Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency

Nur Heri Cahyana[1], Shoffan Saifullah[2], Yuli Fauziah[3], Agus Sasmito Aribowo[4], Rafal Drezewski[5]
Department of Informatics, Universitas Pembangunan Nasional Veteran Yogyakarta Yogyakarta, Indonesia[1,2,3]
Department of Information Systems, Universitas Pembangunan Nasional Veteran Yogyakarta Yogyakarta, Indonesia[3]
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia[4]
Institute of Computer Science, AGH University of Science and Technology, Cracow, Poland[2,5]

***Abstract*—Sentiment analysis can detect hate speech using the Natural Language Processing (NLP) concept. This process requires annotation of the text in the labeling. However, when carried out by people, this process must use experts in the field of hate speech, so there is no subjectivity. In addition, if processed by humans, it will take a long time and allow errors in the annotation process for extensive data. To solve this problem, we propose an automatic annotation process with the concept of semi-supervised learning using the K-Nearest Neighbor algorithm. This process requires feature extraction of term frequency-inverse document frequency (TF-IDF) to obtain optimal results. KNN and TF-IDF were able to annotate and increase the accuracy of < 2% from the initial iteration of 57.25% to 59.68% in detecting hate speech. This process can annotate the initial dataset of 13169 with the distribution of 80:20 of training and testing data. There are 2370 labeled datasets; for testing, there are 1317 unannotated data; after preprocessing, there are 9482. The final results of the KNN and TF-IDF annotation processes have a length of 11235 for annotated data.**

Keywords*—*Natural language processing; text annotation; semi-supervised learning; TF-IDF; K-NN

I. Introduction

The concept of text mining in natural language processing is often experienced in the annotation process, including the length of the human annotation process in data labeling. This annotation process also often causes errors due to time pressure and instructions to complete it [1]. In addition, sometimes, they are not trained and skilled in annotating specific fields. Thus, it is necessary to annotate with little knowledge (data labels) from humans semi-automatically, making complete annotations with machine learning.

Text classification also includes processing, which puts documents into predetermined categories [2]. Text classification can be done for solving several cases, such as sentiment analysis [3], emotion analysis [4], and hate speech detection [5], [6].

This study discusses detecting hate speech in the text, especially in text annotation. The discussion starts with determining the hate speech category, then grouping documents into those categories and validating them. The hate speech detection process in documents uses the basic principles of sentiment analysis, starting with document preprocessing, vectorization, modeling, and validation. There are three models in the classification of sentiment analysis (or hate speech): machine learning, lexicon, and mixed models [7]. The lexicon and mixed models need a hate speech dictionary. This study uses a machine learning approach because the lexicon of hate speech in Indonesian is not widely available. The machine learning approach is divided into three approaches: supervised, unsupervised, and semi-supervised [8]. The unsupervised approach causes hate speech categories not to be directed as needed. It all depends purely on the condition of the document features. However, in supervised and semi-supervised, researchers can direct the categorization of hate speech into two or three categories: very hate speech, low hate speech, and non-hate speech. Based on existing research, annotators generally polarize hate speech into only two categories (hate speech and non-hate speech). Human annotators can assess the presence of hate speech in a document.

The hate speech annotation process is where experts separate or provide information on documents into two categories: groups of documents containing elements of hate (hate speech) and groups of documents that do not contain elements of hate (non-hate speech). This annotation follows the ways of annotating sentiments which are generally in two polarities (positive and negative) as used in [9]–[18]. This annotation process requires an expert (human annotator) who understands the meaning of hate speech and has experience annotating opinion documents.

This study aims to use the model of semi-supervised text annotations automatically by K-Nearest Neighbor. In addition, we have not been able to determine the best vectorizer because, as in [14], we used TF-IDF. This study's results differ from [19] with increased accuracy in the model that produces hate speech annotated datasets.

II. Method

A. Similar Research on Semi-supervised Text Annotation

Typically, in semi-supervised text annotations, the annotator uses the sentiment lexicon to annotate the unlabeled data and manually revise the annotated data sample. This approach requires more time to revise the annotations [20]. AraSenCorpus in [20] is a self-learning approach to automate annotations and reduce human effort. AraSenCorpus is a semi-

supervised framework for annotating a sizeable Arabic text corpus using a small subset of manually annotated tweets and extending it from a large set of unlabeled tweets to reduce human effort in annotating. This process uses the FastText neural network and the LSTM [21] deep learning classifier to manually expand the annotated corpus and ensure the quality of the newly created corpus, respectively (Fig. 1).

