SENTIMENT ANALYSIS OF INDONESIA'S DIGITAL WALLET USING COMBINATION MACHINE LEARNING AND EMOTICON WEIGHT

[1]Gusmariani Tinambunan, [2]Suharjito
[1,2]Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, [1,2]Jakarta, Indonesia 11480
[1]gusmariani.tinambunan@binus.ac.id, [2]Suharjito@binus.edu

Telah dipresentasikan pada Seminar Nasional Ilmu Komputer dan Sistem Informasi (SNIKSI)

Abstract

Opinions on social media can be used to determine user sentiment by using sentiment analysis concept. Sentiment analysis requires several important stages, namely, preprocessing, feature extraction and classification method stages. The preprocessing stage was carried out to eliminate inconsistent data. In previous research, punctuation marks removal was applied at the preprocessing stage which can eliminate the emoticon position. Emoticons are a combination of punctuation marks. According to previous research, the emoticon feature has no contribution in sentiment analysis. There is another suggestion to maintain an emoticon position like converting an emoticon into a more relevant word such as :) into a "smile". However, the feature of emoticon weights has not been considered in the sentiment analysis process. In order to consider the role of emoticons and to improve sentiment analysis performance, we propose using a combination of machine learning and emoticon weights. We perform emoticon weight based on probability and sentiment score. Each probability value and sentiment score of the emoticon will be normalized using the z-score method. There are several machine learning methods that have the best classification success rates, namely, Naïve Bayes and SVM. Based on the evaluation results of the proposed model, the best accuracy is 87% - 89% when using the combination of machine learning and emoticon sentiment score. Based on the results also show that the emoticon sentiment score has a significant effect on the accuracy of sentiment analysis.

***Keywords**: Emoticon Weight, Naïve Bayes, Preprocessing, Sentiment Analysis, SVM*

INTRODUCTION

The number of digital industries in Indonesia has increased significantly, especially in the financial sector. Digital wallets are an example of technology from the financial sector that can facilitate the process of non-cash transactions online [1]. It offers easy accessibility and can be used from mobile devices or smartphones. According to [2], there are more than 40 electronic money operators in Indonesia. According to [3], there are three popular digital wallets in Indonesia, which are GOPAY, OVO and DANA. The order of the popular applications is based on the number of active users. Each digital wallet service provider will use various media such as social media for marketing, interaction, socialization and expressing their opinions in textual and non-textual forms such as emoticons. All opinions can be collected and used to determine a user's sentiment using sentiment analysis technology.

Sentiment analysis is an application of Natural Language Processing (NLP) that focuses on identifying the author's expression of an object [4]. Sentiment analysis can also be called opinion mining, which is a field of study that analyzes people's opinions, sentiments, evaluations, assessments, attitudes, and emotions for entities such as products, services, organizations, individuals, problems, events, topics, and other attributes [5].

There are several important processes in sentiment analysis such as preprocessing, feature extraction, and classification methods. The first important stage is preprocessing that focuses on eliminating inconsistent data. This stage aims to produce the best accuracy in classification. There are several processes on the preprocessing stage such as removal of some meaningless parts, case folding, tokenization, stopwords removal, and stemming based on [6, 7, 8].

Generally, punctuation marks will be removed which can eliminate the position of emoticons. Punctuation and emoticons are considered meaningless and have no contribution to sentiment analysis. According to [9], the emoticon has meaning to express emotions in sentences. There is a proposed model for maintaining the role of emoticons, namely convert emoticon carried out by [10, 11, 12]. Convert emoticon is a method of turning emoticons into more relevant words and have sentimental meanings such as :) become "smile" word.

Based on the methods applied in previous studies, the feature of emoticon weights has not been considered. Thus, this paper proposes using emoticon weights to improve sentiment analysis performance,

especially against the opinions of users of Indonesia's digital wallet services. Two methods of emoticons weights will be implemented, namely probability and sentiment score. The probability emoticon is calculated based on the number of emoticon occurrences in data. While the sentiment score of emoticons is obtained based on a calculation of the probability emoticon in the positive and negative classes. Once both values are obtained, the next stage is normalization with a z-score method. A z-score method is implemented to support improved accuracy in classification.

