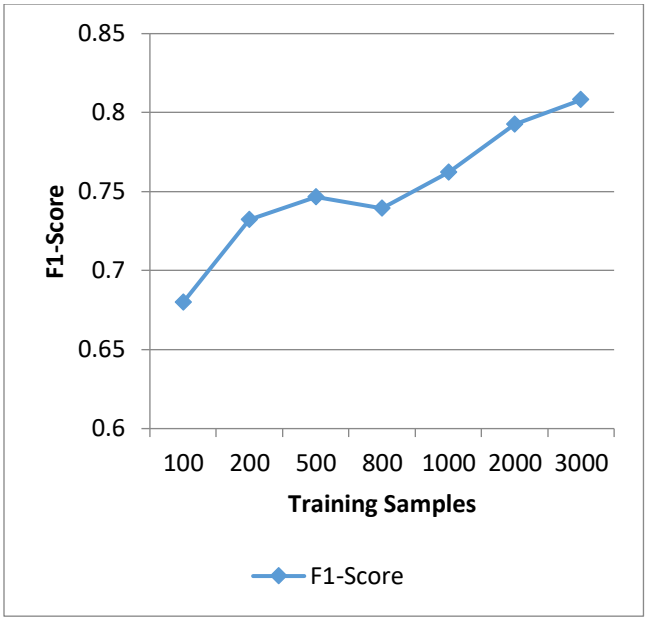

Figure 3.3 F1-Score

D. Memory Usages

The memory usages show the utilization of main memory during the execution of the algorithms. In JAVA based implementation we can compute the memory usages using the following formula:

$$memory\ usage = total\ assigned - total\ free$$

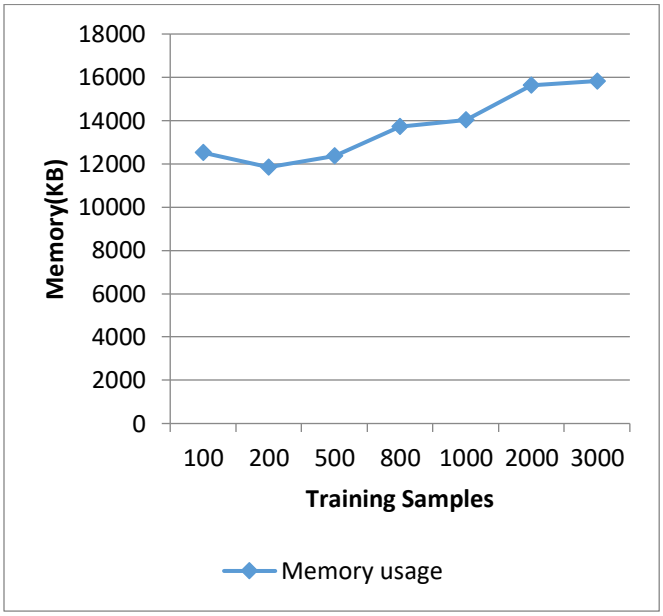

Figure 3.4 Memory usages

The experiments have conducted with the increasing amount of training samples the obtained results during the experiments are observed and reported in table 3.4. Using

International Journal of Research (IJR)
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 9 Issue 02
February 2022

these observations the line graph has been made as given in figure 3.4. The X axis of this diagram contains the training sample size and Y axis shows the memory usages. The memory usage of the algorithm has measured in terms of kilobytes (KB). According to the results the memory usages of the algorithm have increases with the amount of dataset samples were used. But the memory is not varying much thus the memory usages is acceptable for the given performance.

Table 3.4 Memory Usage

Dataset Size	Memory usage
100	12513
200	11847
500	12361
800	13725
1000	14028
2000	15632
3000	15829

E. Time Consumed

Time is an essential parameter for performance evaluation. The time is demonstrating the amount of time consumed during training of the model. The time consumption of the training of classifier is measured using the following formula.