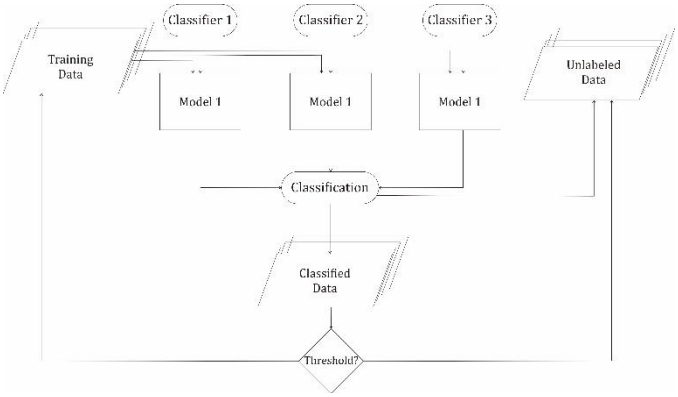

Fig. 1. AraSenCorpus Architecture [20].

This study performs a two-way (positive and negative) and three-way (positive, negative, and neutral) sentiment classification. In the case of two-way classification, AraSenCorpus improved the sentiment classification results from 80.37% to 87.4% using the 2017 SemEval dataset and from 79.77% to 85.2% using the ASTD dataset. The three-way classification gives 69.4% accuracy for the SemEval 2017 dataset, while the best system gives 63.38% using the F1-score and from 64.10% to 68.1% using the ASTD dataset. However, according to our assessment, the classification process is not determined by the type of classifier. Accuracy also depends on the vectorizer used. Another weakness is that if the iteration has been done many times and the classification results are consistently below the threshold, there is no visible solution to whether the dataset will still be included or discarded.

Another semi-supervised annotation study involved two targets: sentiment analysis and emotion analysis on an English textual review of three digital payment applications. The approach used involved supervised and unsupervised machine learning techniques. Data annotation involves three assistants from the field of Psychology. Annotators were recruited to label sentiments and emotions for 3,000 reviews. If the sentiment is neutral, the emotion is labeled neutral and excluded from the emotion analysis because no emotion can be detected from the neutral document. The machine learning algorithms are Support Vector Machine, Random Forest, and Naïve Bayes. Random Forest yielded the best accuracy for sentiment (F1 score = 73.8%; Kappa Cohen = 52.2%) and emotion (F1 score = 58.8%; Kappa Cohen = 44.7%) [22]. The architecture of the model used is shown in Fig. 2. The advantage of the approach used in [22] is that analyzing sentiment and emotion is carried out simultaneously. The downside is that the average accuracy for emotion classification based on Random Forest and SVM is around 61.3% (deficient), even though it uses a quite sophisticated algorithm. Another weakness is in the annotation process step. Sentiment and emotion annotations are carried out simultaneously so that it is prone to ambiguity in labeling positive sentiments with negative emotions (anger, sadness, fear, disgust) or vice versa. This model also cannot anticipate if the annotation results are low in accuracy as in the AraSenCorpus Architecture (Fig. 2) [20].

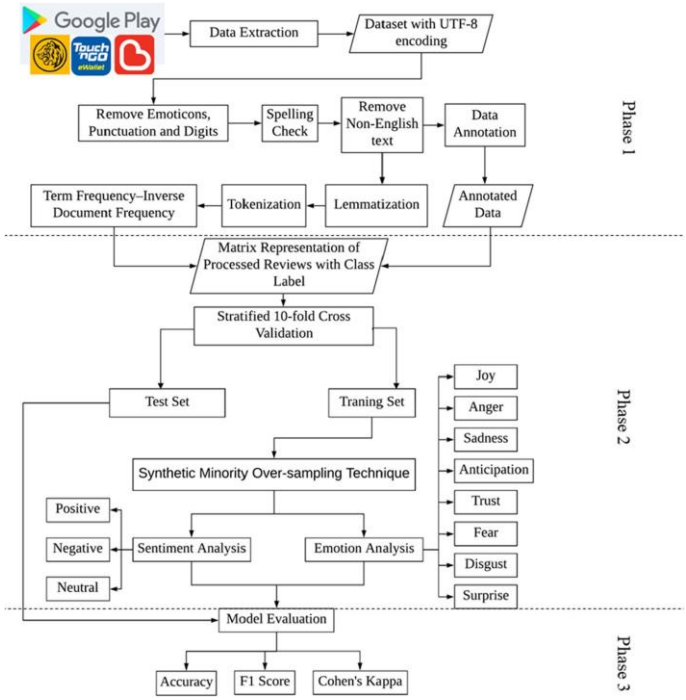

Fig. 2. Semi-Supervised Sentiment–Emotion Architecture [22].