Another important stage in sentiment analysis is the extraction feature. Frequency-Inversion Frequency Term (TFIDF) is a method of extraction feature that can provide value for each word in a document [13]. Motivation to use Naïve Bayes and SVM methods because it has the best performance and fast in classifying sentiment. As can be seen from previous research conducted by [7], the accuracy obtained is 96.46% for Naïve Bayes and 94.16% for SVM. The main contribution in this research, the classification process will propose by combining machine learning results and weight emoticons. The emoticon feature also plays more function to support text sentiment. This combination process aims to produce an efficient and performance-enhancing prediction class. Z-score was also implemented to normalize emoticon weights.

RESEARCH METHODS

In this paper, we use sentiment analysis with some improvement in preprocessing and classification process as indicated in Figure 1. This study methodology, there are four phases to perform the sentiment analysis with data collection, preprocessing and labelling phase, classification phase and testing phase.

1. Data Collection

   Based on Fig. 1, the first stage to performance sentiment analysis is data collection. The source of the data is obtained by web scraping from Twitter. The scraping process is based on the keyword Twitter object of Indonesia's digital wallet. Once the data is successfully collected, the data is then reselected specifically containing emoticons only. The purpose of data selection is to see the effect of emoticon roles clearly.

2. Preprocessing and Labelling

   a. Preprocessing

According to [14], preprocessing is usually applied at the time of collecting text before carrying out further text mining processes. Preprocessing is important to change unstructured data by reducing the number of dirty data. In this paper, the process of convert emoticons remained implemented. The function of convert emoticons is to represent the presence of emoticons and supports the process of calculating emoticon weights.

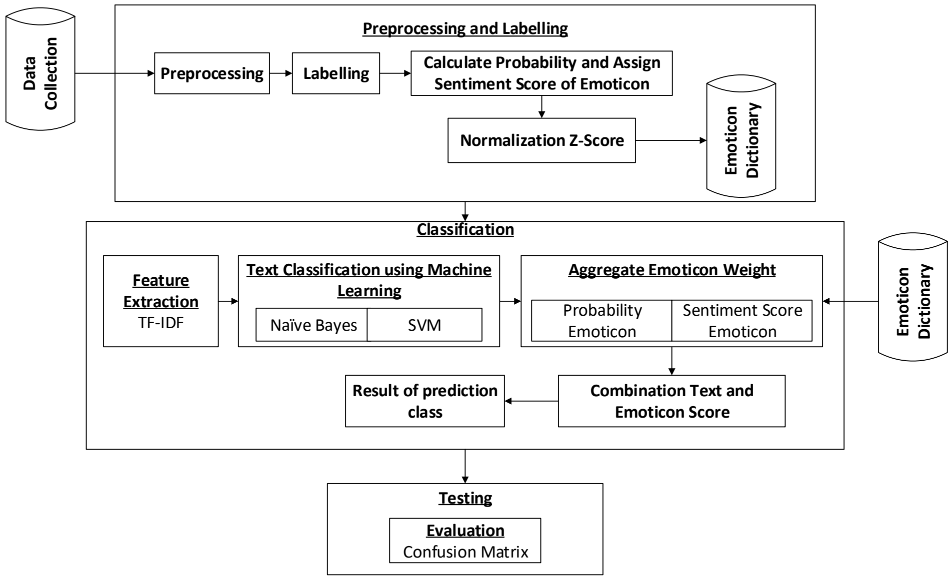

Figure 1. Research methodology of sentiment analysis using combination classification method

Table 1. Sample of Emoticon List

Character	Emoticon Name	Label
:-s	emoticonAnnoyed	negative
xo	emoticonShock	negative
:((	emoticonSad	positive
:))	emoticonSmile	positive
:D	emoticonLaugh	positive

Table 1 shows a list of emoticons that were used according to [15], emoticons collected are Unicode characters that are often used on Twitter. The lexicon was used as the initial label of emoticons used in this study. Other emoticons are used according to [16], where emoticons are derived from ASCII characters. The labels in this study are defined as -5 to 5. A minus value indicates a negative label and another label indicates a positive label. The following is a list of preprocessing processes used, namely cleansing, case folding, tokenization, filter slang words, stopword removal, filter token, stemming, separating emoticon and text. This paper used stopword removal and stemming dictionaries based on research conducted by [17, 18].

b. Labelling

After the data is cleaned, then labeling is the next process as the initial classification. Determination of sentiment is using a dictionary-based approach based on research from [19]. This approach works by calculating the score of each word. In this paper, the dictionary used to determine the class of the word is SentiStrength for Bahasa Indonesia and emoticon dictionary according to [15, 16].

c. Calculate probability and assign sentiment score of emoticon with normalization

Calculate score for emoticon is the next step after get data label. In this paper, we proposed considering the number of occurrences of emoticon in the data. The calculation were used according to [15].

