# Active Learning on Indonesian Twitter Sentiment Analysis Using Uncertainty Sampling

Muhaza Liebenlito [1,*], Nur Inayah[2], Esti Choerunnisa[3], Taufik Edy Sutanto[4], Suma Inna[5]

[1,2,3,4,5] *Department of Mathematics, Faculty of Science and Technology, UIN Syarif Hidayatullah, Jakarta 15412, Indonesia*

(Received: November 22, 2023; Revised: December 18, 2023; Accepted: January 15, 2024; Available online: January 29, 2023)

**Abstract**

Nowadays, sentiment analysis research in social media is rapidly developing. Sentiment analysis typically falls under supervised learning, which requires annotating data. However, the annotation process for sentiment analysis tasks is notoriously time-consuming. An effective strategy to overcome this challenge, known as active learning, has emerged. Active learning involves labeling only a small subset of the dataset, leaving the rest for annotation through sampling strategies. This study focuses on comparing two active learning strategies: random sampling and boundary sampling. These strategies are applied to machine learning models such as logistic regression and random forests. In addition, we present an evaluation of the model performance and data savings achieved by implementing these strategies in the context of traditional machine learning for sentiment analysis on Twitter, and the dataset consists of two labels: positive and negative sentiments. The results of our investigation show that an uncertainty sampling strategy can significantly reduce the amount of training data required, saving up to 65% of the total training data required to achieve peak model accuracy. The best model obtained in this experiment is a random forest with a margin sampling strategy, yielding an accuracy of 81.12% and an F1 score of 88.60%. This research highlights the effectiveness of active learning strategies in sentiment analysis, demonstrating their potential to improve model performance and resource efficiency. The results underscore the viability of employing active learning methods, particularly the combination of random forest models with margin sampling, which can achieve more efficiency regarding data usage in social media sentiment analysis.

*Keywords:* Active Learning, Uncertainty Sampling, Logistic Regression, Random Forest, Sentiment Analysis

## 1. Introduction

Recently, the rapid development of Information Technology has been intensely felt, which means that using the Internet cannot be avoided on a daily basis. According to a report by We Are Social [1], the number of internet users in Indonesia reached 212.9 million as of January 2023. This means that about 77% of the Indonesian population uses the Internet. The internet is used for online socializing through social media, and one of the most popular social media platforms is Twitter. Many posts are uploaded by users, including images, videos, and text. All the posts uploaded to social media represent a considerable amount of data that can be mined and analyzed to uncover new useful information. The data found on Twitter can be used to obtain specific information, making it attractive for deeper analysis. When evaluating public opinions on Twitter, sentiment analysis can be used to categorize those opinions as positive or negative. However, a challenge in sentiment analysis is the informal language used on Twitter [2]. The obtained data must be preprocessed before it is ready for processing. Sentiment analysis has several methods, and one of the approaches is machine learning [3]. The machine learning approach creates models that are used in the classification process. Text classification categorizes text data into predefined groups or classes.

Data annotation or labeling assigns one or more labels to a data set, allowing algorithms to learn and predict labeled data. In supervised learning, labeled training data is used in the training process to produce the output of a model. Obtaining labeled training data can be a time-consuming task, as acquiring labeled data typically requires human assistance, especially when dealing with very large training datasets [4]. Therefore, it is necessary to implement a mechanism for selecting informative and valuable data during model development, thus reducing the required training

***Corresponding author: Muhaza Liebenlito (muhazaliebenlito@uinjkt.ac.id)**
**DOI: https://doi.org/10.47738/jads.v5i1.144**

data while still achieving optimal performance results [5][6]. One of the methods that can be used to address the above problem is active learning. This problem is addressed in data-centric AI frameworks [7].

Active learning is a method where a classifier can actively select the most informative training data to build a model [8]. A common questioning strategy for active learners is uncertainty sampling. Like humans asking questions about things they do not know, uncertainty sampling labels the most uncertain points. In general, the unlabeled data that confuses the algorithm the most is the most valuable and is labeled and added to the training data.

