JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 3, September 2024, Pp. 1622-1627

OPINION MINING OF REGIONAL HEADS IN INDONESIA USING THE SUPPORT VECTOR MACHINE (SVM) METHOD

Dwi Hosanna Bangkalang*[1)]
1. Faculty of information technology UKSW, Indonesia

Article Info
Keywords: opinion mining, sentiment analysis, support vector machine, regional heads

Article history:
Received 2 June 2024
Revised 20 July 2024
Accepted 4 August 2024
Available online 1 September 2024

DOI :
https://doi.org/10.29100/jipi.v9i3.5381

E-mail address:
dwihosanna.bangkalang@uksw.edu

ABSTRACT
Social media is one of the communication mediums commonly used by regional heads to disseminate information, develop their image, and influence society through digital media. As a result, the regional head's opinion on an issue is one of the factors that piques public interest in knowing where the opinions of regional heads lie. Opinion mining is the process of obtaining information or the analysis and summarization of opinions that are automatically voiced on particular topics or issues. A method is required to convert the regional leaders' social media tweets into information and ideas that can be valuable for the community in order to see the trend of regional heads' opinions and discussion topics on social media. One method that can be used is mining the opinion of regional heads to find out their topics and sentiments in the new normal. The opinion mining method used is sentiment analysis using the Support Vector Machine (SVM) algorithm. The SVM algorithm uses a target label that will be predicted from a labeled dataset to find the optimal hyperplane that categorizes sentiment. This study aims to determine the opinion of the regional heads regarding the chosen topic for the current period of time. The findings of this study identify the regional head sentiment tendency based on model evaluations with an accuracy rate of more than 80%.

I. Introduction

SOCIAL media is one of the communication media for regional heads to the public which is often used to convey information related to policies, activities, or certain announcements [1]. Several aspects used by regional heads in posts on social media are local Cultural Approach, Humoristic Collaboration Content, direct instructional messages, interactive dialogue and content, human interest content, humanity content, and creative content [1]. By using social media, the public can quickly learn about the daily activities and viewpoints of regional heads on any issue that arises in society. Therefore, regional heads must consider the perception and aspirations of the general population when conveying information on social media in order to regulate policies [2]and attract public sympathy [3].

Opinion mining is the process of obtaining information or the analysis and summarization of opinions that are automatically voiced on particular topics or issues[4]. One of the method can be used is sentiment analysis. Sentiment analysis is a method that is often used to describe and predict public opinion or trends in any field, such as: predicting election results, predicting market sentiment for a product/brand and can also help in making decisions about a business by knowing customer sentiment [5].

The previous study titled "SVM for sentiment analysis of regional head candidates based on data on video comments on regional election debates on YouTube" used public comments on the debate videos of regional head candidates to ascertain the public's propensity for the debate's outcomes. The SVM method is used for sentiment analysis. The results of public sentiment obtained are used as evaluation material and input for regional head candidates and the winning team to improve their image and find out public expectations [6]. Another study regarding sentiment analysis is "Twitter Sentiment Analysis of Public Opinion in Regional Head Elections". This study aims to classify public opinion taken from the Twitter comments of regional head candidate pairs. The method used in this research is Naïve Bayes. The findings of this study's opinion analysis classified the public's pro and con remarks into the categories of neutral, positive, and negative comments [7]

Based on the background above, the mining of regional heads' opinions was carried out through social media to explore and see trends opinions and discussion topics of regional heads on social media. The purposed method that used in this research is opinion mining. Opinion mining is a data mining technique for extracting, classifying, and analyzing user opinions expressed in various media [8][9]which is usually known as sentiment analysis.

Sentiment analysis in general can be interpreted as a classification process to find out an author's sentiment on several aspects of a document which are usually labeled positively or negatively from the target object [5], in this case, the regional head's tweeter results. Opinion mining is applied in almost every economic and social domain because opinions are at the center of almost all human activities and are the main influence on our behavior [9][10]

In this study, the text Classification approach was used to extract opinions. Text Classification is a model used to categorize text into organized groups which can then be tagged (Positive, Negative, Neutral) on new text that is suitable to be applied based on content called classifiers [11]. The classifier needs to be trained on a machine learning model so that it can make predictions. According to several studies that have been carried out, there are several classification algorithms that can be used, namely Support Vector Machine, Naïve Bayes and Neural Network. According to research results, SVM is one of the algorithms that has the best accuracy value in producing accurate sentiment analysis[5]. The algorithm used to train the model in this research is supervised-learning: Support Vector Machine (SVM). The SVM algorithm uses a target label that will be predicted from a dataset that has been given a label to find the optimal hyperplane that categorizes sentiment [10]. The aim of this research is to acquire the regional leaders' opinions on the topics raised over a specific time period. It is hoped that the results of this research will determine trends of opinion of the regional heads and produce a model that can be used to analyze the sentiments of the regional heads.

