Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7691531

# TWITTER BASED SENTIMENT ANALYSIS FOR PERSPECTIVE AND RANKING OF ENGINEERING COLLEGE USING MACHINE LEARNING TECHNIQUE

**Shanta.H.Biradar***
Research Scholar, Department of CSE, ATME College of Engineering, Mysore,India
shantha_is@sirmvit.edu

**Dr.J.V.Gorabal**
Professor, Department of CSE, ATME College of Engineering, Mysore,India
jvgorabal@gmail.com

**Abstract :** Indian Technical institutes create naval knowledge and support social communities along with regular academic requirements. They play a significant role to increase engineering competitiveness from local to national level. National Institute Ranking framework (NIRF), India evaluates all technical institutions based on teaching, Learning & Resources, Research, Professional Practice and Collaborative Performance, Graduation Outcomes, Outreach and Perception. Sentiment analysis has wide scope in many domains including education, medical etc. Several research reports provide the state of the applications of sentiment analysis in industry, business, social and educational performance. But no / few work focused on ranking of Engineering College ranking using natural language processing, deep learning and machine learning solutions. The aim of the research work was to develop a sentiment analysis of the NIRF in order to enhance the performance of the ranking method. The work investigates the effect of NIRF five factors on ranking of engineering college values and brand attachments based on stakeholders such as students, parents and industries sentimental values. The work used 5002, 3051, 2821, 4252 and 3625 Twitter data for Learning & Resources, Research, Professional Practice and Collaborative Performance, Graduation Outcomes, Outreach and Perception respectively. This research work has applied Natural Language Processing (NLP) operations such as pre-processing, stop-word elimination, tf-idf transformation and n-gram model to bring the textual data to machine learning understandable format. Later state-of-the-art Machine Learning (ML) algorithms had to be applied following the topic modeling and extraction of the sentiment. This research work mostly focused on the features and key terms which will influence the prospective ranking of the educational institutions with their percentage of the contribution. performed machine learning on the 26678 tweets followers of 25 institutes were considered for the ranking process and conducted statistical verification with NIRF rankings.

**Keywords:** Sentiment analysis, Natural Language Process, Machine Learning, Institute ranking

## I. INTRODUCTION

Social networks are a plethora of information gathering and sharing platforms for educational institutes for their promotions and advertisements, which leads to a change in the general perspective of networking, socialization and personalization[1]. With the proliferation of

different social media, students are willing to give their comments on the quality of education and technical facilities after joining institute or graduation. Student's views are part and parcel of e-mouth, which is more credible than that of any advertisement or publicity. More reviews online about the institute, which leads to institutes gradually emerging due to the effort of management, faculty and trainers[2].

Many Indian technical institutes in the last decades promote inter-state and inter-country student mobility have been given high priorities, which is useful to educational top management, teachers and instructors in order to improve teaching and learning activities. Sentimental analysis (SA) used estimate the revenue generation of movies from the box-office[3], prediction of similarity with common attribution[4], estimate votes in Singapore presidential election[5], forecast electoral results[6], current geographical locations[7], public opinion towards massive online open courses, predict US presidential election[8], most active research topics[9], etc. Few SAs help to express students' feelings, opinion and problems with or without showing their identification [10]. Centrist think tank reported that nearly 6% of students drop out after their first year due to bad university selection [11].

Many online resources project university ranking on various websites. It helps to choose the right institute and course but many of these rankings are duped and have serious shortcomings due to it covering only 5% of all the universities in the world [12]. The institute ranking system used to compare institute teaching and learning productivity and performance by different techniques. Many of them criticized methods of estimation, data gathering procedures, data mining, data segregation etc. Therefore, an alternative method must be developed to analyse students, faculty, staff and alumni opinions about the institute. In this direction, the National Institutional Ranking framework (NIRF) developed a methodology to estimate the rank of the institute depending on whether five clusters are assigned weightages in resources, research and stakeholder perception. For stakeholder perception is measured based on social media such as Facebook, Twitter etc. Twitter is one of the promising micro blogging platforms used for efficiently analysing customer's opinions in various domains. Although many researchers [13-15] used Twitter data for various domains such as marketing, products, prediction of election results, research opportunities, only less or no work has been explored in the field of education. To fill the mentioned gap, the objective of the research work was to develop a sentimental analysis of the National Institute Ranking framework (NIRF) in order to enhance the performance of ranking methods.