Previous study found that uncertainty sampling has proven effective and efficient [9], and there have been many developments in uncertainty sampling methods [10]. A study by [11] provides a framework for active learning strategies involving uncertainty sampling. This study concluded that active learning using uncertainty sampling could reduce human labeling. Another study by [12] conducted a survey comparing various active learning strategies and found that, in general, the least confidence and margin sampling strategies performed better than other query strategies, and overall, active learning outperformed passive learning. Furthermore, Agharwal et al. [13] use a margin sampling strategy and conditional random field (CRF) as a classifier to classify disease names in biomedical text datasets.

Based on the previous studies, this study aims to implement active learning with uncertainty sampling in sentiment analysis. The sentiment analysis task on Indonesian Twitter related to COVID-19 and daily conversations will be conducted. In addition, the study will compare the performance of the margin sampling and random sampling that affect the performance of logistic regression and random forests in the active learning scenarios.

## 2. Method

### 2.1. Active Learning

Active learning is a machine-learning approach that allows humans to participate actively in the learning process [14]. The goal of active learning is to make the most informative queries from unlabeled data based on the output of the learning algorithm so that the labeling results of these queries can improve the model's performance [15][5]. Therefore, the desired performance can be achieved with fewer and faster queries with a random selection [16]. Active learning enhances model performance by adding well-selected data to the training data, enabling the model to learn from more relevant data while reducing the overall data required for training [17][18].

This study applies active learning to the data using query strategies, which are part of uncertainty sampling, namely margin and random sampling. The steps of the active learning process in this study are as follows [10]: (a) the model is constructed using pre-defined training data; (b) it evaluates the sentiment of previously unknown text; (c) the model selects data using the specified query strategy; (d) after that, the data selected by the query strategy is added to the training data, and the model is updated; (e) this process is repeated until the model achieves the desired performance.

### 2.2. Margin Sampling

One variant of the uncertainty sampling query strategy is margin sampling. Margin sampling addresses the shortcomings of the least confident strategy, which only considers information about the most likely label and effectively discards the remaining label distribution. This strategy considers the two most likely classes: the highest and second-highest probability [15]. The margin sampling can be written as

$$x_m^* = \text{argmin}_\theta (p_\theta(\hat{y}_1|x) - p_\theta(\hat{y}_2|x)) \quad (1)$$

where $\hat{y}_1$ and $\hat{y}_2$ are the positive and negative labels, which are high probability classes when predicted using the model. It takes the argument of minimum values of the margins because the difference between the two most likely classes indicates the model's uncertainty in predicting the data sample $x_m^*$. Therefore, the lowest difference represents the most uncertain and informative sample. The algorithm for margin sampling is given below:

INPUT: Given dataset containing two classes and a large portion of unlabeled data, also known as pool $P$. The number of batch sizes.

OUTPUT: Model performance with respect to the number of iterations

```
Initialization: Train the model on a small portion of the labeled dataset with respect to the number of batch sizes. After that, remove the trained data from $P$.
While $P$ is not $\emptyset$ do
    Apply (1) to find the candidate of data and label the data manually.
    Update the model and remove the candidate from $P$.
    Calculate the model performance.
```

## 3. Result and Discussion

The data used in this study consists of two datasets: Dataset 1, available in [19], focuses on Indonesian netizens' comments about COVID-19. The keyword used for data collection was "Covid-19 di Indonesia," it was collected from March 2, 2021, to March 29, 2021. The dataset used in this study comprises 26,170 tweets. It is stored in comma-separated values (CSV) format. Dataset 2 [20] contains text content from public conversations on the social media platform Twitter. It was collected between September and December 2018, using common words from everyday conversations as keywords. During that period, a total of 454,559 tweets were collected. These tweets were then preselected to choose those suitable for training sentiment analysis models, resulting in a final dataset of 10,806 tweets. The code and data can be found at https://github.com/Estich85/ActiveLearning_MarginSampling.

