IJARSCT

ISSN (Online) 2581-9429

IJARSCT
Impact Factor: 6.252

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 1, June 2022

A Framework on Online Reviews Ranking Based on Set Theory for Mining Using Automated Pipeline

Mrs. K. Karthika[1], Maheswari. S[2], Karishmaa. S. T[3], Ethayasirphy. S[4]
Assistant Professor, Department of Science and Computer Engineering[1]
Students, Department of Science and Computer Engineering[2,3,4]
Anjalai Ammal Mahalingam Engineering College, Kovilvenni, Tiruvarur, India

Abstract: *With the rapid rise of e-commerce, a big number of things are being sold online, and a growing number of people are making purchases online. Users can get valuable information from online reviews before purchasing a product or making a purchase. We investigate the peculiarities of their behaviour based on their early reviews. In this study, customer feedback linked with various products is collected from several online shopping websites in order to forecast product ratings based on user feedback utilising opinion mining. We classified the product's lifespan into three segments at first (Early, majority and laggards). A person who posts a review at an early stage is considered to be an early responder to the product. The product reviews are analysed using machine learning techniques. They give comments, and products are subsequently recommended for purchase and sale based on that factor. Users can provide product reviews on popular e-commerce platforms like Flipkart, Myntra, Amazon, and many others. To purchase a product, the consumer will investigate to gain a deeper grasp of the product and how it works. The interpretation will be a very straightforward product with inferior, superior, and neutral product checks. This experiment is carried out using machine learning techniques. Sentiment Analysis is a type of market research in which customers are aware of their reaction to a product. Individual decision-makers, businesses, and governments can all benefit from the awareness of feeling.*

Keywords: Machine Learning, Classification, Linear Regression, Naïve Bayes, Random Forest, etc.

I. INTRODUCTION

Everything is going online in this modern era of technology and digitalization. Rather of stepping outside, people turn to the internet for everything from food to clothing, and from house to devices. As a result, the popularity of e-commerce platforms has soared. On these platforms, different brands provide a variety of products.

As a result, selecting a helpful and dependable product will be challenging. A user reads product reviews to learn more about the product and determine whether or not to buy it. A user is more likely to believe in the opinions and experiences of others. A person's decision to buy or cancel a product is usually dependent on the reviews. As a result, it is clear to demonstrate the significance of reviews. Although it will be difficult for vendors to update and enhance their products by going through thousands of evaluations. This is where machine learning enters the picture. Machine learning has been extensively studied in sectors such as healthcare analytics, business, sentiment analysis, and so on. Opinion mining is a computational technique that allows one to learn about a person's opinions on a product or an item. It is, in principle, a classification.

A method that highlights that a given review can communicate substandard, superior, or neutral feedback. This field of research has grown in popularity in this new era of internet expansion. However, there are several disadvantages to having such a vast amount of feedback. To begin with, even all of these online evaluations do not guarantee or have a warranty of the original goods. This is due to the fact that fictitious users can post fictitious comments and express fictitious thoughts. The second flaw is that the majority of online reviews and reviewers are unavailable.

This is oftentimes the cause of these retail platforms' demise. We may utilise sentiment analysis to evaluate the structural product, learn what customers like and dislike, and improve customer service. The difference between our products' opinions and those of our competitors' products, the ability to perceive information about current products, and the ability to save several hours of physical time conversion. The major goal is to categorise user reviews into positive and negative

DOI: 10.48175/IJARSCT-4609

attitudes so that the user can quickly decide whether or not to purchase a product. However, there have been a few different techniques of categorising the reviews.

II. EXISTING SYSTEM

To gather the rating from the consumers, the prior methods used the fuzzy Kano model or classic navel margin-based methods. In the fuzzy Kano model, the sentiment collaborative filtering technique is used. They use this strategy to show consumers several types of cheerful photographs so that they can voice their opinions. The product reviews are calculated using this procedure. Users were polled using the fuzzy Kano model or classic navel margin-based approaches. In the fuzzy Kano model, the sentiment collaborative filtering technique is used. This model is incapable of comprehending various sorts of smileys. The product reviews are calculated using this procedure.

