JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 4 No. 1 (2023)
https://doi.org/10.35877/454RI.jinav1518

Retweet Predictions Regarding COVID-19 Vaccination Tweets Through The Method Of Multi Level Stacking

Vena Erla Candrika, Jondri*, & Indwiarti

Faculty of Informatics, Telkom University, Indonesia

Abstract

The rapid development of technology from day to day indirectly influences increasing social media use. This can be seen from spreading information that is very easily found on social media, one of which is *Twitter*. It is one of the most popular platforms for expressing people's feelings by tweeting and interacting with other users at the same time. Various opinions about the COVID-19 vaccination began to be discussed on the Twitter platform. Moreover, most people take advantage of the feature available on Twitter, namely retweets. Users do retweet because there are many influencing factors. It can be caused by a reason that they have the same opinions and thoughts as the tweet owner, and so on. A retweet feature is also a form of information diffusion on the Twitter platform. The diffusion of information on Twitter has several factors, such as the most influential users, using hashtags or URLs, and others. In this conclusion, retweet predictions have been carried out regarding COVID-19 vaccination tweets using the features user-based and time-based through the Multi-Level Stacking classification method. This method indicates the best results when oversampling with an F1-Score of 96.23%.

Keywords: Multi Level Stacking, Twitter, Retweet, COVID-19, Vaccination, Prediction

1. Introduction

1.1. Background

Social media has become a part of people's lives, along with the development of the age, social media continues to evolve. Its existence makes a lot of information could be easy to get for its users. It has been almost two years since the official announcement of the first case of COVID-19 with 2 people confirmed positive in Indonesia on March 2, 2020 by President Joko Widodo (Rahma & Fadhilia Arvianti, 2020; Rantauni & Sukmawati, 2022). It caused a lot of public information and opinions in the form of pros and cons posted on social media, one of which is Twitter. Various efforts have been made by the government to deal with the pandemic situation that is currently happening, starting from lockdowns, vaccine administration, etc (Pandya & Lodha, 2021; Runacres et al., 2021; WHO, 2020).

Information diffusion or dissemination of information is one of the dynamic processes that are widely studied on a network (Firdaniza et al., 2022; Yujie, 2020). The rapid information dissemination in the community is easy to accept information quickly without knowing the truth which can affect policies in social media. In Twitter, there is information that users can see either in the form of tweets or writing that can be inserted with images, URLs, hastags, and so on (Li et al., 2017). The process of disseminating information on Twitter is highly dependent on the numbers of followers, tweets liked, retweets, and the age of the user's account. There is a retweet feature that can be used by the users which causes information to spread more widely. The information diffusion on Twitter is also influenced by several features, namely user-based, time-based, and content-based (Hoang & Mothe, 2018; Molaei et al., 2020). In research that has been conducted by Adrian, previously using [2]multi-class classification which could get an accuracy value and an f1-score of less than 80%. However, with the use of deep learning algorithms, it is able to make model training that is built even better (Adrian et al., 2021).

* Corresponding author.
E-mail address: jondri@telkomuniversity.ac.id

According to the previous Adrian research, this research used the Multi-Level Stacking method to predict retweets information related to COVID-19 vaccination tweets in Indonesia on Twitter social media using user-based and time-based features (Habibi & Cahyo, 2021). Through combining the Decision Tree, KNN, and Naive Bayes algorithms in the Multi Level Stacking method, it is expected to produce good performance.

1.2. Topics and Limitations

Based on the background that has been previously conveyed, the formulation of the problem in this research is to predict tweets related to COVID-19 vaccination that are then retweeted based on user-based and time-based features and analyze the Multi-Level Stacking method in making predictions. The limitations of the problem aims to make the thesis research carried out specifically, there are several limitations, namely data taken from Twitter regarding COVID-19 vaccination tweets and tweet data from Twitter using Indonesian.

1.3. Objective

The purpose of this research is to create a retweet prediction system related to COVID-19 vaccination tweets based on the features of user-based and time-based as well as implementing and analyzing the Multi-Level Stacking method in making predictions.

