Grenze International Journal of Engineering and Technology, Jan Issue

Sentiment Analysis on Multilingual Tweets using XGBoost Classifiers

C. Suganthini[1], R. Baskaran[2] and Rithvik Senthil[3]
[1-3]College of Engineering Guindy, Anna University, Chennai, India
Email: chinna246@gmail.com, baaski@annauniv.edu, rithvik335@gmail.com

***Abstract*—Twitter is a free social networking microblogging service that allows registered users to send short messages or post so-called tweets. Twitter members can tweet and follow other users' tweets. These tweets are persistent, searchable, and public. Anyone can see tweets on Twitter regardless of whether they are a follower or not. The tweets posted are either good tweets or bad tweets. Sentiment analysis uses natural language processing and computational linguistics techniques to automate the classification of sentiment generated from tweets. In this work, multilingual tweets are collected and these tweets are divided into positive and negative tweets. First, preprocess the tweets to remove unnecessary content from the tweet. Then extract the adjectives that make up the feature vector and then perform sentiment analysis by classifying the tweets using algorithms such as logistic regression. The polarity scores for the multilingual tweets are also calculated. The XGBoost classifier gives better results for the Twitter dataset than the logistic regression model.**

***Index Terms*— Semantic analysis, classifications, multilingual, sentiment analysis, polarity score.**

I. INTRODUCTION

Twitter is a social networking service on which users post and they interact with messages known as tweets. The growing popularity was unexpected in the increasing formation of opinion. But the opinion thus formed remains in the lack of proper strategic analysis, tweets can be opinionated and misleading. Sentiment analysis is currently the most significant field in the development of an organization and hence many machine learning technologies have been applied to automate the work. Sentiment Analysis identifies the sentiment expressed in the tweets and detects the polarity value. The social network acts as a medium for users to post their opinions and these sites are used to disseminate the information. Social media plays a significant role in the level of marketing and people's changing needs. The semantic analysis of the sentence increases the meaning and necessity of the information.

In current years, a variety of functions and methods for training classifiers have been explored, with varying results. The aim of this work is to extend the data set for sentiment analysis by adding semantics as an additional feature by including the corresponding semantic concept as an additional feature for each extracted entity. This approach could result in better accuracy of the classification model.

II. LITERATURE SURVEY

In the past, many researchers have discussed sentiment analysis of tweets using machine learning methods. This

Grenze ID: 01.GIJET.9.1.722

section discusses some of the existing work. Initially, tweets were classified based on Naive Bayesian Maximum Entropy Sentiment and Negation [1]. Each tweet received a score against which it was classified.

The sentiment analysis of tweets was analyzed using the Hadoop framework with the aim of efficiently analyzing the huge amount of existing data [2]. It was successfully shown that the Hadoop framework is immensely effective in classifying particularly large amounts of data. The machine learning methods used in semantic analysis is to classify the sentence and product ratings based on Twitter data were performed [3]. To analyze a large number of reviews using Twitter datasets that are already labeled. The Naive Bayes technique shows a better result than maximum entropy. The polarity score function and the n-gram function are the sentiment feature set of tweets that integrates into Support Vector Machines (SVM) model training and predicts a sentiment classification label [4].

The most effective combination of features for a Chinese microblogging site is identified [5]. Sentiment analysis of tweets using TextBlob, a NLP library, was performed in [6]. A unique method for detecting Twitter communities by improving the quality function of modularity and a Constant Potts model further contributed to further expand the scope of community detection in social networks [7]. Semantic analysis of interpersonal interaction correspondence based on the semantic analysis of food reviews [8]. The first feeling polarity dictionary with a translator and a dictionary consisting of about 159,876 Turkish words was created in [9]. The idea of adding semantically similar words and contextual identities to the feature vectors to increase the accuracy of the prediction was successfully identified [10] [11]. The push into multilingual classification continued as Indonesian Twitter sentiment analysis using Word2Vec [12]. An aspect-based fuzzy logic sentiment analysis on big data in social media was conducted to analyze customers' problems in purchasing quality products [13].