The collected data still contains various elements, such as words, image links, or unnecessary video links. Therefore, data cleaning is necessary. Text preprocessing is a crucial step in text mining because it makes the data cleaner and facilitates further analysis, reducing the chances of errors during model evaluation. The processes involved are as follows: Case folding: this process aims to convert the entire text into a consistent format, usually converting all characters to lowercase. Removing punctuation, hashtags, symbols, and numbers: this process reduces noise by eliminating punctuation marks, hashtags, symbols, and numeric characters. Tokenizing: involves breaking the input string into individual words or tokens. In this step, each word in the text is separated, and spaces are used as separators between words. Stopwords are words that do not have significant meaning and often occur with high frequency. This step removes such words to reduce noise. Lemmatization: to reduce words to their base or root form by removing suffixes. It is performed to reduce the number of unique words and group words with similar meanings together. These preprocessing steps are essential for cleaning and structuring the text data, making it more suitable for analysis and reducing the potential for errors during model evaluation. For example, a sentence contains Indonesian user on Twitter who posted, "Wamenkes Laporkan Temuan 2 Kasus Mutasi Covid-19 dari Inggris di Indonesia: Sebab itu, dia mengatakan, saat ini pandemi akan semakin berat. Sehingga saatnya untuk mengembangkan riset dan studi epideomiologi lebih tepat. https://t.co/HEY91MIET8". After preprocessing, the sentence becomes “wamenkes lapor temu kasus mutasi covid inggris indonesia sebab kata pandemi semakin berat saat kembang riset dan studi epideomiologi lebih tepat".

In the active learning model's initial iteration, 5% of training data is used for the training phase. At the same time, the rest will be available for selection by query strategies such as margin sampling and random sampling to improve the model's performance. This study has several batch sizes, indicating how many data samples are used in one iteration. This research uses a batch size of 15.

Dataset 1 - After preprocessing, there are 4,272 data samples, with 3,249 labeled as positive and 1,023 labeled as negative. The ratio of training data to testing data in this study is 70:30, with 2,990 samples for training and 1,282 samples for testing. A small subset of data comprising 149 samples, or 5% of the entire training data, is used for the initial iteration. In contrast, the remaining data is stored in a variable called "pool," which consists of 2,841 samples. Fig. 1 shows a graph displaying the accuracy results for each model. The logistic regression model achieved a final accuracy of 80.10% using margin sampling, while it obtained a final accuracy of 77.84% when using random sampling. The random forest model achieved a final accuracy of 81.12% using margin sampling and 78.70% using random sampling. The F1-score results are displayed in labels (c) and (d) for each model. The logistic regression model achieved a final F1-score of 88.12% using margin sampling and 86.87% using random sampling. The random forest model achieved a final F1-score of 88.60% using margin sampling and 87.43% using random sampling. The F1-score results are displayed in labels (c) and (d) for each model. The logistic regression model achieved a final F1-score of 88.12% using margin sampling and 86.87% using random sampling. The random forest model achieved a final F1-score of 88.60% using margin sampling and 87.43% using random sampling.

Dataset 2 - There are 5,055 data samples after preprocessing, with 2,638 labeled negative and 2,417 labeled positive. The ratio of training data to testing data in this study is 80:20, with 4,044 samples for training and 1,011 samples for testing. A subset of data comprising 202 samples, 5% of the actual training data, is also used for the initial iteration. At the same time, the remaining data is stored in a variable called "pool," consisting of 3,842 data samples. Based on Fig. 2, the accuracy results are shown in labels (a) and (b) for each model. The logistic regression model achieved a final accuracy of 74.87% using margin sampling, while it obtained a final accuracy of 72.79% when using random sampling. The random forest model achieved a final accuracy of 70.52% using margin sampling and 69.13% using random sampling. The F1-score results are displayed in labels (c) and (d) for each model. The logistic regression model achieved a final F1-score of 77.07% using margin sampling and 74.39% using random sampling. The random forest model achieved a final F1-score of 74.18% using margin sampling and 72.56% using random sampling.