The diagram below illustrates how CSS is a significant advance over the industry's current practises. Approaches that evaluate sentiment strength for individual features of a product and methods that aim to rate a review on a global level can also be distinguished. The majority of global review categorization solutions rely on machine learning techniques and solely examine the polarity of the review (positive/negative). More linguistic elements, such as intensification, negation, modality, and discourse structure, are used in solutions that aim for a more precise classification of reviews (e.g., three- or five-star ratings). A thorough categorization of existing approaches. This grouping is not exhaustive. One solution may be appropriate for multiple categories. The restrictions listed below have an impact on current SA-focused work.

The following are some of the limitations:

1. Knowledge of hierarchical structures;
2. Knowledge of product aspect relationships is not fully utilised.
3. Sentences or reviews that cover a wide range of topics related to complex emotions aren't handled successfully. Furthermore, the overall evaluation of the SA employing ML has not resulted in improved accuracy or more efficient training time. As a result, this research offered a useful SA for online product reviews. In the fuzzy kano model, it uses the sentiment collaborative filtering algorithm. They use this strategy to show consumers several types of cheerful photographs so that they can voice their opinions. The product reviews are calculated using this procedure.

III. DISADVANTAGE

- This strategy is disliked by most users.
- The text-based reviews are not supported. As a result, users are unable to adequately express themselves.
- They are unable to provide information regarding the product.

IV.PROPOSED SYSTEM

The proposed based on review sentiment has been based on a careful investigation of the product review. Machine learning methods such as KNN, support vector machine, Decision tree, and Random Forest are used to analyse the emotional content of the reviews. The sentiment classification is followed by the review analysis. We can forecast the product's future by analysing early reviews. The early reviewers are analysed using K-means with PageRank. The previous system's flaws are addressed using the random forest classification technique. The good and negative reviews are separated using the K-means algorithm. After analysing good and negative feedback, provide recommendations to the user on which aspects of the product could be improved. There have been numerous applications and enhancements to Sentiment Analysis methods suggested and deployed throughout the years.

The purpose of this article is to take a deeper look at the most popular strategies employed in retail, particularly in the E-Commerce sector, and to provide a comprehensive evaluation. There are two main study fields in sentiment classification: Machine Learning and Lexicon, each with its own subgroup. There have also been a few studies that combine these two strategies to get a higher level of efficiency in sentiment analysis.

DOI: 10.48175/IJARSCT-4609

V. ADVANTAGE

- Data Consistency
- User Friendly
- Data Analysis
- Secured Process
- Identified Users

VI. LITERATURE REVIEW

Pang, Lee, and Vaithyanathan proposed a sentiment categorization utilising machine learning techniques in a dataset of cinema reviews in a 2002 study [5]. They used the Nave Bayes, Max Entropy, and Help Vector Machine models to investigate sentiment analysis on monographs and data bigrams. In their experiment, SVM paired with unigram function extraction produced the greatest results. They had an accuracy rate of 82.9 percent.

Mulle and Collier [4] finalised the Sentiment Classification of Jewellery and Footwear Shoe Product Criticism in a 2004 paper. They compared the hybrid Support vector machine, Nave Bayes, Logistic regression, and decision tree approaches to feature extraction methods based on Lemmas and Osgood theory. In this investigation, the support vector machine produces the best results, with an accuracy of 86.6 percent. In the 2017 paper, Elmurngi and Gherbi proposed detecting fraudulent movie reviews.

The performance of SMVs, decision books, and knit for a corpus with stop words and a corpus without stop words was compared to Naive Bayes [11]. SVM looks to be the most precise, with 81.75 percent and 81.35 percent precision in both scenarios. Bijoyan Das and Sarit Gupta conducted an experiment in 2018 using SVM, TF-IDF, and the Next Word Negation with an Amazon Product Opinion dataset, which yielded an accuracy of 89 percent.

[10] Chakraborty. A comparative evaluation of several machine methodologies, morphological-based approaches, and the best with 81% accuracy [13]. We will examine and contrast different approaches for implementing Sentiment Analysis in this study, as well as their accuracies.