2. Literature Review

In this section, several references have been taken as supporters of the background raised and connected to this research. Here is the research taken as the references:

A study conducted in 2018 entitled "*Retweet Predictive Model in Twitter"* aimed to predict the popularity of tweets within a range. The study also contained innovative retweet predictions using [4]*a machine learning* approach through the features (Joukov Costa De Oliveira *et al.*, 2018). This research used the features of user, tweet and extra tweet. The user features used the numbers of followers, favorite tweets, registered users, the age of account ,verified users, and statuses (Joukov Costa De Oliveira *et al.*, 2018). The tweet features used hashtags, URLs, mentions, text length and word count, reply, tweet time, and visual content (video, photo, gif). Meanwhile, the extra tweet features here used sentiment features which could be classified as positive sentiment score, negative sentiment score, and neutral sentiment score (Joukov Costa De Oliveira *et al.*, 2018).

Research conducted by Hoang, T., & Mothe, J in 2018 which predicted a tweet is whether retweeted or not (binary classification) by modeling and predicting new tweet propagation rates (multi-class classification). The research by (Hoang & Mothe, 2018) was conducted using user-based, time-based, and content-based features totaling 29 features. The results of the research obtained for both binary and multi-class prediction types, the performance of the F-measure with the built model increased by 5% of its state of art.

Unlike the previous research, a study conducted by Adrian *et al.*, that compared the Support Vector Machine algorithm with Random Forest with multi-class classification used tweet data related to *PSBB* which amounted to 466 tweet data (Adrian *et al.*, 2021). The study showed an accuracy value of 58% for the Random Forest algorithm and 56% for the Support Vector Machine algorithm. This result indicated the accuracy value was small because it used only few data (Adrian *et al.*, 2021) .

Stacking is a common procedure where learners are trained to combine individual learners which are called as first-lever learners (Fitriansyah & Saparudin, 2016), meanwhile, those who combine are called second-level learners or meta-learners (Fitriansyah & Saparudin, 2016). Stacking process is simply a dataset learned by first-level learners that produces a new dataset that will be used as input for the second-level learner (Fitriansyah & Saparudin, 2016).

2.1. Twitter

Twitter is a popular micro-blogging website that offers the users to send and interact with short messages. Twitter was also chosen as a data source of much research due to its popularity, ease of use to express opinions, and use of its tweets up to 280 characters (Bouazizi & Ohtsuki, 2019; Mohbey, 2020; Tane *et al.*, 2019)

2.2. Features

The feature in this study is the most important aspect to be analyzed. Two features were applied, namely user-based and time-based features.

The user-based features involved the numbers of (Hoang & Mothe, 2018):

- User's past tweets
- Followers accounts
- Following accounts
- User account age
- Tweets liked

The time-based features involved a post at (Hoang & Mothe, 2018):

- Holiday
- Noon
- Night
- Weekend

2.3. Multi Level Stacking

As the name implies, multi-levels stacking consists of many layers: two, three, or more. For example, you want to build a stacking model with three levels. At the first level (layer), a number of basic L models have been selected. In the second level, a number of M meta-models are trained using inputs in the form of predictions produced by the basic L model at the first level. Finally, in the third level, one last meta-model is trained using inputs in the form of predictions generated by the M meta-model of the second level (DR. SUYANTO *et al.*, 2020).