Multilingual representations of words with similar meanings in multilingual [14]. Language recognize common points of view from multilingual comments [15]. Hybrid sentiment classification model using ensemble machine learning methods [16]. Classification model with different text features analyze sentiment polarity [17]. Sentiment analysis is performed for tweets on different domains [18]. Various classification models for sentiment polarity [19]. Graph neural networks are used to extract the information from the tweets and make a prediction based on the extracted graph [20]. The ensemble classifier helps in determining sentiment analysis [21]. The classifier predicts the location of the tweets based on the KNN sentiment analysis [22-24]. Semantic similarity between the comments directly identifies emotional polarity and lexical polarity [25-27]. User-generated sentences contain the sentiment information using the hybrid model classification techniques [28-30]. Sentiment analysis to determine opinions using a lexicon-based algorithm [31].

III. System Design

The system consists of simplified modules and the architecture diagram shown in Figure 1. The Twitter data is pre-processed to convert data into an appropriate format. Then the transformed data is passed through a semantic analyzer to remove and lemmatize stop words. In the feature extraction model, the feature vectors are extracted from the processed tweets. Then the data’s split into training and validation data. Then the training data is passed to the training model. The testing data performs in classifier model. These help predict the output of the senti-strength score, based on which tweets are classified as positive and negative. Finally, the polarity score or degree of positivity or negativity is calculated for the tweets dataset based on the four categories such as strongly positive, weakly positive, weakly negative and strongly negative.

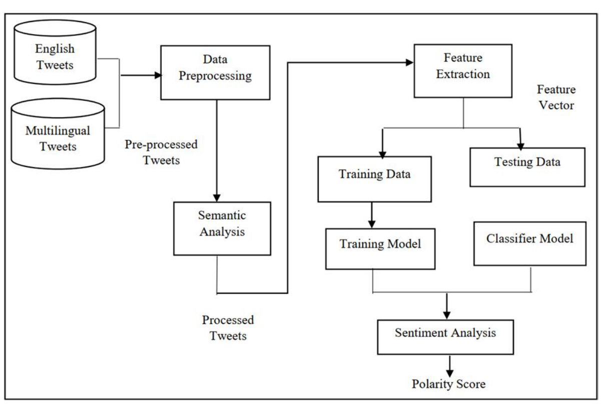

Figure 1. Architecture Diagram

A. Data Preprocessing

Data pre-processing is a technique that allows raw data to be converted into a suitable format. Data collected from different platforms is often unstructured, incomplete, contradictory or lacking specific characteristics and is likely to contain ambiguities. Data pre-processing makes it easier to solve these problems. It prepares pre-processed data for further processing.

B. Semantic Analysis

In this phase, the data is cleaned in terms of the following aspects: stop word removal, word stemming, punctuation, lowercase data, and repeated word/character removal. Remove stop words and lemmatize. The process of lemmatization is to translate a word into its base form. This filtering is more valuable, meaning only those with actual and relevant content are retained. This process plays the main role in tokenizing words. Finally, after removing those useless words, the remaining data is tokenized.

C. Feature Extraction

The main purpose of feature extraction is to reduce model complexity, improve model accuracy and train the machine learning algorithm faster. In sentiment classification, the features are most often the terms or phrases that influence the sentiment of the tweets. The post-processing of the filtered dataset has many distinctive properties. The feature extraction process extracts the part of speech from the data set. TFID feature vectorization is used. It converts text into feature vectors that can be utilized as input for estimators. The vocabulary is a dictionary that converts each token into a feature index in matrix form.

D. Training / Testing Model

The selected functions are divided into training set and test set. In the training set, 80 percent of the actual data set is used as the trained model. In the test set, the 20 percent data sample is a trained model that is evaluated as a final fit.

E. Classification Metrics

The classifier used is a logistic regression classifier. The purpose of classification is to find out whether the tweet is positive or negative. The logistic regression classifier used to identify four classification labels as strongly positive, weakly positive, weakly negative and strongly negative.