VII. PROCEDURES IN SENTIMENT ANALYSIS

When it comes to sentiment analysis, there are two main ways. It employs both machine learning and a word-book-based approach. The former relies on traditional symbolic methods for text classification and lexicon-based approaches, which are also used by machine learning techniques. Text learning can be classified into many categories depending on different learning methodologies and strategies, such as learning through case studies, root learning, and analogy. The machine learning strategy can be loosely divided into supervised and unsupervised learning strategies.

A) Machine Learning Approach

Machine learning is regarded as an important branch of artificial intelligence that operates by enforcing a code that enables the machine to comprehend. This method makes use of both morphological and emotive characteristics. It looks at sentiment analysis as a problem of periodic text categorization, with a variety of priming and categorization documentation. The prototype is told to predict the class grade for the most recent example. The decision tree, neural tree network, Nave Bayes, logistic regression, and Support Vector Machine are all popular classifiers. We use both supervised and unsupervised learning to create these classifiers.

B) Supervised Learning Method

Supervised learning is a machine learning technique that employs a labelled data collection. These tutoring records provide a few input data as well as intended output. New instances are then classified using machine learning classifiers. Many various types of techniques have been introduced and documented, some of which are mentioned in this section. Method of unsupervised learning Supervised learning necessitates a collection of tagged training that is not always available, but is available in some cases.

Such learning algorithms do not necessitate the use of labelled data. Lowly supervised learning makes use of a large percentage of untagged learning data and a small percentage of tagged data. Input training devices are included in non-monitored learning, but no anticipated yield values are communicated to them. Collection analysis and expectation-maximization algorithms are two instances of unsupervised learning.

C) Lexicon Based Approach

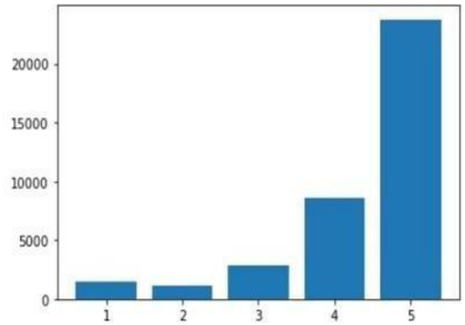
The Lexicon-Based method enables the detection of a point of view. Lexicon contains a large number of known and precompiled words of opinion [1]. This viewpoint lexicon can be used to study the text. The dictionary-based method, the corpus-based method, and the manual approach to opinion are the three ways that make up the lexicon-based approach [2]. This should be used in conjunction with the other two techniques.

VIII. METHODOLOGY

The majority of the steps of sentiment analysis utilising machine learning algorithms include the processing of twitter data. Following are the steps:

- Data Collection
- Data Preprocessing
- Clustering
- Feature Extraction
- Test Query

A) Data Collection



This stage requires retrieving information from Twitter in the form of tweets, which necessitates the setup of a Twitter account. Aside from that, in order to collect tweets, you must have Twitter permission. After obtaining authorization, we extracted tweets and stored them as.csv files. Web scraping was used to get data from an e-commerce site. Beautiful soup is the tool used to scrape the web. The building of the data base is based on the collecting of favourable and negative reviews.

B) Data Pre-processing

It is necessary to pre-process the data and remove any irrelevant information once the data has been collected in.csv files. There are various steps in the pre-processing process, including the following: Tokenization: URLs, hashtags, and at mentions are eliminated, and the string is divided down into tokens. It's usually used to keep track of how many times words appear.

Pre-processing the data is a must, and it necessitates the use of a technique known as data cleaning, which entails converting raw data into a machine-readable format. Separating each word from a sentence with a tokenization list. Stopping the removal of words: It means deleting numerous prepositions and articles that have little influence on the overall tone of the text.

C) Clustering

For clustering positive and negative reviews, the K means segmentation technique is used. We can cluster the group of words and sentences based on the collection group.

D) Extraction of Features

We use stemming tokenization based on sentimental words after collecting the data set. The process of converting raw data into numerical features that can be handled while keeping the original data set's information.