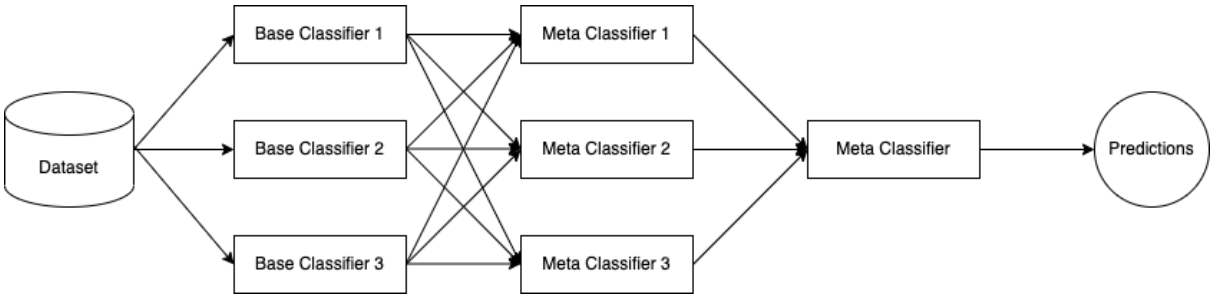

Figure 1. Multi Level Stacking Model

From an illustration above, it can be seen that the multi-levels stacking model contains training a number of M meta-models to produce output based on the output of an L number basic models or the weak learners.

The study entitled "Internet of Energy: Ensemble Learning trough Multilevel Stacking for Load Forecasting" mentione that the multi-levels stacking classification used 4 layers. Each layer used a different algorithm with its selection based on the performance of the model at the previous level. In the selection of algorithms implemented in the study, several algorithms were compared to produce values from error accuracy. If the error accuracy value was getting smaller, it was inserted on the next layer until which algorithm that will be superior (Singh *et al.*, 2020).

2.4. K-Fold Cross Validation

Cross-validation is a technique primarily used to predict a model and estimate how accurate a predictive model is when run in practice. One technique of cross-validation is *k-fold cross validation* which distributing data into k parts of datasets of the same size (Tempola *et al.*, 2018).

Figure 2. K-Fold Illustration

The figure above is an illustration of data sharing with $k = 4$. Where at the time of the first iteration (k=1), subset 1 is used as test data, while the other 3 subsets are used as train data. It will repeat according to the value of k (Purnajaya & Kusuma, 2019).

3. Methods

Analysis related to retweet predictions regarding COVID-19 vaccination tweets was carried out using several features accompanied by multi-level stacking classifications. The following are the stages or descriptions of the system carried out to support this research. User tweet data are obtained from the netlytic.

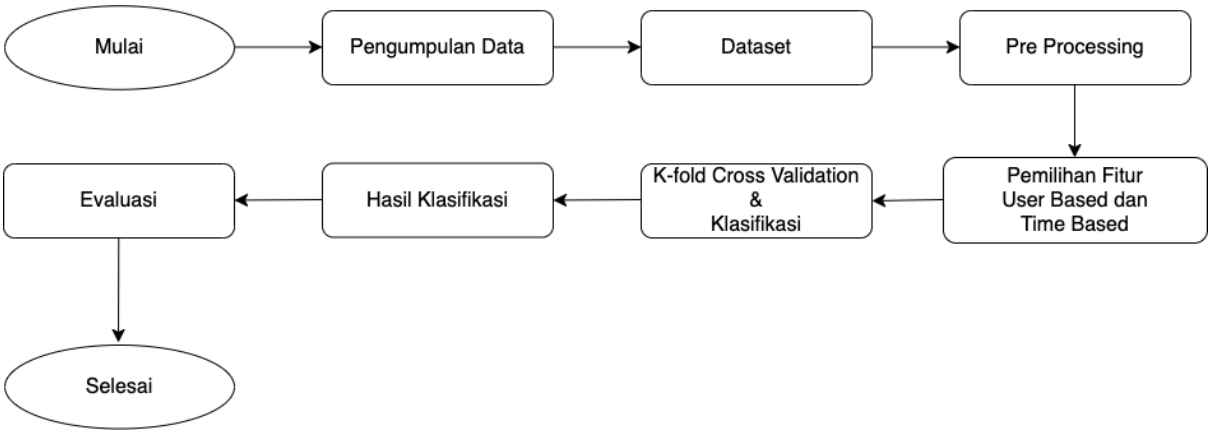

Figure 3. Flow Chart System Overview

3.1. Data Crawling

Data crawling is the retrieval of data from a data source. In this study, data collection was carried out from Twitter social media. The use of API that has been provided by the platform, namely the Twitter API Key (Firdaus *et al.*, 2018). Twitter data were taken using netlytic and obtained 11,718 tweets with the keyword of Indonesian COVID-19 Vaccine during September - December 2022.