II. RESEARCH METHOD

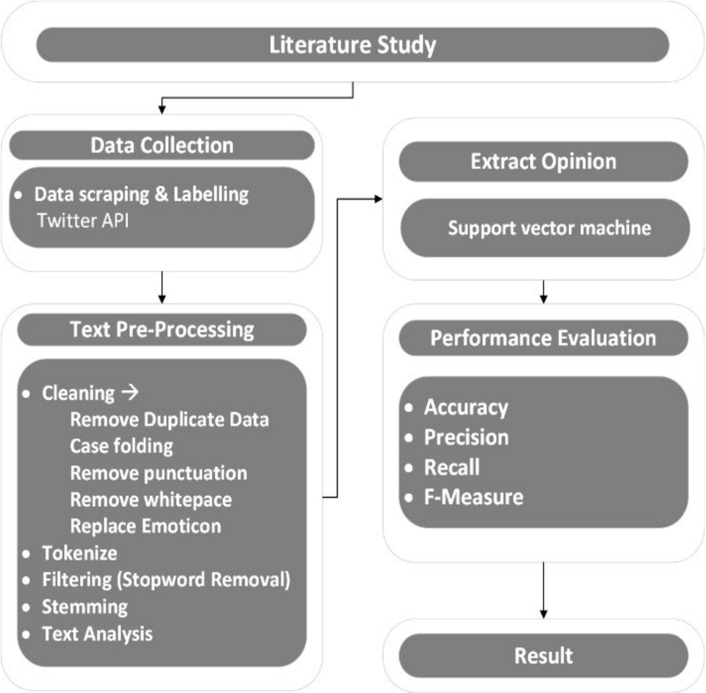

Figure 1. Research Stages

The stages of the research conducted are shown in Figure 1. First, a literature study was carried out to collect previous theories and research as a basis for conducting research. The following stage is data collection.

A. *Data Collection*

Data collection was carried out by scraping data from the Twitter accounts of the three regional heads of the largest provinces in Indonesia, namely the provinces of Central Java, West Java, and DKI Jakarta from 2021-2022. Data Scraping is a method that use to collect data from application using API [12].

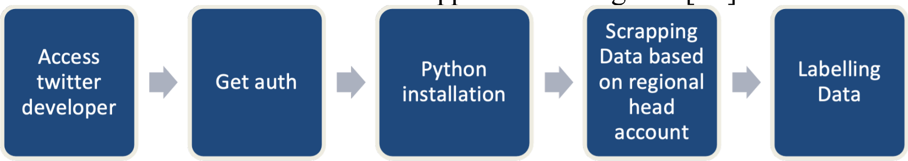

Figure. 2. Data Scrapping Stages

Figure 2 is the stages in scrapping data. First, a developer's Twitter account is required to obtain authentication in the form of key or token when scrapping data from Twitter. Data Scraping was done using python programming language and Twitter API with keywords. The total tweet data obtained is 15011 tweets. Then, pre-processing was carried out to remove duplicate data using as many as 8775 tweets which were then labeled to determine positive, negative, and neutral sentences.

Table I is the labeling result of the dataset. The labels used are positive, negative, and neutral.

TABLE I
DATASET CLASSIFICATION

Regional Heads	Positive	Negative	Neutral
DKI Jakarta	1219	327	1535
Central Java	960	361	1424
West Java	1241	369	1339

Based on the existing dataset, visualization was carried out to see the presentation of differences in sentiment classes for each regional head. The class sentiment images can be seen in Figure 3.