B. Proposed Method

The semi-supervised text annotation method for hate speech that we propose is a new method that has never been used in any research. The semi-supervised text annotation begins with reading 2000 hate-speech training data and 9000 non-annotated testing data. The flow of the semi-supervised text annotation process is shown in Fig. 3.

The semi-supervised hate speech annotation process (Fig. 3) starts from step 1, reading the DT as training data annotated as hate speech (by experts). Step 2 reads unannotated UD data. Step 3 is the text preprocessing of DT and UD. Step 4 is the meta-vectorization process. The meta-vectorization process converts the clean DT and UD datasets into four types of vectors. The first vector is VBoW, created using the Bag-Of-Word method. The second vector is VTFIDF, created using the TF-IDF method [23], [24]. Step 5 is the process of preparing training data. Step 6 is the setup of machine learning algorithms.

The algorithm involved is K-Nearest Neighbor (K-NN). Step 7 is the creation of a meta-learning model. The meta-vector and meta-learning models will produce vectorization and machine-learning approaches. The combination of machine learning will be used for the auto-annotation process. Steps 8 and 9 prepare vector datasets that have not been annotated. In step 10, it will be checked if there is still a dataset that has not been annotated then the process will continue to step 11, namely the process of annotating the dataset. The annotation process is done by predicting labels by the vector and machine

learning combinations. The prediction results are also subject to validation to determine their accuracy—the auto-annotation process results in Step 12 as a meta-labeled dataset. In Step 12, a voting process will be carried out to determine the label for each dataset record. The voting counts used the sample of the voting process for labeling decisions. There are two types of weight: the total W (weight) score of the vectorization-machine-learning model for the hate-speech polarity and the W (weight) score from the vectorized-machine-learning model for the non-hate speech polarity. In Step 13, if the polarity score of hate speech or non-hate speech exceeds the threshold, then the dataset and its annotations will be transferred to data training. If not, it will be re-annotated in the next cycle. In step 14, the data will be looping processed three times for optimization. If it is more than three times, then the rest of the Unlabeled Dataset will go to Step 15, which is manual annotation. The results of manual annotations will be combined with the training data.

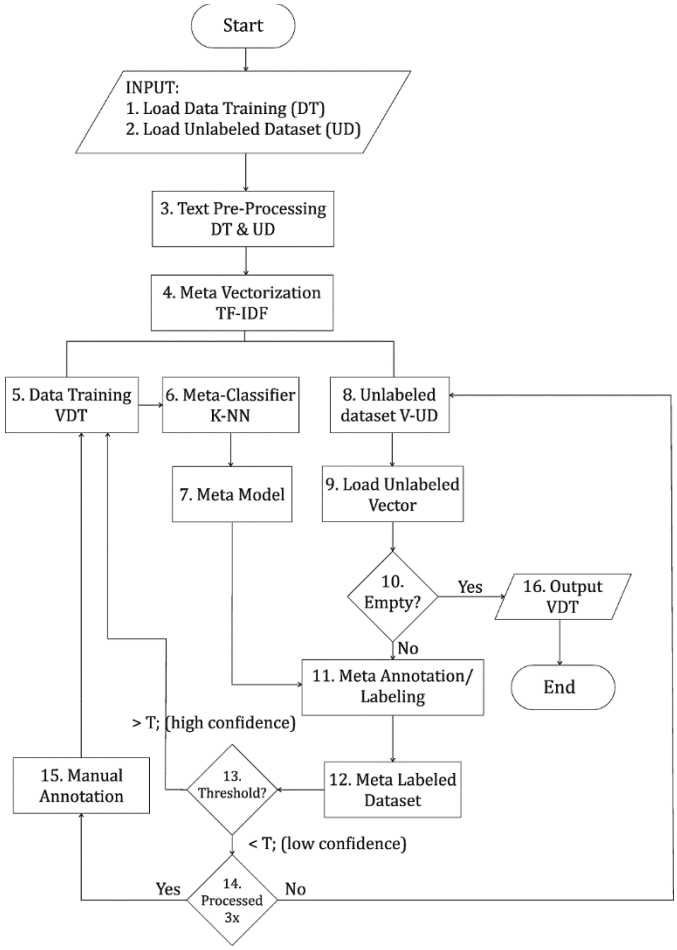

Fig. 3. Semi-Supervised Hate Speech Annotation Model (Proposed).