3.2. Pre-Processing

1) Data normalization used *StandardScaler* in features in the form of numeric data types was scaled to speed up the classification process.
2) Labeling was done to create features that were not in the dataset, for example, the feature in the "retweet_count" to determine whether the tweet was retweeted or not.

3.3. Feature Selection

At this stage, the removal of features that are not related to the classification process was carried out, so that there were 16 features selected for use as follows:

Table 1. Features Selection

Feature	Data Type
Source	int64
favorite_count	float64
retweet_count	float64
Lang	int64
tweet_type	int64
in_reply_to_user_id	float64
in_reply_to_status_id	float64
retweeted_user_id	float64
retweeted_status_id	float64
user_statuses_count	float64
user_friends_count	float64
user_followers_count	float64
Hour	int64
Day	int64
Month	int64
Year	int64

3.4. K-Fold Cross Validation

At this stage, the use of K-Fold Cross Validation 11,718 data were divided by the number of *predetermined* of *k* values, namely k = 3, k = 5, and k = 10.

3.5. Multi Level Stacking Classification

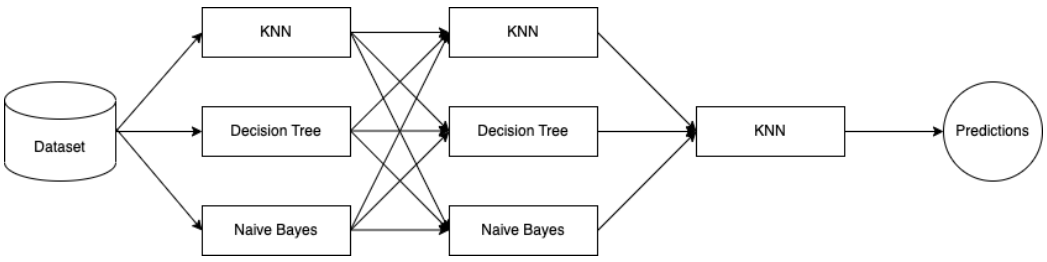

Figure 4. Multi Level Stacking Illustration

In the picture above, there are three layers Multi Level Stacking method. Layer 1 or Base Learner has three classification algorithms, namely: *Decision Tree*, *KNN*, and *Naive Bayes.* In Layer 2 or *Meta Learner*, the similar algorithm is used in Layer 1. Furthermore, Layer 3 is used as KNN algorithm. Each layer is trained with the prediction results of the previous layers.

3.6. Test Scenarios

The dataset obtained amounted to 11.718 tweets and one of them contained a class of 'retweet_count'. Its class is very influential in research, so a visualization is created to find out the distribution of the class. It can be seen in the picture below, that the dataset obtained is too many tweets that are not retweeted (the frequency at 0 value is very high) so that this can affect the results of the study later. This is usually referred to as the imbalanced class state.

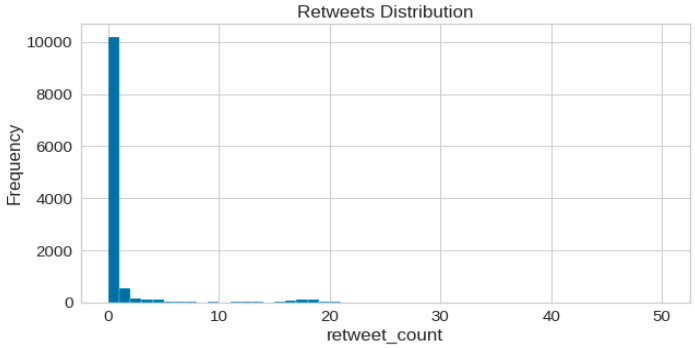

Figure 5. retweet_count

- Test Scenario 1

In test 1, the dataset used is a dataset that is still in an imbalanced class state.