F. Sentiment Analysis

The tweets are separated after classification as positive or negative tweets. A sample data set of 15,000 tweets is tested without a sentiment label (0 or 1) and it is found that the model correctly predicts the output and stores it in the output file as shown in Table I.

TABLE I. SAMPLE TWEETS

S. No.	Tweet	Sentiment
1	@user when a father is dysfunctional and is a…	NEGATIVE
2	@user @user thanks for #lyft credit. I can’t us…	NEGATIVE
3	bihday your majesty	POSITIVE
4	#model i love you take with u all the time in…	POSITIVE
5	Factguide: society now #motivation	POSITIVE

IV. RESULTS AND DISCUSSION

A. Datasets

The dataset contains 160K tweets extracted using the Twitter API, an open-source dataset. The tweets were commented (0 = negative, 1 = positive) and used for sentiment detection. It contains the following 6 fields shown in Figure 2.

	target	id	time	query	user	tweet
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....

Figure 2. Sample Dataset

Data distribution is used to check whether the data set is balanced or not. The conclusion of the data distribution shows that the data set is equally balanced, with the number of positive tweets and negative tweets being equal. The distribution data is presented visually in Figure 3.

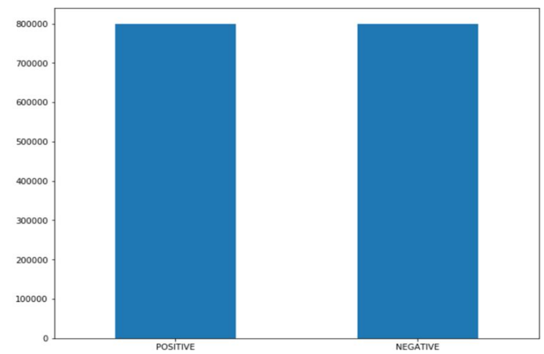

Figure 3. Data Distribution

B. Senti Strength

Sentiment strength estimates the strength of positive and negative sentiments in texts, including for informal language. Polarity scores are used to find Senti strength. Polarity values range from -1 (strongly negative) to +1 (strongly positive). Based on the polarity of each tweet, the Senti strength classifications are shown in Table II. The calculation of polarity scores or values for the processed tweets, is shown in Table III.

TABLE II. CLASSIFICATION OF SENTI-STRENGTH

Senti-Strength	Polarity
Strongly Positive	Polarity ≥ 0.6 to 1
Weakly Positive	Polarity ≥ 0.1 to 0.5
Weakly Negative	Polarity ≤ - 0.1 to -0.5
Strongly Negative	Polarity ≤ - 0.6 to -1

TABLE III. POLARITY SCORES

S. No.	Tweet	Senti strength	Polarity
1	@user when a father is dysfunctional and is a…	Weakly Negative	- 0.500000
2	@user @user thanks for #lyft credit. I can't us…	Weakly Positive	0.200000
3	bihday your majesty	Strongly Negative	- 0.70000
4	#model i love you take with u all the time in…	Strongly Positive	0.978882
5	Factguide: society now #motivation	weakly Positive	0.40000

C. Multilingual Testing

This also applies to a new approach to finding the sentiment analysis of foreign languages. Some of the languages used are Tamil, French, Spanish, Italian, Arabic, Chinese, Hindi and many more. Some of the examples shown below.

- ARABIC:
  الجميع من فضلك لا تطير مع. كان لديه تجارب قليلة رهيبة. لديهم
  خدمة سيئة للغاية ورحلاتهم الجوية تتأخر دائما فدب سفر رجاء الجميع لا يطير مع الخطوط الجوية. كان لديه تجارب قليلة رهيبة.
  لديهم خدمة سيئة للغاية ورحلاتهم الجوية تتأخر دائما
- THAI:
  เจนไน ทำได้ดีมากในวนนี้โ แซม ดยเ เล่นเตะที่ยอดเยยมเพื่อเอาชนะทีม เอ้ย ทุกคนโปรดอย่าบินกั สายการบิน มีประสบการณ์ที่แย่มาก พวกเขามีบรการที่แย่มากและเที่ยวบินมักจะล่าช้า- เสมอ โอ้พระเจ้า
- FRENCH:
  Csk se débrouillait très bien aujourd’hui, surtout que Samcurran a joué un excellent coup de pied pour gagner l’équipe. Oh mon Dieu Tout le monde, s’il vous plaît, ne volez pas avec @airways. J’ai eu quelques terribles expériences. Ils ont un service très médiocre et leurs vols sont tou- jours retardés fedup travelwoes
- ENGLISH:
  Today’s Performance by @csk was great especially @samcurran played an amazing knock to get the win for the team. winningstreak. Everyone please dont fly with @airways. Had a terrible few experiences. They have very poor service and their flights are always delayed.