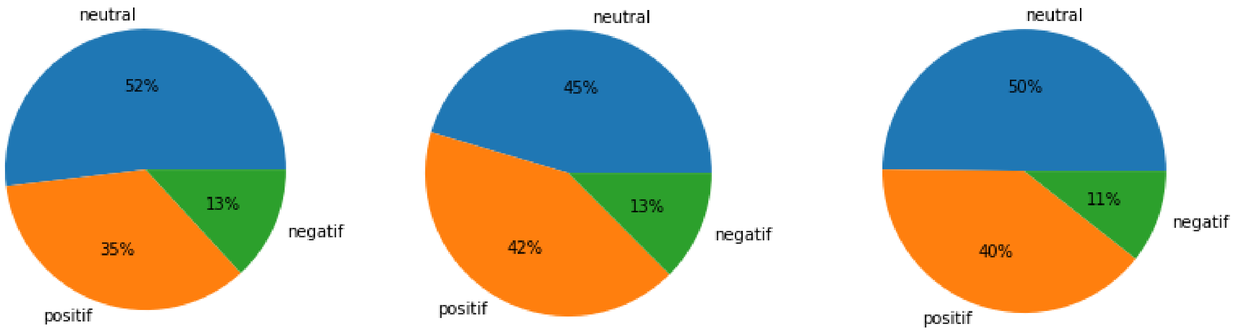

Fig 3(a). Central Java regional head class sentiment

Fig 3(b). West Java regional head class sentiment

Fig 3(c). Sentiment class head of the DKI Jakarta area

B. *Text Pre-Processing*

Text Pre-Processing is one of the important stages in sentiment analysis [13].This process is carried out to eliminate noise in the data [14] [15]so that the opinion extraction process can be processed by a better and more accurate algorithm. First, duplicate tweet data and those that came from the same username were removed to increase the validity of the dataset. Second, case folding was carried out which aims to convert text data into lowercase letters and remove characters other than letters a to z. Third, the elimination of punctuation, spaces, and emoji in the tweet data that will be processed. Fourth, tokenization was carried out. Tokenization is the process of changing sentences into tokens so that they can be analyzed. To acquire meaningful words to be processed in classification, filtering using stop words is necessary. Stopwords are common words that have no important meaning. The goal of filtering itself is to remove words that have little information from the text so that we can focus on important words instead. Such as: “a, able, about, cannot, don’t, get, become, etc.”. The stopword process is one of the important processes in text-processing because of the need to carry out sentiment analysis to focus on words that have important meanings. Lastly, in the pre-processing process is the stemming process. Stemming is the process of eliminating word inflections into basic word forms [16] such as healthy becomes health. This process is carried out to improve the quality of the information from the text to be processed[17]. To visualize words that appear frequently, wordcloud is used. Wordcloud is a method that is usually used to quickly interpret a subject domain [18]. Figure 4 is a positive sentiment wordcloud from the three datasets after labeling and cleaning. Figure 4(a) shows positive sentiment from Central Java regional heads regarding registration, education, the island of Java and competition. In figure 4(b) the positive sentiment from the West Java regional head is a normal post-pandemic adaptation, district administration. Finally, in figure 4(c) the positive sentiment for DKI Jakarta is health, government and monuments.

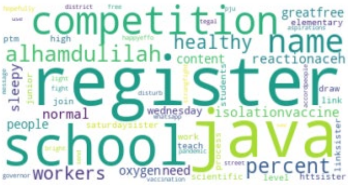

Figure 4(a). Positive wordcloud sentiment of Central Java regional head

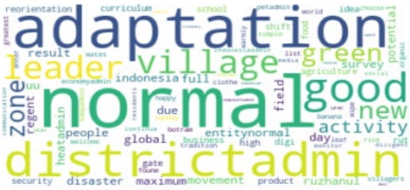

Figure 4(b). Positive wordcloud sentiment of West Java regional head

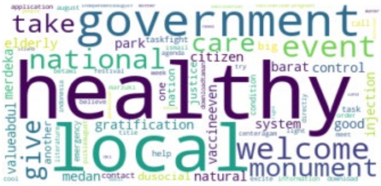

Figure 4(c). Positive wordcloud sentiment of DKI Jakarta regional head

To classify data, it is necessary to transform the data [2]into a form that can be processed by a computer. To carry out a transformation, weighting or giving values is carried out. In previous research using the SVM method, TF-IDF was the weighting that produced the best accuracy [19]. In this research, weighting is carried out using TF-IDF which transforms words into vector form[15] [20] so that they can be processed using the Support Vector Machine method. The TF-IDF formula is seen as below [21].