## II. REVIEW OF RELATED RESEARCH WORK

The research article studies the emotional analysis of institution reviews is mainly to judge whether the students are satisfied or not with the institution online from two directions such as negative and positive emotions. Several studies were conducted to review the universities such as Naugen et.al [16] created university student feedback data for two years and they categorized them into three sensitive data such as positive, negative and neutral from 5000 classified sentences. They used three classifiers such as Naïve Bayes, Maximum Entropy, and support vector machine (SVM) for analysis. Among them, the support vector machine shows best results 91.36%. Mehmet Korkmaz [17] conducted research deployed for classifiers such

Naïve Bayes, SVM, Decision Tree and Random forest to verify the qualitative feedback after each semester. Among all techniques SVM Classification algorithm shows 63.73% accuracy. Duwairi & Qarqaz[18] and Altrabsheh et al[19] SVM classifiers showed highest accuracy and recommended for students feedback. Gottipati et al[20], Guleria et al[21], Hashim [22] and Abaidullah et al [23] showed that decision tree (78.1%), Association Rule Mining (79.8%),Apriori Algorithm(67%) and K-means clustering algorithm (74%) are given better results for student sentimental feedbacks respectively.

## III. APPROACH METHODOLOGY

Present methodology was using the unigram feature extraction technique for analysing twitter data set. The developed framework preprocessor used to transform raw sentences to more understandable sentences. Machine learning techniques used for training the processed data with feature vectors further synonyms and similarity sentences provided by semantic analysis which provides the polarity of the sentences. The detailed methodology of present work has been classified sub sections and the block diagram as shown in Fig. 1.