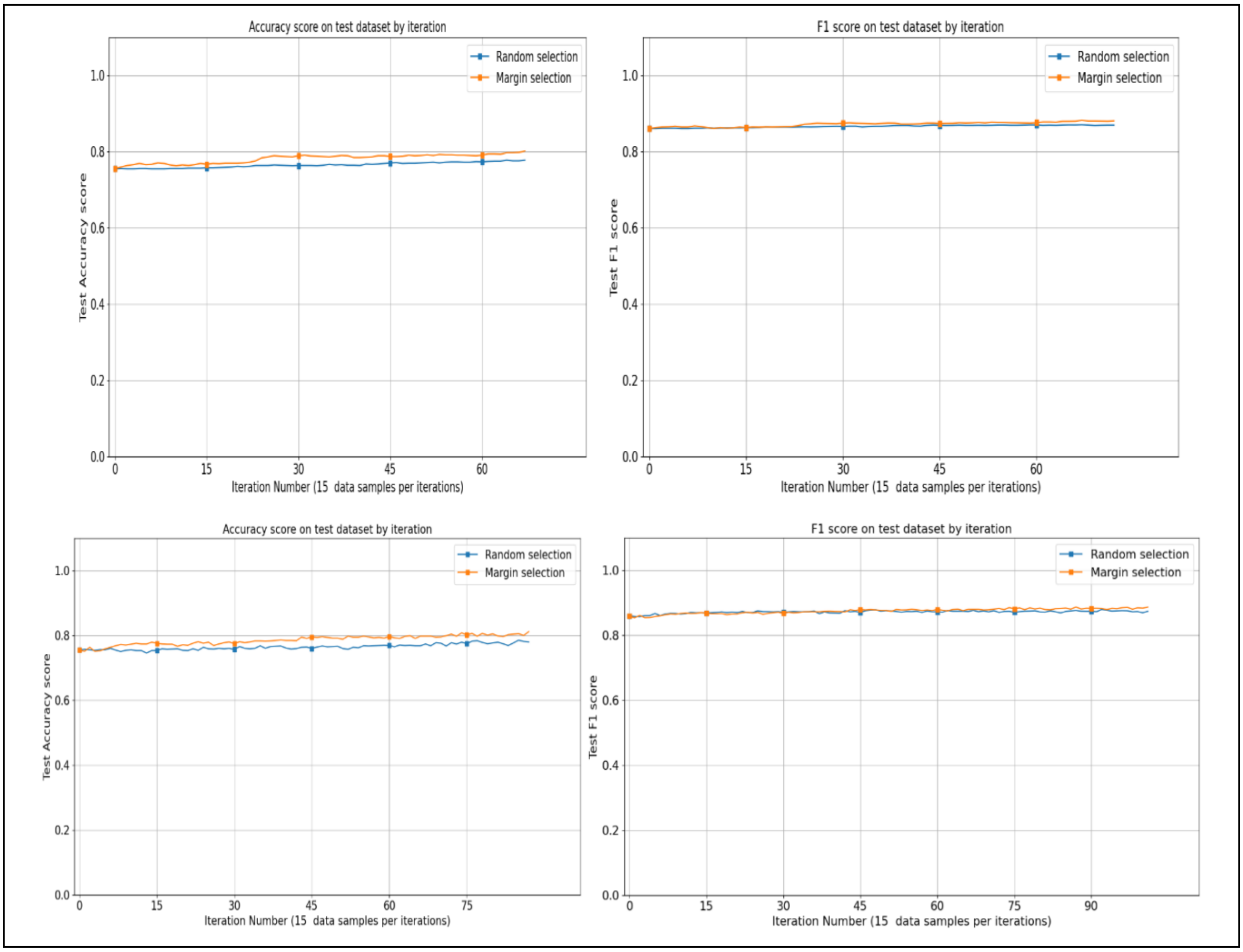

**Figure 1.** Accuracy and F1-score of logistic regression (top) and random forests (bottom) on Dataset 1 with a batch size of 15.