By using a multilingual Tweets dataset for this module, which is an open source dataset. The foreign language is changed to English and then a semantic and classifying task is performed and finally the sentiment and Senti strength of the tweets are predicted as shown in Table IV.

TABLE IV. MULTILINGUAL TESTING SAMPLE DATA

S. No.	Tweet	Sentiment
1	Bugun bulusmami lazimdii	POSITIVE
2	Volkan konak adami tribe sokar yemin ederim :D	POSITIVE
3	Bed	POSITIVE
4	I felt my first flash of violence at some fool…	NEGATIVE
5	Ladies drink and get in free till 10:30	POSITIVE

D. Performance Metrics

In this section, a comparative analysis using the XGBoost classifier and the logistic regression model on iterated random projections is performed to find a suitable machine learning algorithm for the classification component of the framework. The results obtained from the XGBoost classifier and the logistic regression are discussed and the comparison with relative performances based on four parameters: accuracy, precision, recall and F1 score are chosen as classification metrics to compare the performance results of different models in relation to four classes such as strongly positive, weakly positive, weakly negative and strongly negative are shown in Table V.

- Precision:

Precision is defined as the proportion of correctly predicted positive observations to the total of predicted positive observations, as in (1).

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

- Recall:

Recall is defined as the proportion of correctly identified positive observations, as in (2).

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

- F1 Score:

F1 score is the harmonic mean (HM) of precision and recall, as in (3).

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

TABLE V. COMPARISON OF THE XGBOOST CLASSIFIER AND LOGISITIC REGRESSION MODEL

Metrics	XGBoost Classifier	Logistic Regression
Precision	0.81	0.78
Recall	0.84	0.80
F1 Score	0.83	0.79
Accuracy	0.80	0.76

The performance of the model comparison of the XGBoost classifier and the logistic regression is shown in Figure 4.

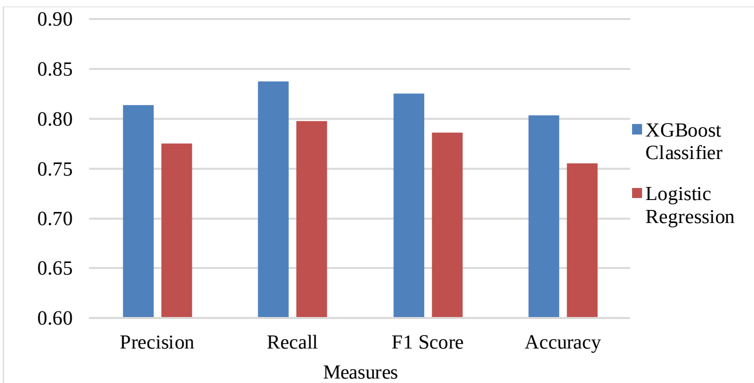

Figure 4. Performance Measure

The logistic regression classifier validated to be the best performing classifier as it had a good accuracy of 81%. Since this is a balanced dataset, the focus is more on accuracy. A confusion matrix is shown below in Figure 5 to show TN, TP, FP, FN.