- Calculate probability emoticon

Emoticon probability is calculated based on the number of emoticons occurrences in each positive and negative class ($N(c)$). Equation 1 defined how to calculate the probability of an emoticon.

- Calculate sentiment score of emoticon

Sentiment scores ($Semo$)are obtained by subtracting positive probability ($Pemo(pos)$) and negative probability ($Pemo(neg)$). Equation 2 is defined how to calculate the sentiment score of emoticons.

$$Pemo(c) = \frac{N(c) + 1}{N + k} \tag{1}$$

$$Semo = Pemo(pos) - Pemo(neg) \tag{2}$$

$$z = \frac{score \;\; - mean}{standard\ deviation} \tag{3}$$

$$P(c_j|w_i) = \; P(c_j)xP(w_i|c_j) \tag{4}$$

$$P(c_j) = \frac{Nc_j}{N} \tag{5}$$

d. Normalization z-score

The next stage is normalization for both values, namely probability and sentiment score emoticons with z-score method. According to [20], z-score normalization is a method to change the value of an attribute in the range 0 to 1. The normalization process is carried out based on mean and standard deviations. Equation 3 is defined how to perform normalization z-score.

3. Classification
   a. Text Classification using Machine Learning

   Machine learning has good performance and effectiveness in classifying data, namely Naive Bayes and SVM. The purpose of using both methods is to determine which performs better with accuracy parameters. To support the classification process, TF-IDF was used to calculate the weight of each word. Classification process requires data training and testing, then cross validation method used with k-fold is 10.

   - Naïve Bayes

The classification process of the Naïve Bayes algorithm is carried out with Equation (4). This equation is to determine the probability of data testing in each class. Where the value of $P(c_j)$ is a category opportunity from data training. While the value of $P(w_i|c_j)$ is the probability of a word in the positive or negative category.

The following is Equation (5) to obtain the probability of category. The process is to divide the number of words in a category (Nc_j) by the number of all words used in the training data (N).

The following as Equation (6) to obtain the probability of a word in each category. Process stages by summing the occurrence of test words from each category ($count(w_i, c_j) + 1$). Then the result is divided by the sum of the entire occurrence of the word from each category $\sum_{w \epsilon V} count(w, c_j))$ added the total number of words ($|V|$).

$$P(w_i|c_j) = \frac{count(w_i, c_j) + 1}{(\sum_{w \epsilon V} count(w, c_j)) + |V|} \tag{6}$$

$$EWeight(c) = \sum \text{ProbabilitasEmoticon}(c) \tag{7}$$

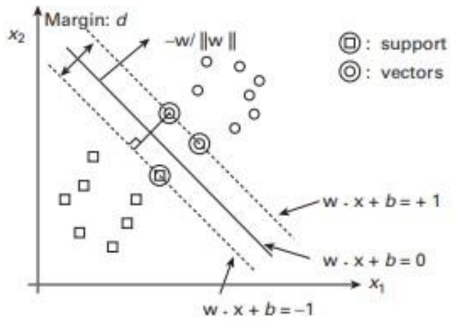

Figure 2. SVM model to find the suitable hyperplane

- SVM

  SVM is classification technique works with the hyperplane by finding the largest margin. Margin is the distance between the closest data of each class to a hyperplane. Figure 2 represent how the suitable hyperplane.

b. Combination Classification of Text and Emoticon

   To determine the role of emoticons and improve classification performance, the weight of the emoticon will be combined with the results of the text classification. The emoticon weight will be calculated based on the probability and sentiment score in each tweet.

- Calculate emoticon weight based on polarity

  Calculate the weight of an emoticon by adding the emoticon probability value of each icon that appears on a tweet. Equation (7) is defined how to calculate the emoticon weight based on polarity.