The evaluation of the two trained models aims to select the best-performing model based on the required performance during training. The results of the active learning implementation, including accuracy and F1-score, for the first and last iterations of each model using random sampling and margin sampling query strategies, are shown in Table 1. It can be observed that when using passive learning with the entire training data, the accuracy and F1-score values are lower compared to active learning. In active learning, the model is initially built using a small portion of the training data, specifically 5% of the training data, which amounts to 149 data points. Then, the active learning process is repeated until the final iteration.

For the logistic regression model, using 5% of the training data initially and conducting 67 iterations resulted in an accuracy of 35% of the pool set, totaling 1005 data points selected by the query strategy. Similarly, the F1-score for logistic regression, starting with the same 5% of the training data, reached its final iteration by including 38% of the pool set, corresponding to 1080 data points selected by the query strategy.

The random forest model selected 1305 data points from the pool set after 87 iterations, starting with 5% of the training data. Similarly, the F1-score for the random forest model selected 1515 data points from the pool set after 101 iterations, starting with the same 5% of the training data. Similarly, the F1-score for the random forest model selected 1515 data points from the pool set after 101 iterations, starting with the same 5% of the training data. The results indicate the advantages of active learning, where the model is trained on a smaller initial dataset but can achieve higher performance by iteratively selecting and incorporating informative data points from the pool set.