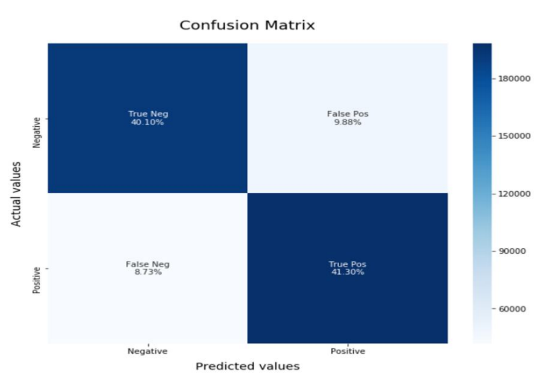

Figure 5. Confusion Matrix

V. CONCLUSION

In this paper, the proposed method uses a set of machine learning methods with semantic analysis. Based on the Tweet Sentiment Score, the words from each sentence are rated, which can be utilized to determine positive or negative intention behind the tweet. Then the tweets were categorized into four different classes, based on the sentiment identified in the tweets such as strongly positive, weakly positive, weakly negative and strongly negative, based on the sentiment identified in the tweets. The sentiment strength is also defined for the tweets using polarity scores. Finally, foreign languages were included for sentiment analysis, showing good results. The experimental result of this study illustrated that the XGBoost classifier algorithm performed well compared to the logistic regression algorithm in terms of the applied evaluation metrics. Future work includes comparing the performance of other more powerful models such as the Inception models for sentiment classification and using different semantic analysis techniques to process the data.

REFERENCES

[1] S. Kale and V. Padmadas, "Sentiment analysis of tweets using semantic analysis," *in 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), IEEE*, pp. 1-3, 2017.

[2] M. Acharya and S. Sharma, "Semantic Analysis of Twitter Posts," *in 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), IEEE*, pp. 104-109, 2018.

[3] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," *in 2014 Seventh international conference on contemporary computing (IC3), IEEE*, pp. 437-442, 2014.

[4] Z. Jianqiang and C. Xueliang, "Combining semantic and prior polarity for boosting twitter sentiment analysis," *in 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), IEEE,* pp. 832-837, 2015.

[5] H. Sui, Y. Jianping, Z. Hongxian and Z. Wei, "Sentiment analysis of Chinese micro-blog using semantic sentiment space model," *in Proceedings of 2012 2nd International Conference on Computer Science and Network Technology, IEEE*, pp. 1443-1447, 2012.

[6] H. M. Alash and G. A. Al-Sultany, "Enhanced Twitter Community Detection using Node Content and Attributes," *in 2021 1st Babylon International Conference on Information Technology and Science (BICITS), IEEE*, pp. 5-10, 2021.

[7] B. Sophia, R. Rajaraman, S. Mathan and G. R. Shankar, "Semantic approach to live data mining using clustering technique," *in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE*, vol. 1, pp. 1577-1581, 2021.

[8] N. Islam, N. Akter and A. Sattar, "Sentiment analysis on food review using machine learning approach," *in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE*, pp. 157-164, 2021.

[9] A. B. Altınel, K. Buzlu, K. Ipek, "Performance Analysis of Different Sentiment Polarity Dictionaries on Turkish Sentiment Detection," *in 2022 International Conference on Innovations in Intelligent Systems and Applications (INISTA), IEEE*, pp. 1-6, 2022.

[10] M. Kanakaraj and R. M. R. Guddeti, "NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers," *in 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), IEEE*, pp. 1-5, 2015.

[11] M. Kanakaraj and R. M. R. Guddeti, "Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques," *in 2015 IEEE 9th International Conference on Semantic Computing (ICSC), IEEE*, pp. 1-5, 2015.

[12] F. W. Kurniawan and W. Maharani, "Indonesian Twitter Sentiment Analysis Using Word2Vec," *in 2020 International Conference on Data Science and its Applications (ICoDSA)", IEEE*, pp. 1-6, 2020.

[13] S. U. Maheswari and S. S. Dhenakaran, "Aspect based Fuzzy Logic Sentiment Analysis on Social Media Big Data," *in International Conference on Communication and Signal Processing, IEEE*, pp. 157-164, 2021.pp. 0971-0975, 2020.