DOI: 10.48175/IJARSCT-4609

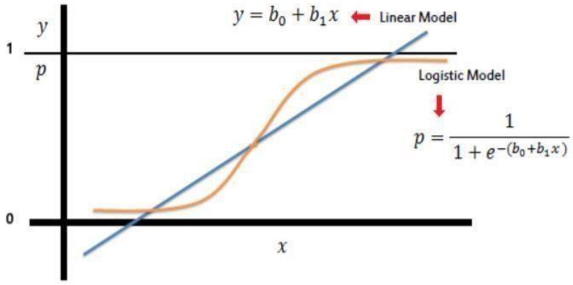
E) Test Query

The user purchases a product from the website and then leaves a review. In this proposed system validation of the outcome, the accuracy and validation graph are assessed.

IX. LOGISTIC REGRESSION

Calculated Relapse predicts the probability of a categorical subordinate variable. A parallel variable coded as yes or no is included in the subordinate. The huge test estimate is outperformed by calculated relapse. The calculated work could be a sigmoid work (as shown in the diagram), which accepts any genuine input x and returns an esteem value between 0 and 1.

$$\square (t) = \frac{1}{1 + e^{-t}}$$



Multinomial Naive Bayes

This is a supervised algorithm that categorises text into one of two categories: positive or negative [9] using a probabilistic technique. Before sorting tweets or text into different groups, this algorithm evaluates the probability of each word in the dataset. This algorithm employs the Bayesrule.

P(X/Y) = [P(X) P(Y/X)] / P(Y)

Where, x,y = Events
P(x|y) = Probability of x given y is true
P(y|x) = Probability of y given x is true
P(x),P(y) = Independent probabilities of x and y

Steps Involved:

- **a.** Data separation into training and test sets.
- **b.** Creating a vocabulary from the words in the training sets.
- **c.** Matching the content of the tweet with the vocabulary.
- **d.** Creating a feature vector.
- **e.** Training the classifier using the feature vector i.e., training the model.
- **f.** Testing the model with the test set.

Random Forest

The random forest classifier was chosen because it was at the top of a single decision tree in terms of reliability and efficiency [7]. It is an ensemble method that is based on bulging. This is how the categorizer works: (as shown in figure 3) The disposer generates k bootstrap D specimens initially, with Di representing each of the specimens, given D. A Di has a nearly equal number of D rows that are selected via D-substitution. By sampling with replacement, it reveals that a few real D rows may not be included in Di, although other tuples may appear multiple times. The classifier will then generate a decision tree based on each Di. Consequently, a "forest" made up of k decision trees is created. Each tree produces a single ballot enumerating its genre prognosis in order to identify an unknown tuple, X.

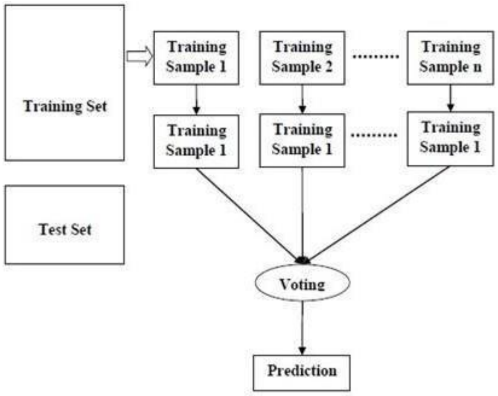


Figure 3: Working of Random Forest

X. PERFORMANCE EVALUATION

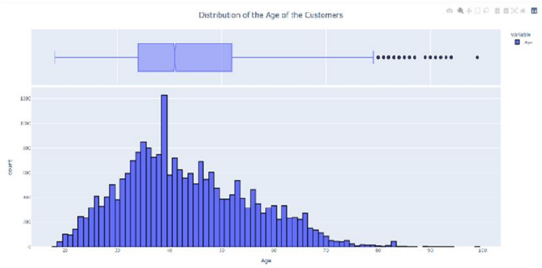
Evaluation metrics play a crucial role in determining a model's categorization efficiency. Precision measurement is the most extensively utilised approach for this aim. The efficiency of a sequencer on a given experimenting dataset is the proportion of those datasets that are properly grouped by the sequencer, and the accuracy measure of the text mining method is often insufficient to provide appropriate decision or result, so other metrics should be used to evaluate the classifier's output. Memory, precision, and F-measurement are three more important measures that are often employed.