- Calculate emoticon weight based on sentiment score

  Calculate the weight of an emoticon by adding each sentiment score of the emoticon from each icon that appears on a tweet. The calculation process is based on each class. Especially weighting with sentiment score, will also be used sigmoid function method. In this paper, we use the sigmoid function to change the amount of data with a range of 0 to 1. The sigmoid function also plays a role in supporting the process of combining values from machine learning results that derive grades from both classes. Equation (8) and Equation (9) are defined how to calculate the emoticon weight based on sentiment score.

$$EWeight(positif) = sigmoid(\sum EmoticonScore) \quad (8)$$

$$EWeight(negatif) = 1 - sigmoid(\sum EmoticonScore) \quad (9)$$

$$CombinationScore(c) = \alpha \; x \; TextScore(c) + (1 - \alpha) \; x \; EWeight(c) \quad (10)$$

- Combination text and emoticon
  After the weight of emoticons in each class is obtained, then continued with the process of a combination of sentiment score of text and emoticons. The α value used in this paper was 0.5. Equation (10) is defined how to calculate combination text and emoticon for classification.
- Determination prediction class
  After the result of the combination of scores obtained from Equation (10), the determination of classification with the following conditions:
  • Data is predicted as a positive class, if the combination of positive scores is greater than or equal to the combination of negative scores.
  • Data is predicted as a negative class, if the combination of positive score is less than negative score combination.

4. Experiment Design

Sentiment classification performance results will be evaluated using the confusion matrix method. The experimental design that will be carried out in this paper from the proposed model are as follows:

- Evaluates the classification results of a combination of machine learning results and emoticon probability.
- Evaluates the classification results of a combination of machine learning results and emoticon sentiment scores.

The parameters used to evaluate classification performance are accuracy, precision, and recall. The classification results will be mapped into a confusion matrix table.

RESULTS AND DISCUSSIONS

The following are the design and analysis results of the models used to achieve the goal of improving classification performance.

1. Data Collection
   - Scraping and Preprocessing Result
     The data source was collected from Twitter with Indonesian digital wallet objects namely GOPAY, OVO, and DANA. Table 2 represent total data collection.

Table 2. Total Data Collection

Objects of digital wallet in Indonesia	Total Overall Data	Total Data Containing Emoticon
@gopayindonesia	31.570	4.793

@ovo_id	90.586	9.025
@datawallet	38.575	8.457

Table 3. Total Data with Labelling Using Dictionary-Based Approach

Objects of digital wallet in Indonesia	Total Positive	Total Negative
@gopayindonesia	3.502	1.291
@ovo_id	6.198	2.827
@datawallet	7.036	1.421

Data retrieval using scrape techniques by utilizing snscrape library applications in python. The data retrieval period is January 2020 to December 2020. Table 2 shows that the overall scraping of data from GOPAY account is 31.570 records, OVO account is 90.587 records and DANA account is 38.575 records. Once the data is collected, the next process is preprocessing. The collected data will be reselected only containing emoticons and text. Based on the result, total data containing emoticon and from GOPAY account has 4.793 records, OVO account has 9.025 records and DANA account has 8.457 records

- Labelling Result

All data has been obtained and cleaned, then continued with labeling using a dictionary-based approach. Table 3 shows the number of data that has positive and negative labels from each Twitter account.

The dictionary-based approach performs by calculating the score of positive words and negative words contained in sentences. Data are labeled as positive sentiment if the score is more than 0. Data are labeled as negative sentiment if the score is less than 0. Table 3 shows that GOPAY account has 3.502 records for positive tweet and 1.291 records for negative tweet. OVO account has 6.198 records for positive tweet and 2.827 records for negative tweet. DANA account has 7.036 records for positive tweet and 1.421 records for negative tweet.

2. Performance Result

In order to support the classification process with a machine learning approach, 10-fold cross-validation was used to divide training and testing data. Table 4 shows that the proposed model produces the best accuracy.