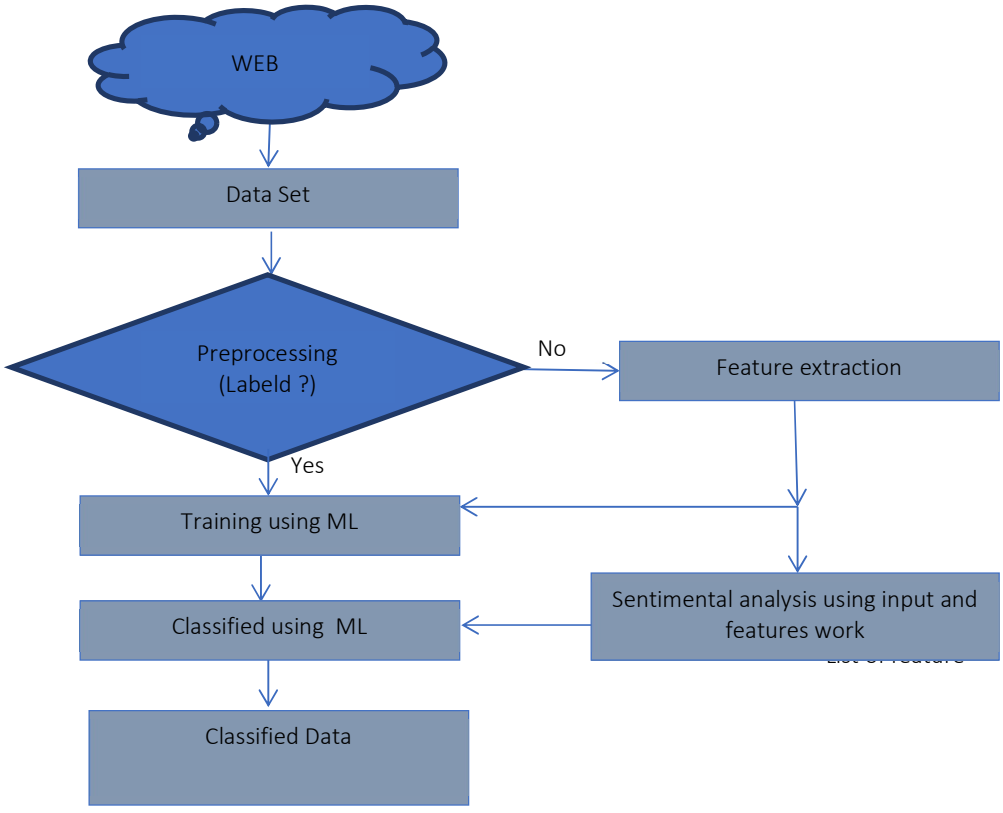


Fig. 1 Block diagaram of the Methodology for Ranking of Institute based on sentimental analysis.

The tweeter data was collected from the standard Twitter API to the chosen institutes in Bengaluru. The data was collected between July 1, 2020 to June 30, 2021, academic year consists of the summer and winter semester. The information data was used for present research work gathered from twitter various accounts. Nearly 5000 Twitter data for Learning and Resources post in English languages about different Karnataka Technical Institutes. Reviews were classified as per nature of the review under the heads of positive and negative sentences. Reviews are also categorized into five such as 1) Learning & Resources, 2) Research, 3) Professional Practice & Collaborative Performance, 4) Graduation Outcomes, 5) Outreach and Perception. Data of each category is given in Table 1.

Table 1 Twitter dataset for category classification

| Sl | Lable | Count |
|---|---|---|
| 01 | Learning & Resources | 5002 |
| 02 | Research | 3051 |
| 03 | Professional Practice &Collaborative Performance | 2821 |
| 04 | Graduation Outcomes | 4252 |
| 05 | Outreach and Perception | 3625 |

In the pre-processing task, the data contains various opinions about the institute, they are expressed differently. The twitter dataset may be labelled with polarity or raw data without polarity and redundancy, which leads to affect on the result or prediction hence it very much improves the quality of raw data. To improve the quality of the raw data is pre processed, such as removing the repeated words, punctuations and new symbols. To improve the quality of the dataset after preprocessing, which has more of distinctive properties, the feature extraction technique was used. It extracts adjectives and differentiate positive and negative polarity in a collected sentence, which is finally determining the opinion of individual stakeholders. The collected information data are two columns namely, Review, which given by the students and Liked column has either 0 (negative reiew) or 1 (positive review) shown in Table 2.

Table 2. The Twitter message considered for Sentimental analysis

| | | |
|---|---|---|
| 01 | Loved this Institute | 1 |
| 02 | Teaching and learning is not good | 0 |
| 03 | Selection of innovative clubs are great | 1 |
| . | | |
| . | | |

| | | |
|---|---|---|
| 751 | The whole experience was underwhelming | 0 |

The Twitter data was processed and analysed rreviews in natural language (NLP) using Natural Language Toolkits, which handles interaction between human language and computer program. The Twitter data was cleaned in five steps, to remove punctuation and special characters from the data set from the substrings. Then cleaning all stop-words which are connecting the sentence by using three NLP techniques such Lexical, Syntactic and Semantic analyses.

```
In [7]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import Technical Education Reviews Datasets
        data = pd.read_table(r"C:\\Users\\KRISHNA M\\Desktop\\Institutional_data.csv")
        print(data['liked'].value_counts())
        plt.figure(figsize =(3,5))
        sns.countplot(x=data.liked)
        x=data['Review'].values
        y=data['Liked'].values
        from sklearn.model_selection import train_test_split
        x_train, x_test, y_train, test=train_test_split(x,y,random_state=0)
        x_train.shape
        x_test.shape
        y_train.shape
        y_test.shape
        from skelearn.feature_extraction.text import countvectorizer
        vect=CountVectorizer(stop_words='english')
        x_train_vect=vect.fit_transform(x_train)
        x_test_vect=vect.transform(x_test)
        from sklearn.svm import SVC
        model = SVC()
        model.fit(x_train_vect, y_train)
        y_pred=model.predict(x_test_vect)
        accuracy_score(y_pred, y_test)
```

Fig. 2 A typical Python coading for senstimental analysis of three different distributions of training and testing modified the dataset in the ratio of 80:20, 75:25 and 70:30. The complete algorithm is given in Fig. 1 and Python code is given in Fig. 2.components.