$$time\ consumed = End\ Time - Start\ Time$$

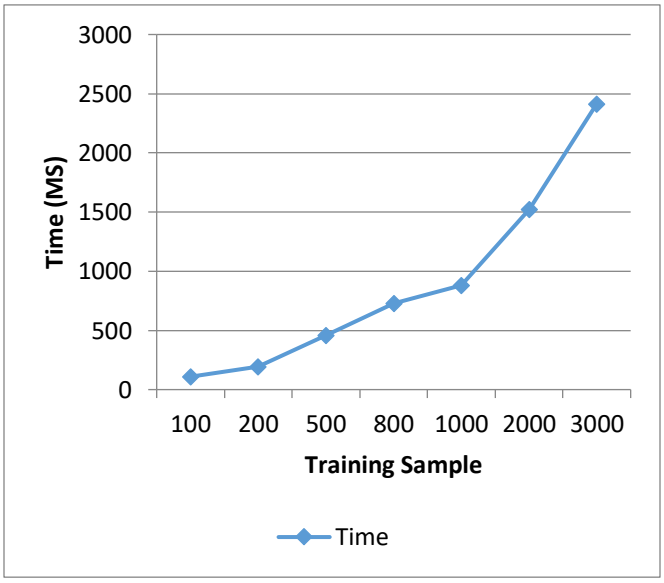

Figure 3.5 Training Time

The time consumed is also known as time complexity of the algorithms. The time requirement of the algorithm for increasing size of training sample is explained in figure 3.5 and table 3.5. The time consumption of the model is measured in terms of milliseconds (MS). The X axis of this diagram contains the training sample size and Y axis contains the time consumed. According to the results the time consumption of the algorithm increases with data size.

Table 3.5 Training Time

Dataset Size	Time
100	109
200	194
500	461
800	728
1000	883
2000	1524
3000	2415

II. CONCLUSION

The Amazon product review analysis model has implemented successfully. This section provides the conclusion of the entire effort carried out during study. Additionally based on experimental observations the future extensions of the work have also been proposed.

A. Conclusion

In business intelligence and other processes the product review is an essential part of identify the brand image of product manufacture. The product reviews can help to brand owners to improve the product quality and service. But manually reviewing a significant amount text data may lead the time consumption and manual mistakes. Therefore we need an automated product review analysis system. In this context the proposed work is motivated to design and develop an intelligent model which classifies the negative and positive reviews. In this context the classical text feature selection technique and NLP based features are used for improving the classification accuracy. Additionally supervised learning model has been used for finding accurate outcomes.

The proposed model is GUI based application which involves the interactive application interface for operating the model with custom datasets. The proposed work includes a real world review of the products based on Amazon. The dataset is downloaded from the Kaggle database. The product review is first used with a preprocessing technique which filters the noisy contents in terms of stop words and special characters. The filtered data is further used with two different feature selection techniques namely TF-IDF and POS tagging. Further the combined features are utilized with the SVM classifiers for taking training and testing. The trained model can be used for classifying the reviews in terms of positive and negative.

The implementation of the proposed system has been carried out using the JAVA technology. Additionally to store the performance the MySQL server is used. The implemented model has evaluated with different size of data and the recorded performance is summarized in table 4.1.

Table 4.1 performance summary

S. No.	Parameters	Observation
1	Precision	The precision of the model has varying with the different size of training samples
2	Recall	The recall is increasing with the amount of training sample size but the noisy contents can impact the performance of classifier
3	F1-Score	The performance of the model found consistent and shows the increasing trends
4	Memory	The memory consumption of the model is increasing in proportion with dataset samples
5	Time	The time consumption is also increases with the size of training sample. Here the time consumption has measured after obtaining the filtered features

According to the obtained results as summarized in table 6.1, the performance of the implemented model have found acceptable. That works fine but the current model still needs some corrections to be made for real world application development. Thus the next section provides the future extension of the proposed work.

B. Future work

The proposed work is motivated to classify the Amazon product review text into the negative and positive classes in order to know the brand image of a company. The model is successfully implemented and tested with the supervised learning classifier and real Amazon reviews. The model is promising but the following improvement has been proposed for work.

1. Modification in the use of extracted features from the text obtained from TF-IDF and POS tagging for performance improvement in terms of accuracy
2. Apply some ensemble learning technique for improving classification accuracy.

REFERENCES

[1] T. U. Haque, N. N. Saber, F. M. Shah, “Sentiment Analysis on Large Scale Amazon Product Reviews”, 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 978-1-5386-5283-1/18/$31.00 ©2018 IEEE

[2] J. Han, J. Pei, M. Kamber, “Data mining: concepts and techniques”, Elsevier, 2011.

[3] B. M. Ramageri, “Data Mining Techniques and Applications”, Indian Journal of Computer Science and Engineering, Volume 1 Number 4, pp. 301-305

[4] N. Venkata Sailaja and L. Padmasree, “Survey of Text Mining Techniques, Challenges and their Applications”, International Journal of Computer Applications (IJCA), Volume 146 – No.11, July 2016.

[5] E. M.G. Younis, “Sentiment Analysis and Text Mining for Social Media Micro-blogs using Open Source Tools: An Empirical Study”, International Journal of Computer Applications, Volume 112 – No. 5, February 2015.

[6] V. Gupta, G. S. Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, Volume 1, No. 1, PP. 60-76, August 2009.