III. Results and Discussion

A. Population and Sampling

Based on the monitoring of data sources, it is known that the dataset size is dynamic, meaning that these opinions continue to grow, even though the presidential-vice presidential debate or the Covid-19 pandemic has already occurred. A new video on the official channel also appeared, followed by viewers' opinions. So this study concludes that the population size is unknown, and it is impossible to download all comments from the YouTube repository. So the research determined that the sampling method used was purposive sampling. This purposive sampling method was chosen for the following reasons:

1) Our initial observations found hate speech in YouTube video comments on the topic of the 2019 Indonesian presidential debate and Covid-19. So, the data are suitable for our study.

2) The data from YouTube video comments are public and can be downloaded for free

The number of videos related to the presidential debate samples required refers to previous similar studies. In the process of getting the population of comments from videos, this study uses videos that meet the following criteria:

1) The data collection stage downloads all comments from the presidential debates one to five, each broadcast in full by two official channels. So this study downloads comments from 10 presidential debate videos and five Covid-19 news.

2) The data collection stage also downloads from the official channel comments from videos that do not show the whole presidential debate but are considered necessary to download because of the high number of views, more than 10,000 views, and comments above 1000 comments on exciting topics.

B. Data Annotation

In supervised learning, opinion data must be annotated by experts responsible for labeling hate speech on opinion data. So the data labeling process is the first step of knowledge transfer before categorization is carried out. The level of hate speech used for the labeling process is shown in Table I.

TABLE I. Hate Speech Levels in this Research

No	Level	Code	Information
1.	Very Hate	VH	Hate speech that has the potential to cause dangerous social unrest
2.	Hate	H	Hate speech does not have the potential to cause harmful social unrest
3.	Non-Hate	NH	No hate speech

An annotator is an expert with expertise and knowledge following the political realm. Experts know the fields of social humanities and information technology (social media). The method of annotating opinions has also been determined. There are two experts and a team, all of which annotate hate speech, sentiment, and emotion. Steps of the annotation process:

1) The first expert will take around 2,000 opinions and then describe Very Hate, Hate or non-Hate.

2) The first expert's annotation results will be kept and not given to the second expert. Only comments from the first-choice expert are given to the second expert for hate speech annotation. The second expert does not know the first expert's annotation label. Then proceed with automatic annotation by

algorithms on data sets that experts do not annotate. This method is a semi-supervised text annotation based on meta-learning, which will automatically perform hate speech annotation, as shown in Fig. 4.

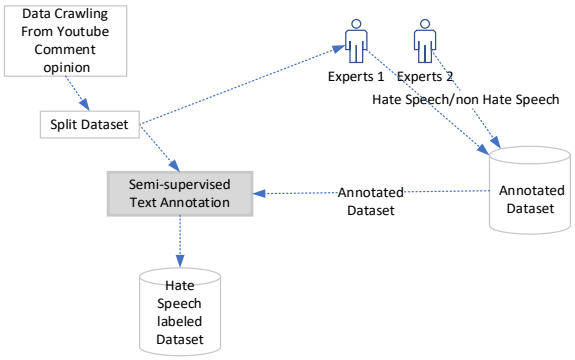

Fig. 4. Hate Speech Levels in this Research.

C. Scenarios and the Results of Text Annotation using K-NN and TF-IDF

This study uses a scenario with the composition of training data, testing data, and threshold: 20%, 80%, and 80%. As for the training data, the initial labeling process was carried out by experts. This annotation process is tested using data, as shown in Table II. In contrast, 80% of data is testing data used to verify the accuracy of the annotation process with a data-limiting threshold of 80%.