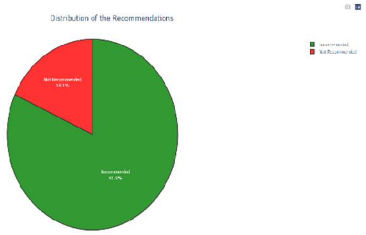
Method	Accuracy
Logistic Regression	90.88
Multinomial Naïve Bayes	87.09
Bernoulli Naive Bayes	86.46
Random Forest	93.17

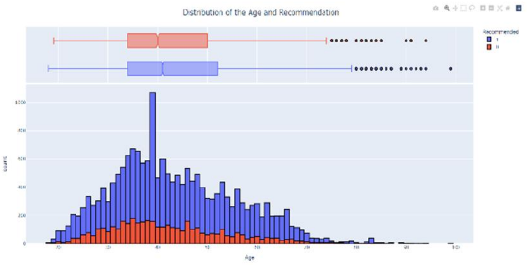
Figure 4: Depicts the Overall Conclusion of All Algorithms

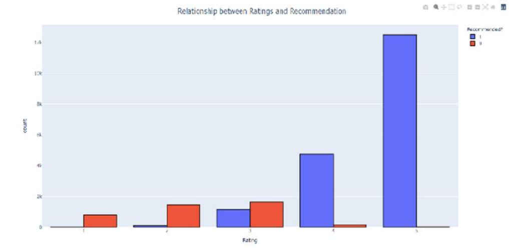
It comes to the conclusion that random forest is more efficient. Many advanced mathematical computations and algorithms have been discovered for sentiment recognition and analysis, which can optimise the data as accurately as feasible**.**

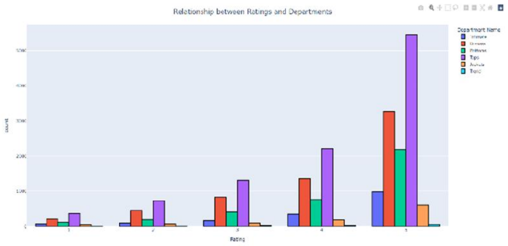
XI. IMPLEMENTATION AND RESULT

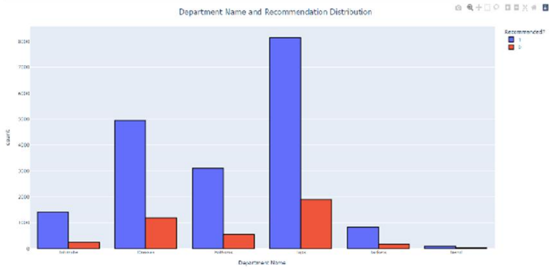


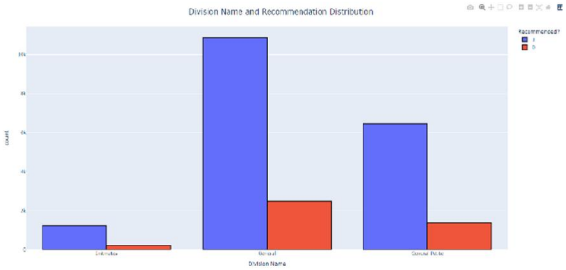


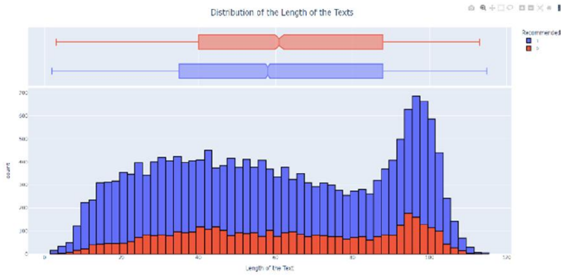












DOI: 10.48175/IJARSCT-4609

XII. CONCLUSION

Sentiment Analysis, also known as Opinion Mining, is the computer analysis of feelings, assumptions, and feelings expressed in the form of material, such as customer evaluations on a web-based shopping website. Opinion mining is becoming increasingly popular because the information it generates can help legitimise the products and services that are being used. After integrating the data with some neutral and negative opinions, four classification models were employed to identify reviews. i.e., Multinomial and Bernoulli Nave Bayes, Logistic Regression, and Random Forest, the prediction accuracy of Random Forest is determined to be the highest, with 93.17 percent accuracy.