$$TF = \frac{f}{n_1} \tag{1}$$

TF = the frequency of occurrence of words in a document.
f = Frequency of Words in the Document
n1 = Number of Words in the Document

$$IDF = log\ log\ (\frac{Sum\ (n_1)}{n_2}) \tag{2}$$

IDF = the measure of the relative importance of the word across the entire corpus
n2 = Number of Documents Containing the Word

III. RESULT AND DISCUSSION

SVM is a method that uses a hyperplane to classify data in a higher dimensional space[14] [22] [23]. SVM describes a hyperplane using a mathematical function called a kernel. Several types of kernels are linear, sigmoid, RBF, non-linear and polynomial. Before classifying, the training data is divided by 80:20 for data testing. Several parameter experiments, including kernel, gamma, and C experiments, were conducted to obtain reliable classification results. To produce accurate classification results, several experiments were carried out with several parameters. Parameter tuning is carried out to obtain the best accuracy results.

TABLE II
EXPERIMENTAL PARAMETER C

	0.01	0.05	0.25	0.5	0.75	1	
DKI Jakarta%	74	82	85	85	85	**86**	**86**
Central Java%	73	79	81	**83**	82	82	**83**
West Java%	71	75	79	**80**	**80**	79	**80**

In table II, the best results for each dataset are different. For the DKI Jakarta area dataset, C=1 is used, for the Central Java and West Java datasets C=0.5 is used. Based on the results of the best parameter tuning, accuracy measurements were carried out using a confusion matrix. The confusion matrix has two dimensions which describe the actual output and predicted output [24] [25] The results of measuring accuracy using the confusion matrix (3) are calculated using the following formula.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \tag{3}$$

Where,
TN = True negative is the number of negative samples that are classified as accurate
TP = True positives are the number of positive samples that are classified accurately
FN = False negatives are the number of true positive samples that are classified as negative
FP = False positives are the number of true negative samples that were classified as positive

In addition, the calculation of the confusion matrix also produces precision and recall. Precision is a measurement of the model's accuracy in predicting the actual value of an item according to the item's predicated value.

Meanwhile, recall is a measurement of model accuracy based on the actual value of the item compared to the class predication results. The formula for calculating precision (4) and recall (5) is as follows.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

TABLE III (A)
CONFUSION MATRIX, CENTRAL JAVA REGIONAL HEAD

	Actual Negative	Actual Neutral	Actual Positive	Class precision
Prediction Negative	41	20	6	76%
Prediction Neutral	7	273	11	81%
Prediction Positif	6	46	139	90%
Class Recall	61%	94%	73%	

TABLE III (B)
CONFUSION MATRIX, WEST JAVA REGIONAL HEAD

	Actual Negative	Actual Neutral	Actual Positive	Class precision
Prediction Negative	29	33	9	57%
Prediction Neutral	14	234	14	76%
Prediction Positif	8	41	208	90%
Class Recall	41%	89%	81%	

TABLE III (C)
CONFUSION MATRIX, REGIONAL HEAD OF DKI JAKARTA

	Actual Negative	Actual Neutral	Actual Positive	Class precision
Prediction Negative	42	20	8	86%
Prediction Neutral	5	284	20	84%
Prediction Positif	2	36	200	88%
Class Recall	60%	92%	84%	

Table III is the result of the confusion matrix for each dataset. Of the three datasets, it was found that the positive class prediction results obtained the greatest precision accuracy, namely 90% and 80%, while for the calculation the smallest precision was in the negative class predication, namely 57% and 76% and in the DKI Jakarta provincial head dataset, neutral 84%. For recall measurements, the largest measurement values are seen in the neutral predication class, namely 89%, 94% and 92% and the smallest recall measurements are in the negative predication class, namely 41%, 61% and 60%. Based on the existing dataset, the negative class has the least amount of data, namely 11% of the total dataset. The results of the evaluation of this model can be seen in table 4 with an accuracy level above 80%, with macro average precision values of 74%, 82% and 85% and macro average recall of 70%, 76% and 78%. The performarce evaluation on Table IV show an increase in accuracy compared to previous research entitled "SVM for sentiment analysis of regional head candidates based on data on video comments on regional election debates on YouTube". There are several factors that provide increased accuracy results. One of the differences in the amount of training and testing data used is that in the previous study there were 1251 datasets, whereas in this study 8775 datasets were used. Furthermore, the parameter tuning experiments carried out also have an impact on finding the most optimal accuracy results. Based on the performance evaluation on Table IV, this research has obtained better results when compared to previous research that used the same method [6].