TABLE II. SCENARIO

No	Parameters	Value	Information
1.	Threshold	80	Presents
2.	Data training annotated	2370	Length of datasets
3.	Data training un-annotated	9482	Length of datasets
4.	Data testing	1317	Length of datasets
5.	Total of datasets	13169	Length of datasets

The KNN method can perform initial training data with a sample of 20% and has an accuracy of 57.25%. After validating using TF-IDF vectorization, the best iteration is 59.68% so this method can increase the accuracy by 2.43%.

The implementation of KNN and TF-IDF with this scenario resulted in 11,235 annotated data from the total dataset processed after preprocessing of 11,852. So in this process, there are still 617 data that have not been annotated. It happens because the process still has shortcomings from the initial annotation process. In the future, the initial accuracy of the annotation will be improved to optimize the final annotation. In addition, other vectorization and classification methods will also be implemented.

IV. CONCLUSION

The results of the presented research are the KNN and TF-IDF models of speech classifiers with semi-supervised hate speech annotations to detect hate speech on social media with low accuracy. The applied TF-IDF vectorization method increased the accuracy by 2.43%. Thus, future works will increase by using the percentage variation of data labeling on the initial annotation (5%, 10%, 20%) and threshold (0.6, 0.7, 0.8, and 0.9). Besides, the vectorization will be improved by applying methods such as Bag-Of-Word and Word2Vec. The classification methods will be improved and compared with other methods like Random Forest (RF), Extra Tree (ET), Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT).

ACKNOWLEDGMENT

This research was supported by LPPM Universitas Pembangunan Nasional “Veteran” Yogyakarta.

REFERENCES

[1] K. Miok, G. Pirs, and M. Robnik-Sikonja, “Bayesian Methods for Semi-supervised Text Annotation,” 14th Linguist. Annot. Work., pp. 1–12, 2020, [Online]. Available: http://arxiv.org/abs/2010.14872.

[2] M. Bouazizi and T. Ohtsuki, “Multi-class sentiment analysis on twitter: Classification performance and challenges,” Big Data Min. Anal., vol. 2, no. 3, pp. 181–194, 2019, doi: 10.26599/BDMA.2019.9020002.

[3] M. Chen, K. Ubul, X. Xu, A. Aysa, and M. Muhammat, “Connecting Text Classification with Image Classification: A New Preprocessing Method for Implicit Sentiment Text Classification,” Sensors, vol. 22, no. 5, p. 1899, Feb. 2022, doi: 10.3390/s22051899.

[4] M. Z. Asghar et al., “A Deep Neural Network Model for the Detection and Classification of Emotions from Textual Content,” Complexity, vol. 2022, pp. 1–12, Jan. 2022, doi: 10.1155/2022/8221121.

[5] P. William, R. Gade, R. esh Chaudhari, A. B. Pawar, and M. A. Jawale, “Machine Learning based Automatic Hate Speech Recognition System,” 2022 Int. Conf. Sustain. Comput. Data Commun. Syst., pp. 315–318, Apr. 2022, doi: 10.1109/ICSCDS53736.2022.9760959.

[6] K. Miok, B. Škrlj, D. Zaharie, and M. Robnik-Šikonja, “To BAN or Not to BAN: Bayesian Attention Networks for Reliable Hate Speech Detection,” Cognit. Comput., vol. 14, no. 1, pp. 353–371, Jan. 2022, doi: 10.1007/s12559-021-09826-9.

[7] P. Sudhir and V. D. Suresh, “Comparative study of various approaches, applications and classifiers for sentiment analysis,” Glob. Transitions Proc., vol. 2, no. 2, pp. 205–211, 2021, doi: 10.1016/j.gltp.2021.08.004.

[8] C. R. Aydln and T. Güngör, “Sentiment analysis in Turkish: Supervised, semi-supervised, and unsupervised techniques,” Nat. Lang. Eng., vol. 27, no. 4, pp. 455–483, 2021, doi: 10.1017/S1351324920000200.

[9] S. Aman and S. Szpakowicz, “Identifying expressions of emotion in text,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4629 LNAI, no. September 2007, pp. 196–205, 2007, doi: 10.1007/978-3-540-74628-7_27.

[10] A. Krouska, C. Troussas, and M. Virvou, “The effect of preprocessing techniques on Twitter sentiment analysis,” IISA 2016 - 7th Int. Conf. Information, Intell. Syst. Appl., 2016, doi: 10.1109/IISA.2016.7785373.