The study could be expanded in the future to estimate a product's grades based on the viewpoint. Because the grades sustained by the goods and the emotion of the review sometimes do not fit each other, this would provide buyers with a reliable grading. By integrating our data in a more equal way. Item input is also important for creating better things, gifting, and analysing competitor contributions, all of which can have an instant impact on income.

REFERENCES

[1] Y. Liu, J.-W. Bi, and Z.-P. Fan, ‘‘Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory,’’ Inf. Fusion, vol. 36, pp. 149–161, Jul. 2017, DOI: 10.1016/j.inffus.2016.11.012.

[2] Y. Liu, J.-W. Bi, and Z.-P. Fan, ‘‘A method for ranking products through online reviews based on sentiment classification and interval-valued intuitionistic fuzzy TOPSIS,’’ Int. J. Inf. Technol. Decis. Making, vol. 16, no. 06, pp. 1497–1522, Nov. 2017, DOI: 10.1142/S021962201750033X.

[3] X. Liang, P. Liu, and Z. Liu, ‘‘Selecting products considering the regret behavior of consumer: A decision support model based on online ratings,’’ Symmetry, vol. 10, no. 5, p. 178, May 2018, DOI: 10.3390/sym10050178.

[4] E. Najmi, K. Hashmi, Z. Malik, A. Rezgui, and H. U. Khan, ‘‘CAPRA: A comprehensive approach to product ranking using customer reviews,’’ Computing, vol. 97, no. 8, pp. 843–867, Aug. 2015, DOI: 10.1007/s00607-015-0439-8.

[5] C. Wu and D. Zhang, ‘‘Ranking products with IF-based sentiment word framework and TODIM method,’’ Kybernetes, vol. 48, no. 5, pp. 990–1010, May 2019, DOI: 10.1108/K-01-2018-0029.

[6] D. Zhang, C. Wu, and J. Liu, ‘‘Ranking products with online reviews: A novel method based on hesitant fuzzy set and sentiment word framework,’’ J. Oper. Res. Soc., vol. 71, no. 3, pp. 528–542, Mar. 2020, DOI: 10. 1080/01605682.2018.1557021.

[7] C. Guo, Z. Du, and X. Kou, ‘‘Products ranking through aspect-based sentiment analysis of online heterogeneous reviews,’’ J. Syst. Sci. Syst. Eng., 542–558, 2018, DOI: 10.1007/s11518-018- 5388-2.

[8] B. S. H. Karpurapu and L. Jololian, ‘‘A framework for social network sentiment analysis using big data analytics,’’ in Big Data and Visual Analytics. Springer, 2017, pp. 203–217.

[9] B. K. Shah, V. Kedia, R. Raut, S. Ansari, and A. Shroff, “Evaluation and Comparative Study of Edge Detection Techniques,” vol. 22, no. 5, pp. 6–15, 2020, DOI: 10.9790/0661- 2205030615.

[10] S. Thapa, P. Singh, D. K. Jain, N. Bharill, A. Gupta, and M. Prasad, “Data-Driven Approach based on Feature Selection Technique for Early Diagnosis of Alzheimer's Disease”, in 2020 International Joint Conference on Neural Networks (IJCNN), 2020: IEEE, 1-8, DOI: 10.1109/IJCNN48605.2020.9207359.

[11] P. D. Turney, ‘‘Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,’’ 2002, arXiv:cs/0212032. [Online]. Available: https://arxiv.org/abs/cs/0212032

[12] V. Vargas-Calderón, N. A. V. Sánchez, L. Calderón-Benavides, and J. E. Camargo, ‘‘Sentiment polarity classification of tweets using an extended dictionary,’’ Intel. Artif., vol. 21, no. 62, pp. 1–11, 2018, DOI: 10. 4114/intartif.vol21iss62pp1-12.

[13] F. H. Khan, U. Qamar, and S. Bashir, ‘‘ESAP: A decision support framework for enhanced sentiment analysis and polarity classification,’’ Inf. Sci., vols. 367–368, pp. 862–873, Nov. 2016.

[14] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, R. R. Suárez, and O. S. Siordia, ‘‘A simple approach to multilingual polarity classification in Twitter,’’ Pattern Recognit. Lett., vol. 94, pp. 68–74, Jul. 2017.