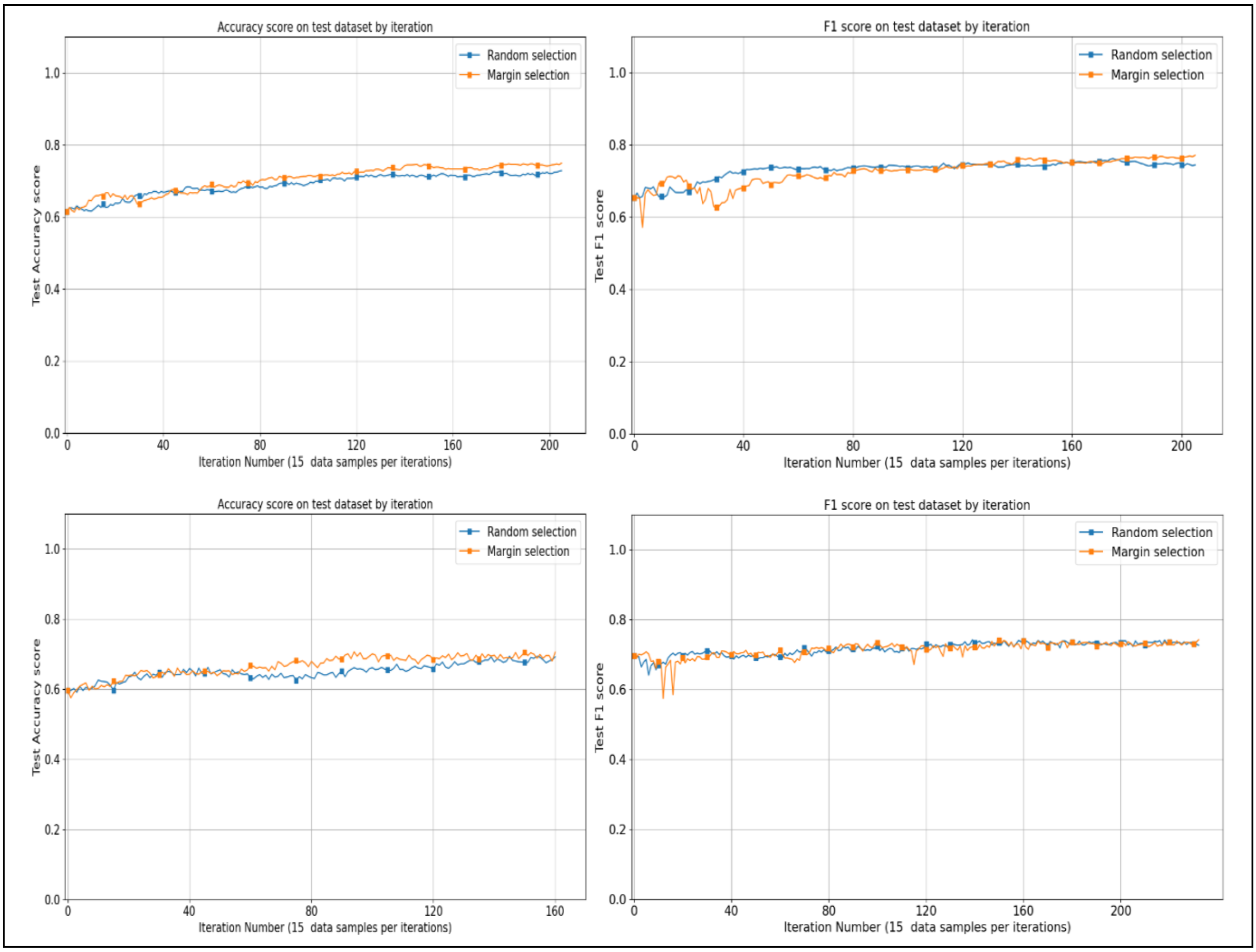

**Figure 2.** Accuracy and F1-score of logistic regression (top) and random forests (bottom) on Dataset 2 with batch size of 15.

**Table 1.** Model Evaluation on Dataset 1

| **Models** | **Metrics** | **Passing Learning (%)** | **Random Sampling** | | **Margin Sampling** | |
|---|---|---|---|---|---|---|
| | | | ***Initial (%)*** | ***End (%)*** | ***Initial (%)*** | ***End (%)*** |
| **Logistic Regression** | Accuracy | 78.70 | 75.43 | 77.84 | 75.43 | 80.10 |
| | F1-score | 87.50 | 85.99 | 86.87 | 85.99 | 88.02 |
| **Random Forests** | Accuracy | 79.32 | 75.43 | 78.70 | 75.43 | 81.12 |
| | F1-score | 87.84 | 85.89 | 87.43 | 85.89 | 88.60 |

**Table 2.** Model Evaluation on Dataset 2

| **Models** | **Metrics** | **Passing Learning (%)** | **Random Sampling** | | **Margin Sampling** | |
|---|---|---|---|---|---|---|
| | | | ***Initial (%)*** | ***End (%)*** | ***Initial (%)*** | ***End (%)*** |
| **Logistic Regression** | Accuracy | 72.40 | 61.42 | 72.79 | 61.42 | 74.87 |
| | F1-score | 75.15 | 65.36 | 74.39 | 65.36 | 77.07 |
| **Random Forests** | Accuracy | 67.16 | 59.54 | 69.13 | 59.54 | 70.52 |
| | F1-score | 72.65 | 69.59 | 72.56 | 69.59 | 74.18 |

Based on Table 2, it can be observed that when using passive learning with the entire training data, the accuracy and F1-score values are lower than active learning. In active learning, the model is initially built using a small portion of the training data, specifically 5% of the training data, which amounts to 202 data points. Then, the active learning process is repeated until the final iteration.

The logistic regression model, starting with 5% of the training data, reached its final iteration by including 80% of the pool set after 205 iterations, totaling 3075 data points selected by the query strategy. Similarly, the F1-score for logistic regression, starting with the same 5% of the training data, reached its final iteration by including 80% of the pool set after 205 iterations, resulting in 3075 data points selected by the query strategy.

The random forest model reached its final iteration after 160 iterations, including 62% of the pool set, which consisted of 2400 data points selected by the query strategy. Similarly, the F1-score for the random forest model reached its final iteration after 232 iterations, including 90% of the pool set, which consisted of 3480 data points selected by the query strategy, starting with the same 5% of the training data. Similarly, the F1-score for the random forest model reached its final iteration after 232 iterations, including 90% of the pool set, which consisted of 3480 data points selected by the query strategy, starting with the same 5% of the training data.