Table 4. Performance Evaluation Result Using Confusion Matrix

Dataset	Machine Learning	Emoticon Treatment	Accuracy	Precision	Recall
GOPAY	Naïve	Sentiment	89,82%	91,71%	94,43%

	Bayes	Score			
		Probability	88,21%	94,16%	89,20%
	SVM	Sentiment Score	89,09%	92,75%	92,07%
		Probability	88,19%	94,39%	88,92%
OVO	Naïve Bayes	Sentiment Score	90,99%	93,33%	93,52%
		Probability	89,30%	94,27%	89,80%
	SVM	Sentiment Score	90,23%	94,64%	90,71%
		Probability	89,24%	94,30%	89,66%
DANA	Naïve Bayes	Sentiment Score	93,07%	93,83%	97,91%
		Probability	84,41%	84,41%	99,61%
	SVM	Sentiment Score	93,32%	95,46%	96,32%
		Probability	93,58%	96,53%	95,55%

Accuracy

Accuracy is the percentage of test data correctly classified by the classifier. In Table 4, GOPAY datasets show that the best accuracy when using Naïve Bayes and weighting sentiment score emoticons is 89,82%. OVO datasets show that the best accuracy when using Naïve Bayes and weighting sentiment score emoticons is 90,99%. DANA datasets show that the best accuracy when using SVM and weighting emoticon with probability is 93,58%. A combination score using emoticon sentiment score obtained the best accuracy in two datasets. This shows that the sentiment score of emoticons has the best effect in increasing accuracy.

Precision

Precision is a parameter that measures the proper proximity of a sentence In Table 4, GOPAY datasets show that the best precision when using SVM and weighting emoticon with probability is 94,39%. OVO datasets show that the best precision when using SVM and weighting sentiment score emoticons is 94,64%. DANA datasets show that the best precision when using SVM and weighting sentiment score emoticons is 95,46%. Based on the result shows that the sentiment score of emoticons has the best effect in increasing precision.

Recall

Recall is the correct number of sentences predicted in a document. In Table 4, GOPAY datasets show that the best recall when using Naïve Bayes and weighting sentiment score emoticons is 94,43%. OVO datasets show that the best recall when using Naïve Bayes and weighting sentiment score emoticons is 93,52%. DANA datasets show that the best recall when using Naïve Bayes and weighting emoticon with probability is 99,61%. Based on the result shows that the sentiment score of emoticons has the best effect in increasing recall.

3. Performance Result Comparasion with Previous Research

   Table 5 shows a comparison of classification performance results between the proposed model and previous research.

   Table 5 shows that research conducted by [11], with proposed models with convert emoticons and SVM produced accuracy of 85%. Research conducted by [21] with proposed of three preprocessing models and the combination of scores produced accuracy of 83%. Research conducted by [22], with the proposed polarity of emoticons and Naïve Bayes method produced accuracy of 81%. However, this study proposed a different model that is a combination of scores of probability or sentiment score emoticons that have been normalized with z-score method. The combination of two types of features, textual and emoticons, has the potential to support performance improvement. Emoticons represent faces in the text that play an important role in changing and improving sentiment. Weighting emoticons are used from the number of emoticons occurrences on tweets. The more emoticons occur, the more significant the emoticon weight will be and support sentiment determination. Based on accuracy results obtained is 87% - 79%, shows that the application of the proposed model in this study supports improved classification performance. The accuracy in this study also perform better compared to the other methods implemented in previous study.

Table 5. Comparison Result with Previous Research

Author	Emoticon Treatment	Machine Learning	Accuracy
Windasari, Uzzi, & Satoto, 2017 [11]	Convert cmoticon	SVM	85%
Wegrzyn-Wolska, Bougueroua, Yu, & Zhong, 2016 [21]	EmoDeletion, Emo2label, Emo2expla-nation	Naïve Bayes and combination score using Emoticon-Wight Lexicon Model	80% - 83%
Pal, Pawar, Zambare, & Hole, 2019 [22]	Polarity Class Emoticon	Naïve Bayes	81%
Proposed model	Emoticon probability and sentiment score	Naïve Bayes, SVM	87% - 89%

CONCLUSION AND FUTURE RESEARCH

This research was conducted to analyze the public sentiment regarding digital wallets in Indonesia. In the process of sentiment analysis, we proposed to apply a combination score of machine learning and emoticon approach and improve the preprocessing model. The improvement of the preprocessing model offered was emoticon handling where emoticons became part of the features in sentiment analysis. The purpose of applying this model was to improve classification performance. To support the hypothesis, this paper evaluates performance results with a confusion matrix. Based on the confusion matrix result, the best accuracy was obtained at 87% - 89%. Based on the result, it can be concluded that emoticons had a significant effect and able to improve performance. In the future, the classification performance can be further improved by adding more emoticon dictionaries, adding more slang words for the Indonesian language, and using the classification method.