[15] B. Pang and L. Lee, ‘‘Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,’’ in Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics - ACL, Jun. 2005, pp. 115–124, DOI:

10.3115/1219840.1219855.

[16] S.-M. Kim and E. Hovy, ‘‘Determining the sentiment of opinions,’’ in Proc. 20th Int. Conf. Comput. Linguistics - COLING, 2004, p. 1367, DOI: 10.3115/1220355.1220555.

[17] O. Appel, F. Chiclana, J. Carter, and H. Fujita, ‘‘A hybrid approach to the sentiment analysis problem at the sentence level,’’ Knowl-Based Syst., vol. 108, pp. 110–124, Sep. 2016.

[18] T. Hayashi and H. Fujita, ‘‘Word embeddings-based sentence-level sentiment analysis considering word importance,’’ Acta Polytech. Hungarica, vol. 16, no. 7, p. 152, 2019.

[19] Y. Xia, E. Cambria, and A. Hussain, ‘‘AspNet: Aspect extraction by bootstrapping generalization and propagation using an aspect network,’’ Cognit. Comput., vol. 7, no. 2, pp. 241–253, Apr. 2015.

[20] Y. Wang, A. Sun, M. Huang, and X. Zhu, ‘‘Aspect-level sentiment analysis using AS-capsules,’’ in Proc. World Wide Web Conf. - WWW, 2019, pp. 2033–2044.

[21] M. D. P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. Á. Rodríguez-García, and R. Valencia-García, ‘‘Sentiment analysis on tweets about diabetes: An aspect-level approach,’’ Comput. Math. Methods Med., vol. 2017, pp. 1–9, Feb. 2017.

[22] D. Bollegala, D. Weir, and J. Carroll, ‘‘Cross-domain sentiment classification using a sentiment sensitive thesaurus,’’ IEEE Trans. Knowl. Data Eng., vol. 25, no. 8, pp. 1719–1731, Aug. 2013.

[23] J. Blitzer, M. Dredze, and F. Pereira, ‘‘Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification,’’ in Proc. 45th Annu. Meeting Assoc. Comput. Linguistics, Jun. 2007, pp. 440–447.

[24] N. Li, S. Zhai, Z. Zhang, and B. Liu, ‘‘Structural correspondence learning for cross-lingual sentiment classification with one-to-many mappings,’’ 2016, arXiv:1611.08737. [Online]. Available: http://arxiv.org/ abs/1611.08737

[25] R. González-Ibánez, S. Muresan, and N. Wacholder, ‘‘Identifying sarcasm in twitter: A closer look,’’ in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol., 2011, pp. 581–586.

[26] S. Thapa, S. Adhikari, U. Naseem, P. Singh, G. Bharathy, and M. Prasad, "Detecting Alzheimer’s Disease by Exploiting Linguistic Information from Nepali Transcript," in International Conference on Neural Information Processing, 2020: Springer, pp. 176-184.

[27] A. Ghimire, S. Thapa, A. K. Jha, S. Adhikari, and A. Kumar, “Accelerating Business Growth with Big Data and Artificial Intelligence”, in 2020 Fourth International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), 2020: IEEE, pp. 441-448, DOI: 10.1109/I-SMAC49090.2020.9243318.

[28] A. Ghimire, S. Thapa, A. K. Jha, A. Kumar, A. Kumar, and S. Adhikari, “AI and IoT Solutions for Tackling COVID-19 Pandemic”, in 2020 International Conference on Electronics, communication and Aerospace Technology, 2020: IEEE.

[29] J.-J. Peng, J.-Q. Wang, J. Wang, H.-Y. Zhang, and X.-H. Chen, ‘‘Simplified neuromorphic sets and their applications in multi-criteria group decision-making problems,’’ Int. J. Syst. Sci., vol. 47, no. 10, pp. 2342–2358, Jul. 2016, DOI: 10.1080/00207721.2014.994050.

[30] F. Altun, R. Şahin, and C. Güler, ‘‘Multi-criteria decision-making approach based on PROMETHEE with probabilistic simplified neutrosophic sets,’’ Soft Comput., vol. 24, no. 7, pp. 4899–4915, Apr. 2020, DOI: 10.1007/s00500-019-04244-4.