## 4. Conclusion

Active learning can reduce the training data used in model training. Despite reducing the amount of training data, active learning can enhance model performance. In Dataset 1, active learning can save 45% to 65% of the total training data, meaning an average of only 1323 data points are needed to train each model. In Dataset 2, active learning can save 10% to 44% of the total training data, meaning an average of only 2810 data points are needed to train each model. This study achieved the best performance results using a batch size of 15. Therefore, the more data used in a single iteration of the active learning query strategy, the better the performance obtained.

The margin sampling can improve the classification performance of logistic regression and random forest models for both datasets in sentiment analysis on Twitter data. It achieved higher performance compared to models using random sampling query strategies. Accuracy values improved performance by about 2% to 4%, and F1-score values also increased by approximately 2% to 4% compared to models using random sampling query strategies. In Dataset 1, the best model obtained was the random forest classification model with margin sampling query strategy, achieving the highest accuracy and F1-score values compared to the logistic regression model. In Dataset 1, using an average of 1323 data points, the random forest model achieved an accuracy of 81.12% and an F1-score of 88.60%. In Dataset 2, the best model obtained was the logistic regression classification model with the margin sampling query strategy, achieving the highest accuracy and F1-score values compared to the random forest model. In Dataset 2, using an average of 2810 data points, the logistic regression model achieved an accuracy of 74.87% and an F1-score of 77.07%.

This paper demonstrates the combination of active learning strategies and machine learning algorithms for sentiment analysis task with two classes. This can extend to multiclass classification problems in the future. The investigation of the potential overfitting and underfitting during training phase is also crucial topics which is not cover yet. In addition, various active learning query strategies can be explored, such as other variants of uncertainty sampling, like least confident and entropy sampling, or different query strategies.

## 5. Declarations

### 5.1. Author Contributions

Conceptualization: M.L., N.I., and E.C.; Methodology: T.E.S.; Software: M.L.; Validation: M.L., N.I., and E.C.; Formal Analysis: M.L., N.I., and E.C.; Investigation: N.I.; Resources: S.I.; Data Curation: N.I.; Writing Original Draft Preparation: N.I., T.E.S., and M.L.; Writing Review and Editing: N.I. and T.E.S.; Visualization: S.I.; All authors have read and agreed to the published version of the manuscript.

### 5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 5.3. Funding

This work is supported by the research grant with Contract Number UN.01/PA/223/2022 and Registration ID. 221200000050368, provided by UIN Syarif Hidayatullah, Jakarta, Indonesia. The authors acknowledge the financial support from UIN Syarif Hidayatullah for the research, authorship, and/or publication of this article.

### 5.4. Institutional Review Board Statement

Not applicable.

### 5.5. Informed Consent Statement

Not applicable.

### 5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This work is supported by research grant number UN.01/PA/223/2022, Registration ID. 221200000050368 from UIN Syarif Hidayatullah, Jakarta, Indonesia.

## References

[1] B. Arafah et al., “The Digital Culture Literacy of Generation Z Netizens as Readers, Producers and Publishers of Text on Social Media,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 3, pp. 112–123, Jul. 2023.

[2] D. Elangovan and V. Subedha, “Adaptive Particle Grey Wolf optimizer with deep learning-based sentiment analysis on online product reviews,” *Engineering, Technology &amp; Applied Science Research*, vol. 13, no. 3, pp. 10989–10993, 2023. doi:10.48084/etasr.5787.

[3] S. Malviya, A. K. Tiwari, R. Srivastava, and V. Tiwari, “Machine Learning Techniques for Sentiment Analysis: A Review,” *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, vol. 12, no. 02, pp. 72–78, Dec. 2020, doi: 10.18090/samriddhi.v12i02.03.

[4] [1] P. Zhang, T. Chai, and Y. Xu, "Adaptive prompt learning-based few-shot sentiment analysis," *Neural Processing Letters*, vol. 55, no. 6, pp. 7259–7272, 2023. doi:10.1007/s11063-023-11259-4.