- Test Scenario 2

In test 2, the dataset used is a dataset that RandomUnderSampler has performed to overcome the imbalanced class.

- Test Scenario 3

In test 3, the dataset used is a dataset that has been oversampling. This was done because the distribution of data in retweet_count class is imbalanced.

4. Result and Discussion

This section is an exposure and analysis of test results. It was done in line with the objective of the research that has been stated in the previous section.

4.1. Test Results

In the first test, the dataset used was an imbalanced class so that undersampling and oversampling had not been carried out. The F1-Score result obtained by the Multi Level Stacking method is 75.47%

Table 2. Test Results 1

Method	Precision	Recall	F1-Score
KNN	80.97%	70.67%	75.47%
Decision Tree	82.35%	89.04%	85.57%
Naïve Bayes	74.11%	51.59%	60.83%
Stacking	80.97%	70.67%	75.47%

In the second test, the results obtained to predict retweets used the Multi Level Stacking method of *RandomUnderSampler* that showed the highest F1-Score when K-Fold Cross Validation was set with a value of *k=10* with a result of 83.61%.

Table 3. Test Results 2

Number of Folds	Precision	Recall	F1-Score
K=3	82.74%	82.74%	82.74%
K=5	83.40%	83.40%	83.40%
K=10	83.61%	83.61%	83.61%

In the third test, the results obtained to predict retweets using the Multi Level Stacking method used *Oversampling* that showed the highest F1-Score, precision, and recall when K-Fold Cross Validation was set with a value of *k=10* with a result of 96.23%.

Table 3. Test Results 3

Number of Folds	Precision	Recall	F1-Score
K=3	94.48%	94.48%%	94.48%
K=5	95.34%	95.34%	95.34%
K=10	96.23%	96.23%	96.23%

5. Conclusion

In the research that has been carried out, it can be concluded that in the prediction of retweets regarding COVID-19 vaccination tweets using the Multi Level Stacking method obtained an increase in performance when oversampling has been carried out. Through this oversampling, the Multi Level Stacking method succeeded in increasing the F1-Score value with a result of 96.23% using K-fold Cross Validation.

For further research, this Multi Level Stacking method can be further developed with a combination of other classification algorithms and add the number of meta-learners. In addition, the author also suggests to explore this Multi Level Stacking method with other features.

References

Adrian, M. R., Putra, M. P., Rafialdy, M. H., & Rakhmawati, N. A. (2021). Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB. *Jurnal Informatika Upgris*, *7*(1). https://doi.org/10.26877/jiu.v7i1.7099

Bouazizi, M., & Ohtsuki, T. (2019). Multi-class sentiment analysis on twitter: Classification performance and challenges. *Big Data Mining and Analytics*, *2*(3), 181–194. https://doi.org/10.26599/BDMA.2019.9020002

DR. SUYANTO, S. T. . M. S., ANDITYA ARIFANTO, S. T. . M. T., RITA RISMALA, S. T. . M. T., & DR. ANDI SUNYOTO., M. K. (2020). *Evolutionary Machine Learning - Pembelajaran Mesin Otonom Berbasis Komputasi Evolusioner*. 1–467.

Firdaniza, F., Ruchjana, B. N., Chaerani, D., & Radianti, J. (2022). Information Diffusion Model in Twitter: A Systematic Literature Review. *Information (Switzerland)*, *13*(1). https://doi.org/10.3390/info13010013

Firdaus, S. N., Ding, C., & Sadeghian, A. (2018). Retweet: A popular information diffusion mechanism – A survey paper. *Online Social Networks and Media*, *6*, 26–40. https://doi.org/10.1016/j.osnem.2018.04.001

Fitriansyah, R. A., & Saparudin. (2016). *Penerapan Ensemble Stacking Untuk Klasifikasi Multi Kelas*. *2*(1), 240–243.