TABLE IV
PERFORMANCE EVALUATION

Criteria	Micro Average Regional Head of Central Java Province	Micro Average Regional Head of West Java Province	Micro Average Regional Head of DKI Jakarta Province
Accuracy	83%	80%	86%
Macro Average Precision	82%	74%	85%
Macro Average Recall	76%	70%	78%

IV. CONCLUSION

Based on the results and discussion, it can be concluded that the opinion mining of regional heads uses sentiment analysis with the SVM method. The data used is the results of tweets from the official Twitter accounts of each regional head. The level of accuracy in this model is above 80%. Based on the research results, it was found that the respective opinion tendencies of regional heads are as follows. Regional heads of DKI Jakarta Province have a negative opinion of 11%, a neutral opinion of 50% and a positive opinion of 39%. The words that appear most often in positive opinions are related to health and government. This means that regional heads of DKI Jakarta tend to give positive opinions on focused issues related to health and governance. The head of the Central Java Province has a negative opinion of 13%, a neutral opinion of 52% and a positive opinion of 35%. Words that appear in many positive opinions are related to education and competition. This means that regional heads of Central Java Province tend to give positive opinions on focus issues related to education and competition. Regional heads of West Java Province have a negative opinion of 13%, a neutral opinion of 45% and a positive opinion of 41%. The words that appear in many of these positive opinions are related to normal adaptation and district admin. This means that the regional head of West Java Province tends to give a positive opinion on the focus of issues related to normal adaptation and district administration.

REFERENCES

[1] M. Arisanty and M. Irmayanti, “Strategi Mass Self Communication Kepala Daerah di media Sosial dalam mewujudkan partisipasi masyarakat untuk menyukseskan program pemerintah daerah,” in *Akselerasi pembangunan masyarakat lokal melalui komunikasi dan teknologi informasi*, Jurusan Ilmu Komunikasi - FISIP UNIVERSITAS LAMPUNG, 2016, pp. 225–237.

[2] T. Bintang, S. Silalahi, and A. Toni, “Pengaruh Pro & Kontra Pilkada 2020 Pada Media Sosial Twitter (Drone Emprit: Pilkada 2020-Pro & Kontra).”

[3] Y. Buluamang, “Hubungan Antara Perilaku Komunikasi Kepala Daerah Dengan Citra Publik Dan Ekspektasi Publik The Relationship Between Communication Behaviors Of District Head With Public Images And Public Expectation,” *Jurnal Studi Komunikasi dan Media*, 2018.

[4] P. S and S. F. S, “Opinion Mining and Sentiment Analysis - An Assessment of Peoples’ Belief: A Survey,” *International Journal of Ad hoc, Sensor & Ubiquitous Computing*, vol. 4, no. 1, pp. 21–33, Feb. 2013, doi: 10.5121/ijasuc.2013.4102.

[5] F. Aftab *et al.*, “A Comprehensive Survey on Sentiment Analysis Techniques,” *International Journal of Technology*, vol. 14, no. 6, pp. 1288–1298, 2023, doi: 10.14716/ijtech.v14i6.6632.

[6] M. Syafa’at, E. Setyaningsih, and Kristian Yosi, “SVM_sentiment analysis_youtube,” *ANTIVIRUS: Jurnal Ilmiah Teknik Informatika*, vol. 15, no. 2, pp. 262–276, 2021.

[7] Nurfaizah, I. Prawitasari, and Fathuzaen, “Analisis Sentimen Twitter Terhadap Opini Publik Pemilihan Kepala Daerah,” in *CITISEE* , 2018.

[8] T. Y. Kim and H. J. Kim, “Opinion Mining-Based Term Extraction Sentiment Classification Modeling,” *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/5593147.

[9] B. Liu, “Sentiment Analysis and Opinion Mining,” Morgan & Claypool Publishers, 2012.