[5] P. Kumar and A. Gupta, "Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey," *J. Comput. Sci. Technol.*, vol. 35, no. 4, pp. 913–945, Jul. 2020, doi: 10.1007/s11390-020-9487-4.

[6] P. Ren et al., "A Survey of Deep Active Learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 180:1-180:40, Oct. 2021, doi: 10.1145/3472291.

[7] M. H. Jarrahi, A. Memariani, and S. Guha, "The Principles of Data-Centric AI," *Commun. ACM*, vol. 66, no. 8, pp. 84–92, Jul. 2023, doi: 10.1145/3571724.

[8] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," *Artif Intell Rev*, vol. 56, no. 4, pp. 3005–3054, Apr. 2023, doi: 10.1007/s10462-022-10246-w.

[9] A. Raj and F. Bach, "Convergence of Uncertainty Sampling for Active Learning," *in Proceedings of the 39th International Conference on Machine Learning, PMLR*, vol. 1, no. 1, pp. 18310–18331, 2022. Accessed: Dec. 17, 2023. [Online]. Available: https://proceedings.mlr.press/v162/raj22a.html

[10] J. Shao, Q. Wang, and F. Liu, "Learning to Sample: An Active Learning Framework," in *2019 IEEE International Conference on Data Mining (ICDM)*, vol. 1, no. 1, pp. 538–547, 2019. doi: 10.1109/ICDM.2019.00064.

[11] V.-L. Nguyen, M. H. Shaker, and E. Hüllermeier, "How to measure uncertainty in uncertainty sampling for active learning," *Mach Learn*, vol. 111, no. 1, pp. 89–122, Jan. 2022, doi: 10.1007/s10994-021-06003-9.

[12] U. Naseem, M. Khushi, S. K. Khan, K. Shaukat, and M. A. Moni, "A Comparative Analysis of Active Learning for Biomedical Text Mining," *Applied System Innovation*, vol. 4, no. 1, pp. 1-18, Mar. 2021, doi: 10.3390/asi4010023.

[13] A. Agrawal and S. Tripathi, "Active learning using margin sampling strategy for entity recognition," *Lecture Notes in Electrical Engineering*, vol. 1, no. 1, pp. 163–169, 2020. doi:10.1007/978-981-15-3125-5_18.

[14] J. Han, M. Kamber, and J. Pei, "*Data Mining: Concepts and Techniques*," The Morgan Kaufmann Series in Data Management Systems., Boston: Morgan Kaufmann, 2011, pp. 327–391. doi: 10.1016/B978-0-12-381479-1.00008-3.

[15] B. Settles, "*Active Learning*," Synthesis Lectures on Artificial Intelligence and Machine Learning, Cham: Springer International Publishing, 2012, pp. 11–20. doi: 10.1007/978-3-031-01560-1_2.

[16] R. Monarch and R. Munro, Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI. Simon and Schuster, 2021.

[17] A. Sauer, R. B. Gramacy, and D. Higdon, "Active learning for Deep Gaussian process surrogates," *Technometrics*, vol. 65, no. 1, pp. 4–18, 2022. doi:10.1080/00401706.2021.2008505

[18] W. Gao and C. Wang, "Active learning based sampling for high-dimensional nonlinear partial differential equations," *Journal of Computational Physics*, vol. 475, no. 1, pp. 111848–111862, 2023. doi:10.1016/j.jcp.2022.111848

[19] I. L. Rahayu, "Identifikasi aktor terpenting penyebaran Informasi covid-19 berdasarkan komentar netizen di twitter menggunakan metode Katz centrality," bachelor Thesis, Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta, 2021. Accessed: Sep. 15, 2023. [Online]. Available: https://repository.uinjkt.ac.id/dspace/handle/123456789/56882

[20] Ridi Ferdiana, Fahim Jatmiko, Desi Dwi Purwanti, Artmita Sekar Tri Ayu, and Wiliam Fajar Dicka, "Dataset Indonesia untuk Analisis Sentimen", *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 8, no. 4, pp. 334-339, Nov. 2019.