Habibi, M., & Cahyo, P. W. (2021). A Social Network Analysis: Identifying Influencers in The COVID-19 Vaccination Discussion on Twitter. *Compiler*, *10*(2). https://doi.org/10.28989/compiler.v10i2.1074

Hoang, T. B. N., & Mothe, J. (2018). Predicting information diffusion on Twitter – Analysis of predictive features. *Journal of Computational Science*, *28*, 257–264. https://doi.org/10.1016/j.jocs.2017.10.010

Joukov Costa De Oliveira, N., Prof, A., Ribeiro, B., Prof, J., Rupino Da Cunha, P., & Abreu, P. (2018). *Retweet Predictive Model in Twitter*.

Li, M., Wang, X., Gao, K., & Zhang, S. (2017). A survey on information diffusion in online social networks: Models

and methods. In *Information (Switzerland)* (Vol. 8, Issue 4). https://doi.org/10.3390/info8040118

Mohbey, K. K. (2020). Multi-class approach for user behavior prediction using deep learning framework on twitter election dataset. *Journal of Data, Information and Management*, *2*(1), 1–14. https://doi.org/10.1007/s42488-019-00013-y

Molaei, S., Zare, H., & Veisi, H. (2020). Deep learning approach on information diffusion in heterogeneous networks. *Knowledge-Based Systems*, *189*. https://doi.org/10.1016/j.knosys.2019.105153

Pandya, A., & Lodha, P. (2021). Social Connectedness, Excessive Screen Time During COVID-19 and Mental Health: A Review of Current Evidence. *Frontiers in Human Dynamics*, *3*. https://doi.org/10.3389/fhumd.2021.684137

Purnajaya, A. R., & Kusuma, W. A. (2019). *Prediksi Interaksi pada Jejaring Bipartite Senyawa dan Protein pada Data yang Tidak Seimbang*. *January*, 1–41.

Rahma, V. S., & Fadhilia Arvianti, G. (2020). THE IMPACTS OF COVID-19 PANDEMIC IN INDONESIA AND CHINA'S HOTEL INDUSTRY: HOW TO OVERCOME IT? *JELAJAH: Journal of Tourism and Hospitality*, *2*(1). https://doi.org/10.33830/jelajah.v2i1.864

Rantauni, D. A., & Sukmawati, E. (2022). Correlation of Knowledge and Compliance of Implementing 5m Health Protocols in the Post-Covid-19 Pandemic Period. In *Science Midwifery* (Vol. 10, Issue 4). Online. www.midwifery.iocspublisher.orgjournalhomepage:www.midwifery.iocspublisher.org

Runacres, A., Mackintosh, K. A., Knight, R. L., Sheeran, L., Thatcher, R., Shelley, J., & McNarry, M. A. (2021). Impact of the covid-19 pandemic on sedentary time and behaviour in children and adults: A systematic review and meta-analysis. *International Journal of Environmental Research and Public Health*, *18*(21). https://doi.org/10.3390/ijerph182111286

Singh, S., Yassine, A., & Benlamri, R. (2020). Internet of Energy: Ensemble Learning through Multilevel Stacking for Load Forecasting. *Proceedings - IEEE 18th International Conference on Dependable, Autonomic and Secure Computing, IEEE 18th International Conference on Pervasive Intelligence and Computing, IEEE 6th International Conference on Cloud and Big Data Computing and IEEE 5th Cybe*, 658–664. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00113

Tane, O. Z. A., Lhaksmana, K. M., & Nhita, F. (2019). Analisis Sentimen pada Twitter Tentang Calon Presiden 2019 Menggunakan Metode SVM (Support Vector Machine). *Seminar Nasional Teknologi Fakultas Teknik Universitas Krisnadwipayana*, *1*(1), 739–742.

Tempola, F., Muhammad, M., & Khairan, A. (2018). Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(5), 577. https://doi.org/10.25126/jtiik.201855983

WHO. (2020). WHO and UNICEF warn of a decline in vaccinations during COVID-19. In *Who* (Vol. 41, Issue 8).

Yujie, Y. (2020). A Survey on Information Diffusion in Online Social Networks. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3393822.3432322