[7] L. kumar, P. Bhatia, “Text Mining: concepts, process and applications “, Journal of global research in computer science, PP.36-39, March 2013.

[8] M. Radovanović, M. Ivanović, “Text Mining: Approaches And Applications”, Abstract Methods and Applications in Computer Science (no. 144017A), Novi Sad, Serbia, Vol. 38, No. 3, 2008, 227-234

[9] K. L. Sumathy and M. Chidambaram, “Text Mining: Concepts, Applications, Tools and Issues – An Overview”, International Journal of Computer Applications (IJCA) Volume 80 – No.4, October 2013.

[10] “What are the main Applications of Text Data Mining and Analysis?”, https://www.promptcloud.com/blog/9-best-examples-of-text-mining-analysis/

[11] M. Syamala, N. J. Nalini, “A Filter Based Improved Decision Tree Sentiment Classification Model for RealTime Amazon

Product Review Data", International Journal of Intelligent Engineering and Systems, Vol.13, No.1, 2020
[12] N. Shrestha, F. Nasoz, "Deep Learning Sentiment Analysis of Amazon.Com Reviews and Ratings", International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.8, No.1, February 2019
[13] C. S. G. Khoo, S. B. Johnkhan, "Lexicon-Based Sentiment Analysis: Comparative Evaluation of Six Sentiment Lexicons", Journal of Information Science, 1–21, 2016 permissions: sagepub.co.uk/journalsPermissions.nav DOI: 10.1177/0165551510000000
[14] N. Nandal, R. Tanwar, J. Pruthi, "Machine learning based aspect level sentiment analysis for Amazon products", Spat. Inf. Res., Korean Spatial Information Society 2020, https://doi.org/10.1007/s41324-020-00320-2
[15] M. Mishra, J. Chopde, M. Shah, P. Parikh, R. C. Babu, J. Woo, "Big Data Predictive Analysis of Amazon Product Review", KSII, The 14th Asia Pacific International Conference on Information Science and Technology (APIC-IST) 2019.
[16] A. S. Rathor, A. Agarwal, P. Dimri, "Comparative Study of Machine Learning Approaches for Amazon Reviews", Procedia Computer Science 132 (2018) 1552–1561
[17] X. Wei, H. Lin, Y. Yu, L. Yang, "Low-Resource Cross-Domain Product Review Sentiment Classification Based on a CNN with an Auxiliary Large-Scale Corpus", Algorithms 2017, 10, 81; doi:10.3390/a10030081
[18] R. S. Jagdale, V. S. Shirsat, S. N. Deshmukh, "Sentiment Analysis on Product Reviews Using Machine Learning Techniques", Cognitive Informatics and Soft Computing, Advances in Intelligent Systems and Computing 768, © Springer Nature Singapore Pte Ltd. 2019
[19] K. S. Srujan, S. S. Nikhil, H. Raghav Rao, K. Karthik, B. S. Harish, and H. M. Keerthi Kumar, "Classification of Amazon Book Reviews Based on Sentiment Analysis", Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing 672, Springer Nature Singapore Pte Ltd. 2018
[20] N. Z. Dina, N. Juniarta, "Aspect based Sentiment Analysis of Employee's Review Experience", Journal of Information Systems Engineering and Business Intelligence Vol.6, No.1, April 2020
[21] "The TF*IDF Algorithm Explained", https://www.onely.com/blog/what-is-tf-idf/#:~:text=TF*IDF%20is%20used%20by,b)%20Cocaine.
[22] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-canada: Building the state-of-the-art in sentiment analysis of tweets", in Proceedings of the seventh international workshop on Semantic Evaluation Exercises, 2013, pp. 321–327.
[23] R. Berwick, Village Idiot, "An Idiot's guide to Support vector machines (SVMs)", http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf
[24] V. A. Kharde, S.S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications (IJCA) Volume 139 – No.11, April 2016
[25] Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment Treebank" Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2013.
[26] Pang, B.and Lee, L. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts", 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04). 2004, PP. 271-278.
[27] F. Sebastiani. Machine learning in automated text categorization, ACM Computer Survey, 34(1):1–47, March 2002.
[28] K. Nigam, A. K. Mccallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled documents using Machine Learning, 39(2-3), PP. 103– 134, 2000.
[29] T. Cover, P. Hart, Nearest neighbor pattern classification, Information Theory, IEEE Transactions on, 13(1):21–27, 1967.
[30] A. Hotho, A. Nrnberger, G. Paa, A brief survey of text mining, LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 20(1):19– 62, May 2005.
[31] J.R. Quinlan. Induction of decision trees, Machine Learning, 1(1):81–106, 1986.
[32] N. Cristianini, J. S. Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.