[14] M. Ahmed, C. Dixit, R. E. Mercer, A. Khan and M. R. Samee et al., "Multilingual Semantic Textual Similarity using Multilingual Word Representations," *2020 IEEE 14th International Conference on Semantic Computing (ICSC),* pp. 194-198, 2020.

[15] Bei Shi, Lam Wai, Lidong Bing and Yinqing Xu, "Multilingual Viewpoint Detection from news comments," *2015 International Conference on Asian Language Processing (IALP)*, pp. 189-192, 2015.

[16] M. Rathi, A. Malik, D. Varshney, R. Sharma and S. Mendiratta, "Sentiment Analysis of Tweets Using Machine Learning Approach," *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pp. 1-3, 2018.

[17] S. Naz, A. Sharan and N. Malik, "Sentiment Classification on Twitter Data Using Support Vector Machine," *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 676-679, 2018.

[18] S. Anil Phand and V. A. Chakkarwar, "Enhanced Sentiment Classification Using Geo Location Tweets," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 881-886, 2018.

[19] T. Jain, S. Mathur, A. Ninnad, B. Nikshep and N. Chalil, "Analyzing of Political Tweets in Hindi Language Using Machine Learning and Deep Learning," *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, pp. 1-5, 2022.

[20] A. Nguyen, A. Longa, M. Luca, J. Kaul and G. Lopez, "Emotion Analysis Using Multilayered Networks for Graphical Representation of Tweets," *in IEEE Access*, vol. 10, pp. 99467-99478, 2022.

[21] J. N. V. R. Swarup Kumar, J. D. Sri, K. M. S. Manoj, D. Srividya and D. Prathyusha et al., "Sentiment Analysis on Textual Tweets using Ensemble Classifier (LSTM-GRU)," *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pp. 926-932, 2022.

[22] A. Mostafa, W. Gad, T. Abdelkader and N. Badr, "Predicting the Tweet Location Based on KNN-Sentimental Analysis," *2020 15th International Conference on Computer Engineering and Systems (ICCES)*, pp. 1-6, 2020.

[23] F. Akbarian and F. Z. Boroujeni, "An Improved Feature Selection Method for Sentiments Analysis in Social Networks," *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 181-186, 2020.

[24] C. J. Varshney, A. Sharma and D. P. Yadav, "Sentiment Analysis using Ensemble Classification Technique," *2020 IEEE Students Conference on Engineering & Systems (SCES)*, pp. 1-6, 2020.

[25] S. H. Dhahi and J. Waleed, "Emotions Polarity of Tweets Based on Semantic Similarity and User Behavior Features," *2020 1st Information Technology to Enhance e-learning and Other Application (IT-ELA), IEEE*, pp. 1-6, 2020.

[26] D. Gowsikhaa, S. Abirami and R. Baskaran, "Automated human behavior analysis from surveillance videos: a survey," *Artificial Intelligence Review*, vol. 42, no. 4, pp. 747–765, 2014.

[27] C. Suganthini and R. Baskaran, "Identifying the Influential User Based on User Interaction Model for Twitter Data," *Advances in Signal Processing and Intelligent Recognition Systems, SIRS 2019, Communications in Computer and Information Science, Springer*, vol. 1209, pp. 48–63, 2020.

[28] M. H. Abd El-Jawad, R. Hodhod and Y. M. K. Omar, "Sentiment Analysis of Social Media Networks Using Machine Learning," *2018 14th International Computer Engineering Conference (ICENCO)*, pp. 174-176, 2018.

[29] C. Suganthini, and R. Sridhar, "A Survey on Community Detection in Social network analysis," *International Journal of Applied Engineering Research*, vol. 10, no.75, pp. 1–6, 2015.

[30] R. Sridhar, A. Kumar, S. B. Roshini, R. K. Sundaresan and S. Chinnasamy, "Feature based Community Detection by extracting Facebook profile details," *ICTACT Journal on Soft Computing*, vol. 8, No. 4, pp. 1706–1713, 2018.

[31] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," *2019 International Conference on Computer and Information Sciences (ICCIS)*, pp. 1-4, 2019.