Third step combines the words in meaningful sentences for analysing at a single instance. In the final step all prefixes and suffixes are eleminted using NLTK Finally cleaning of all the ambiguities, the bag-of-words are prepared along with root words. Further, aspects like Learning & Resources, Research, Professional Practice and Collaborative Performance, Graduation Outcomes, Outreach and Perception are defined. At the beginning, two variables such as positive-data and negative_data were gathered for each predefined aspect, which are assigned to zero score. After reviews pertaining to each category are picked up from the review set then corresponding key words are identified. Then the key words are counted against positive or negative sentiment analysis. Finally both positive and negative review of each aspect was computed.

## IV. RESULT AND DISCUSSION

The present work classification model has taken Technical Institute review as from Twitter account. The model was built for one of the prestigious institutes in Bangalore for extracting that institute reviews. The developed model is categorising each public review into one the categories such as excellent, Good, Moderate, Satisfactory and poor) based on the sentimental data. The developed model used to train 75% of the review data and the remaining 25% data

used for testing. The Fig. 3 (a) shows visualization of sentimental analysis from obtained Twitter data from the institution. Fig. 3(b) shows Python code for catograisation and visualization.

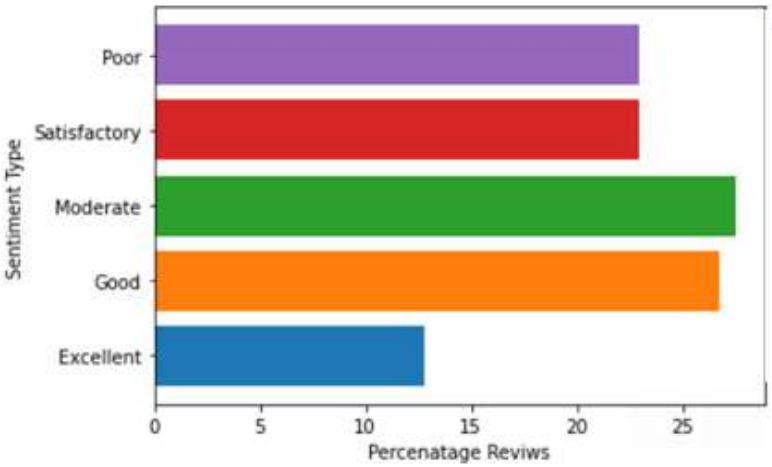


Fig. 3(a). Institutional Twitter data Analysis

```
In [23]: import matplotlib.pyplot as plt
         import numpy as np

         def inst_twitter(Exc_data, Good_data, Mod_data, Sat_data, poor_data)
         x = np.array(["Excellent"])
         y = np.array([Exc_data])
         plt.barh(x,y)
         x = np.array(["Good"])
         y = np.array([Good_data])
         plt.barh(x,y)
         x = np.array(["Moderate"])
         y = np.array([Mod_data])
         plt.barh(x,y)
         x = np.array(["Satisfactory"])
         y = np.array([Sat_data])
         plt.barh(x,y)
         x = np.array(["Poor",])
         y = np.array([poor_data])
         plt.barh(x,y)
         font1={'family' : 'serif', 'color': 'darkred', 'size':15}
         plt.xlabel("Percenatage Reviws")
         plt.ylabel("Sentiment Type")
         plt.show()
```