[11] A. M. Ningtyas and G. B. Herwanto, “The Influence of Negation Handling on Sentiment Analysis in Bahasa Indonesia,” Proc. 2018 5th Int. Conf. Data Softw. Eng. ICoDSE 2018, pp. 1–6, 2018, doi: 10.1109/ICODSE.2018.8705802.

[12] J. Savigny and A. Purwarianti, “Emotion classification on Youtube comments using word embedding,” in International Conference on Advanced Informatics: Concepts, Theory and Applications, 2017, pp. 1–5, doi: 10.1109/ICAICTA.2017.8090986.

[13] K. Mulcrone, “Detecting Emotion in Text,” 2012.

[14] W. C. F. Mariel, S. Mariyah, and S. Pramana, “Sentiment analysis: A comparison of deep learning neural network algorithm with SVM and naïve Bayes for Indonesian text,” J. Phys. Conf. Ser., vol. 971, no. 1, 2018, doi: 10.1088/1742-6596/971/1/012049.

[15] T. Sutabri, A. Suryatno, D. Setiadi, and E. S. Negara, “Improving Naïve Bayes in Sentiment Analysis For Hotel Industry in Indonesia,” in Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018, 2018, pp. 1–6, doi: 10.1109/IAC.2018.8780444.

[16] M. Lailiyah, S. Sumpeno, and I. K. E. Purnama, “Sentiment analysis of public complaints using lexical resources between Indonesian sentiment

lexicon and sentiwordnet," in 2017 International Seminar on Intelligent Technology and Its Application: Strengthening the Link Between University Research and Industry to Support ASEAN Energy Sector, ISITIA 2017 - Proceeding, 2017, vol. 2017-Janua, pp. 307–312, doi: 10.1109/ISITIA.2017.8124100.

[17] U. Makhmudah, S. Bukhori, J. A. Putra, and B. A. B. Yudha, "Sentiment Analysis of Indonesian Homosexual Tweets Using Support Vector Machine Method," Proc. - 2019 Int. Conf. Comput. Sci. Inf. Technol. Electr. Eng. ICOMITEE 2019, pp. 183–186, 2019, doi: 10.1109/ICOMITEE.2019.8920940.

[18] T. F. Abidin, M. Hasanuddin, and V. Mutiawani, "N-grams based features for Indonesian tweets classification problems," Proc. - 2017 Int. Conf. Electr. Eng. Informatics Adv. Knowledge, Res. Technol. Humanit. ICELTICs 2017, vol. 2018-Janua, no. ICELTICs, pp. 307–310, 2017, doi: 10.1109/ICELTICS.2017.8253287.

[19] Nurfaizah, T. Hariguna, and Y. I. Romadon, "The accuracy comparison of vector support machine and decision tree methods in sentiment analysis," J. Phys. Conf. Ser., vol. 1367, no. 1, 2019, doi: 10.1088/1742-6596/1367/1/012025.

[20] A. Al-Laith, M. Shahbaz, H. F. Alaskar, and A. Rehmat, "AraSenCorpus: A Semi-Supervised Approach for Sentiment Annotation of a Large Arabic Text Corpus," Appl. Sci., vol. 11, no. 5, pp. 1–19, 2021, doi: 10.3390/app11052434.

[21] Y. Mao, A. Pranolo, A. P. Wibawa, A. B. Putra Utama, F. A. Dwiyanto, and S. Saifullah, "Selection of Precise Long Short Term Memory (LSTM) Hyperparameters based on Particle Swarm Optimization," 2022 Int. Conf. Appl. Artif. Intell. Comput., pp. 1114–1121, May 2022, doi: 10.1109/ICAAIC53929.2022.9792708.

[22] V. Balakrishnan, P. Y. Lok, and H. Abdul Rahim, "A semi-supervised approach in detecting sentiment and emotion based on digital payment reviews," J. Supercomput., vol. 77, no. 4, pp. 3795–3810, 2021, doi: 10.1007/s11227-020-03412-w.

[23] Y. Fauziah, S. Saifullah, and A. S. Aribowo, "Design Text Mining for Anxiety Detection using Machine Learning based-on Social Media Data during COVID-19 pandemic," in Proceeding of LPPM UPN "Veteran" Yogyakarta Conference Series 2020–Engineering and Science Series, 2020, vol. 1, no. 1, pp. 253–261, doi: 10.31098/ess.v1i1.117.

[24] S. Saifullah, Y. Fauziah, and A. S. Aribowo, "Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety Based on Social Media Data," Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.06353.