REFERENCES

[1] A. J. Levitin, "Pandora's Digital Box: The Promise and Perils of Digital Wallets," *U. Pa. L. Rev.*, pp. 305-376, 2017.

[2] Bank Indonesia, "Payment System License Information," 14 December 2019. [Online]. Available: https://www.bi.go.id/id/sistem-pembayaran/informasi-perizinan/uang-elektronik/penyelenggara-berizin/Contents/Default.aspx.

[3] V. D. Devita, "Who is the E-wallet Application with the Most Users in Indonesia?," 12 December 2020. [Online]. Available: https://iprice.co.id/trend/insights/e-wallet-terbaik-di-indonesia/.

[4] E. Cambria, D. Das, S. Bandyopadhyay and A. Feraco, A Practical Guide to Sentiment Analysis, Switzerland: Springer International Publishing, 2017.

[5] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis lectures on human language technologies,* vol. 5, no. 1, pp. 1-167, 2012.

[6] N. I. Prabaningtyas, I. Surjandari and E. Laoh, "Mining Customers Opinion on Services and Applications of Mobile Payment Companies in Indonesia Using Sentiment Analysis Approach," in *The 16th International Conference on Service Systems and Service Management (ICSSSM)*, 2019.

[7] V. S. Shirsat, R. S. Jagdale and S. N. Deshmukh, "Sentence Level Sentiment Identification and Calculation from News Articles Using Machine Learning Techniques," in *International Conference Computing, Communication and Signal Processing (ICCASP 2018)*, Singapore, 2019.

[8] M. I. Zul, F. Yulia and D. Nurmalasari, "Social Media Sentiment Analysis Using K-Means and Naïve Bayes Algorithm," in *2018 2nd International Conference on*

Electrical Engineering and Informatics (ICon EEI), Batam, 2018.

[9] T. Hu, H. Guo, H. Sun, T. V. T. Nguyen and J. Luo, "Spice Up Your Chat: The Intentions and Sentiment Effects of Using Emojis," in *In Eleventh International AAAI Conference on Web and Social Media*, Canada, 2017.

[10] H. Wang and J. A. Castanon, "Sentiment Expression via Emoticons on Social Media," in *IEEE International Conference on Big Data (Big Data)*, 2015.

[11] I. P. Windasari, F. N. Uzzi and K. I. Satoto, "Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek," in *2017 4th international conference on information technology, computer, and electrical engineering (ICITACEE)*, Semarang, 2017.

[12] P. S. Dandannavar, S. R. Mangalwede and S. B. Deshpande, "Emoticons and Their Effects on Sentiment Analysis of Twitter Data," in *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, 2020.

[13] R. Garreta and G. Moncecchi, Learning scikit-learn: machine learning in python, Birmingham: Packt Publishing Ltd, 2013.

[14] G. Ignatow and R. Mihalcea, An introduction to text mining: Research design, data collection, and analysis, New Delhi: Sage Publications, 2017.

[15] P. K. Novak, J. Smailovic, B. Sluban and I. Mozetic, "Sentiment of emojis," *PLoS ONE*, vol. 10, no. 12, 2015.

[16] D. H. Wahid and S. N. Azhari, "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, pp. 207-218, 2016.

[17] F. Z. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia," *M.S. thesis. M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteti van Amsterdam The Netherlands.*, 2003.

[18] A. Librian, "JSastrawi," 18 July 2020. [Online]. Available: https://github.com/jsastrawi/jsastrawi/.

[19] F. K. Chopra and R. Bhatia, "Sentiment analyzing by dictionary based approach," *International Journal of Computer Applications*, vol. 152, no. 5, pp. 32-34, 2016.

[20] J. Han, M. Kamber and J. Pei, "Data mining concepts and techniques third edition," *The Morgan Kaufmann Series in Data Management Systems*, vol. 5, no. 4, pp. 83-124, 2011.

[21] K. Wegrzyn-Wolska, L. Bougueroua, H. Yu and J. Zhong, "Explore the effects of emoticons on Twitter sentiment analysis," *Comput. Sci. Inf. Technol*, vol. 2, pp. 65-77, 2016.

[22] R. Pal, U. Pawar, K. Zambare and V.

Hole, "Predicting Stock Market Movement Based on Twitter Data and News Articles Using Sentiment Analysis and Fuzzy Logic," in *International Conference on Computer Networks and Inventive Communication Technologies*, 2019.