Fig. 3(a). Coading for visualisation

The present work classification model has taken Technical Institute review as from Twitter account. The model was built for one of the prestigious institutes in Bangalore for extracting that institute reviews. The developed model is categorising each public review into one the categories such as excellent, Good, Moderate, Satisfactory and poor) based on the sentimental data. The developed model used to train 75% of the review data and the Python horizontal plot shown in Fig. 3(a) and 12.8, 26.8, 27.8, 23.5 and 9.1 % of reviewers excellent, Good, Moderate, Satisfactory and poor respectively. It is observed that categories "Good", "Moderate" and "Satisfactory" have high precision scores because one number of reviews are likely to fall under these categories. But on the other hand "excellent" and "poor" have very low precision scores due to less number of reviews.

```
LR-Learning & Resources, R-Research, Pr-Professional practice,
Pe-Formance, GO-Graduation outcome
P-pasitive and N-Negetive
```

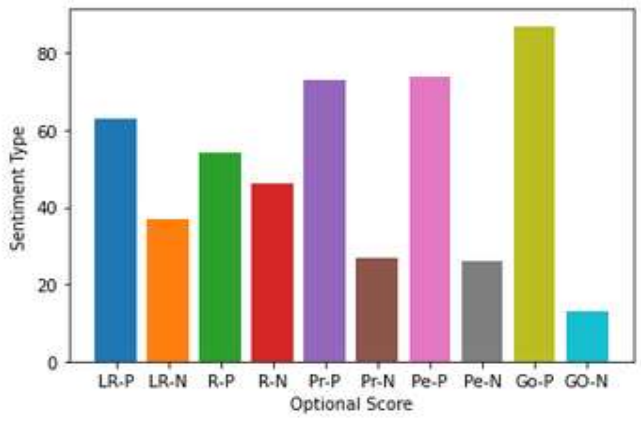


Fig. 4(a). Aspects of parameters of Institution

```
In [40]: import matplotlib.pyplot as plt
         import numpy as np
         x = np.array(["LR-P"])
         y = np.array([LR_P_Data])
         plt.barh(x,y)
         x = np.array(["LR-N"])
         y = np.array([LR_N_Data])
         plt.barh(x,y)
         x = np.array(["R-P"])
         y = np.array([R_P_Data])
         plt.barh(x,y)
         x = np.array(["R-N"])
         y = np.array([R_N_Data])
         plt.barh(x,y)
         x = np.array(["Pr-P"])
         y = np.array([Pr_P_Data])
         plt.barh(x,y)
         x = np.array(["Pr-N"])
         y = np.array([Pr_N_Data])
         plt.barh(x,y)
         x = np.array(["Pe-P"])
         y = np.array([Pe_P_Data])
         plt.barh(x,y)
         x = np.array(["Pe-N"])
         y = np.array([Pe_N_Data])
         plt.barh(x,y)
         x = np.array(["Go-P"])
         y = np.array([Go_P_Data])
         plt.barh(x,y)
         x = np.array(["GO-N"])
         y = np.array([Go_N_Data])
         plt.barh(x,y)
         print ("LR-Learning & Resources, R-Research, Pr-Professional practice, \nPe-Formance, GO-Graduation outcome")
         print("P-pasitive and N-Negetive")
         plt.xlabel("Optional Score")
         plt.ylabel("Sentiment Type")
         plt.show()
```

Fig. 4(b). Coading for visualisation

Fig. 4(a) shows the bar plot of the opinion score of aspects of Learning & Resources, Research, Professional Practice and Collaborative Performance, Graduation Outcomes. It clearly states positive and negative scores of all aspects as shown in Fig. 4(a) and code for distribution given in Fig. 4(b). The graph shows clearly the positive scores 63% and negative score 37% for Learning and Resources. Similarly for research positive score 54% and negative score 46%, Professional Practice for positive score 73 and negative score 27, Peformance positive 74% and negative -26 %, and Graduation outcome positive 87% and 13% negative scored.