[10] M. Ahmad, S. Aftab, and I. Ali, “Sentiment Analysis of Tweets using SVM,” *Int J Comput Appl*, vol. 177, no. 5, pp. 25–29, 2017, doi: 10.5120/ijca2017915758.

[11] N. Naw, “Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers,” *International Journal of Scientific and Research Publications (IJSRP)*, vol. 8, no. 10, Oct. 2018, doi: 10.29322/ijsrp.8.10.2018.p8252.

[12] N. Adila, “Implementation of Web Scraping for Journal Data Collection on the SINTA Website,” *Sinkron*, vol. 7, no. 4, pp. 2478–2485, Oct. 2022, doi: 10.33395/sinkron.v7i4.11576.

[13] M. Aqib, R. Mehmood, A. Alzahrani, I. Katib, A. Albeshri, and S. M. Altowaijri, *Smarter traffic prediction using big data, in-memory computing, deep learning and gpus*, vol. 19, no. 9. 2019. doi: 10.3390/s19092206.

[14] S. Fransiska and A. Irham Gufroni, “Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method,” *Scientific Journal of Informatics*, vol. 7, no. 2, 2020, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/sji

[15] V. I. Santoso, G. Virginia, and Y. Lukito, “Penerapan Sentiment Analysis Pada Hasil Evaluasi Dosen Dengan Metode Support Vector Machine,” *Jurnal Transformatika*, vol. 14, no. 2, p. 72, 2017, doi: 10.26623/transformatika.v14i2.439.

[16] Imamah, Husni, M. Rachman, I. Suzanti, and F. Muffaroha, “Text Mining and Support Vector Machine for Sentiment Analysis of Tourist Reviews in Bangkalan Regency,” in *ICComSET*, Institute of Physics Publishing, 2020. doi: 10.1088/1742-6596/1477/2/022023.

[17] B. Indriyono, E. Utami, and A. Sunyoto, “Pemanfaatan Algoritma Porter Stemmer Untuk Bahasa Indonesia Dalam Proses Klasifikasi Jenis Buku,” *Jurnal Buana Informatika*, vol. 6, no. 4, pp. 301–310, 2015.

[18] Y. Kalmukov, “Using Word Clouds for Fast Identification of Papers’ subject domain and Reviewer’ Competences.” [Online]. Available: www.compsystech.org

[19] Muhammad Kiko Aulia Reiki, Y. Sibaroni, and E. B. Setiawan, “Comparison of Term Weighting Methods in Sentiment Analysis of the New State Capital of Indonesia with the SVM Method,” *International Journal on Information and Communication Technology (IJoICT)*, vol. 8, no. 2, pp. 53–65, Jan. 2023, doi: 10.21108/ijoict.v8i2.681.

[20] R. Wijayanti and A. Arisal, “Ensemble Approach for Sentiment Polarity Analysis in User-Generated Indonesian Text,” in *International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2017, pp. 158–163. doi: 10.1109/IC3INA.2017.8251759.

[21] A. Veluchamy, H. Nguyen, M. L. Diop, and R. Iqbal, “Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches,” *SMU Data Science Review*, vol. 1, no. 4, 2018, [Online]. Available: https://scholar.smu.edu/cgi/viewcontent.cgi?article=1051&context=datasciencereview

[22] S. Styawati and K. Mustofa, “A Support Vector Machine-Firefly Algorithm for Movie Opinion Data Classification,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 3, p. 219, 2019, doi: 10.22146/ijccs.41302.

[23] W. Muhammad, M. Mushtaq, K. N. Junejo, and M. Y. Khan, “Sentiment analysis of product reviews in the absence of labelled data using supervised learning approaches,” *Malaysian Journal of Computer Science*, vol. 33, no. 2, pp. 118–132, 2020, doi: 10.22452/mjcs.vol33no2.3.

[24] A. Kulkarni, D. Chong, and F. A. Batarseh, “Foundations of data imbalance and solutions for a data democracy,” in *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*, Elsevier, 2020, pp. 83–106. doi: 10.1016/B978-0-12-818366-3.00005-8.

[25] Xinyang Deng, Qi Liu, Yong Deng, and Sangkaran Mahadevan, “An improved method to construct basic probability assignment based on the confusion matrix for classification problem,” *Inf Sci (N Y)*, pp. 250–261, 2016.