Table 3. Statastically analysis of Sentimental Data

| | Nature of Score | Positive Score | Negative Score |
|---|---|---|---|
| 0 | Parameter Data | [63,54,73,74,87] | [37,46,27,26,15] |
| 1 | Mean | 70.2 | 30.2 |
| 2 | Median | 73.0 | 27.0 |
| 3 | Mode | ([54].[1]) | ([15].[1]) |
| 4 | Standard division | 11.124747 | 10.533755 |
| 5 | Variation | 123.76 | 110.96 |

```
In [76]: import numpy as np
         from scipy import stats
         import pandas as pd
         Pastive_score= pd.read_table(r"C:\\Users\\KRISHNA M\\Desktop\\P_Instituional_data.csv")
         p_mean = np.mean(pasitive_scores)
         p_median = np.median(pasitive_scores)
         p_mode=stats.mode(pasitive_scores)
         P_st = np.std(pasitive_scores)
         P_var = np.var(pasitive_scores)
         Negative_scores=pd.read_table(r"C:\\Users\\KRISHNA M\\Desktop\\N_Instituional_data.csv")
         N_mean = np.mean(Negative_scores)
         N_median = np.median(Negative_scores)
         N_mode=stats.mode(Negative_scores)
         N_st = np.std(Negative_scores)
         N_var = np.var(Negative_scores)
         data = {
             "Nature of Score" : ["Parameter Data", "Mean", "Median", "Mode", "Standard division", "Variatic
             "Positive_Score" : [ pasitive_scores,p_mean, p_median,p_mode, P_st, P_var ],
             "Negative_Score" : [ Negative_scores,N_mean, N_median,N_mode, N_st, N_var ]
```

Fig. 5 Python code for statastically analysis of Sentimental data

Table 3 shows the statistically analysis of sentimental dataset, which observed that machine learning statistics values such as mean, median, mode standard deviation and variance for collected data. Standard deviation is used to measure the accuracy. It observed “Excelent” and “Poor”are having higher deviation due to lesser number of reviews. The classier performance was verified and more than 73% of the tweets sentiment were guessed correctly by the classifier.

## V. CONCLUSION

The proposed model classifies the twitter feedback statement into five different categories like Excellent, Good, moderate, Satisfactory and poor shows the various percentage of reviews which was drawn using a horizontal bar plot. The developed Python code executed well for smaller or larger data from twitter bank. The model analyses and interprets the student or alumni feedback and classifies them successfully. It also finds the various aspects such as Learning & Resources, Research, Professional Practice and Collaborative Performance, Graduation Outcomes, Outreach and Perception online feedback given by them for a institutions. Also, the model visualizes every aspect's opinion score in terms of the various bar plots by clicking the particular aspect.

## REFERENCES

1. Yogesh K. Dwivedi, ElviraIsmagilova D., LaurieHughes, RaffaeleFilieri, JennaJacobson, VarshaJain, HeikkiKarjaluoto, HajerKef, iAnjala S.Krishen, VikramKumar. lMohammad M.Rahman, Ramakrishnan Raman, lPhilipp A.Rauschnabel, Jennifer Rowley, JariSalo, , Yichuan Wang "Setting the future of digital and social media marketing research: Perspectives and research propositions", International Journal of Information Management, Volume 59, August 2021, https://doi.org/10.1016/j.ijinfomgt.2020.102168
2. Haiping Qiao, A brief introduction to institutional review boards in the United State, Pediatr Investig. 2018 Mar; 2(1): 46–51.
3. Tirath Prasad Sahu, Sanjeev Ahuja, "Sentiment analysis of movie reviews: A study on feature selection & classification algorithms", 2016 International Conference on Microelectronics, Computing and Communications. https://doi.org/10.1109/microcom.2016.7522583
4. Mtetwa, Nhamo; Awukam, Awukam Ojang; Yousefi, Mehdi, Feature extraction and classification of movie reviews, 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI) https://doi.org/10.1109/iscmi.2018.8703235
5. Ankita Sharmaa, Udayan Ghoseb, S entimental Analysis of Twitter Data with respect to General Elections in India, nternational Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020, Procedia Computer Science 173 (2020) 325–334 https://doi.org/10.1016/j.procs.2020.06.038
6. Rezapour, Rezvaneh et al., "Identifying the Overlap between Election Result and Candidates' Ranking Based on Hashtag-Enhanced Lexicon-Based Sentiment Analysis", Semantic Computing (ICSC) 2017 IEEE 11th International Conference, 2017. https://doi.org/10.1109/icsc.2017.92
7. Bouazizi, Mondher and Tomoaki Ohtsuki, "Sentiment analysis in twitter: From classification to quantification of sentiments within tweets", Global Communications Conference (GLOBECOM) 2016 IEEE, 2016. https://doi.org/10.1109/glocom.2016.7842262
8. B. Bansal and S. Srivastava, "On predicting elections with hybrid topic based sentiment analysis of tweets", Procedia Comput. Sci, vol. 135, pp. 346-353, 2018. https://doi.org/10.1016/j.procs.2018.08.183

9. Pultar, E., Raubal, M., Cova, T. J., & Goodchild, M. F. Dynamic GIS case studies: Wildfire evacuation and volunteered geographic information. Transactions in GIS, 13(SUPPL. 1), (2009). 85–104. https://doi.org/10.1111/j.1467-9671.2009.01157.x
10. Anas Abdelrazeq, Daniela Janssen, Sabina Jeschke, Anja Simone Richert, Sentiment Analysis of Social Media for Evaluating Universities, Conference: 2nd International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC 2015), Dubai, UAE, 16-18 December 2015 https://doi.org/10.1109/dipdmwc37531.2016
11. Claudia Brauer and Edward WN Bernroider. So-cial media analytics with facebook-the case ofhigher education institutions. In HCI in Business,pages 3–12. Springer, 2015.
12. Diana Maynard and Adam Funk. Automatic de-tection of political opinions in tweets. In The se-mantic web: ESWC 2011 workshops, pages 88–99.Springer, 2012 https://doi.org/10.1007/978-3-642-25953-1_8
13. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, OwenRambow, and Rebecca Passonneau. Sentimentanalysis of twitter data. In Proceedings of theWorkshop on Languages in Social Media, pages30–38. Association for Computational Linguistics,2011
14. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86 (2002) https://doi.org/10.3115/1118693.1118704
15. Abbasi, A., France, S., Zhang, Z., Chen, H.: Selecting Attributes for Sentiment Classification Using Feature Relation Networks. IEEE Transactions on Knowledge and Data Engineering 23, 447–462 (2011) https://doi.org/10.1109/tkde.2010.110
16. Nguyen Thi Phuong Giang, Tran Thanh Dien & Tran Thi Minh Khoa, Sentiment Analysis for University Students' Feedback 2020: Advances in Information and Communication pp 55–66 https://doi.org/10.34071/jmp.2020.3.14
17. Mehmet Korkmaz, Sentiment analysis on university satisfaction in social media,Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2018. https://doi.org/10.1109/ebbt.2018.8391463
18. Duwairi, R. M., & Qarqaz, I. (2014). Arabic sentiment analysis using supervised classification. Proceedings – 2014 International Conference on Future Internet of Things and Cloud, FiCloud 2014, 579–583. https://doi.org/10.1109/ficloud.2014.100
19. Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2014). Learning sentiment from students' feedback for real-time interventions in classrooms. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8779 LNAI, 40–49. https://doi.org/10.1007/978-3-319-11298-5_5
20. Gottipati, S., Shankararaman, V., & Lin, J. R. (2018). Text analytics approach to extract course improvement suggestions from students' feedback. Research and Practice in Technology Enhanced Learning, 13(1) https://doi.org/10.1186/s41039-018-0073-0
21. Guleria, P., Sharma, A., & Sood, M. (2015). Analysis and Association Rule Mining. International Journal of Data Mining & Knowledge Management Process, 5(6), 35–44. https://doi.org/10.5121/ijdkp.2015.5603

22. Hashim, S. A., Hamoud, K. A., & Awadh, A. W. (2018). Analysing Students' Answers Using Association Rule Mining Based On Feature Selection.Journal Of Southwest Jiaotong University, 53(5)
23. Abaidullah, A. M., Ahmed, N., & Ali, E. (2014). Identifying Hidden Patterns in Students' Feedback through Cluster Analysis. International Journal of Computer Theory and Engineering, 7(1), 16-20 https://doi.org/10.7763/ijcte.2015.v7.923