Classification of Tweets for Video Streaming Services' Content Recommendation on Twitter

Kiki Ferawati[1] and Sa'idah Zahrotul Jannah[2]
[1]Statistics Study Program, Universitas Sebelas Maret
[2]Statistics Study Program, Universitas Airlangga

kferawati@staff.uns.ac.id, saidahzj@stmm.unair.ac.id

Abstract. Streaming services were popular platforms often visited by internet users. However, the abundance of content can be confusing for its users, prompting them to look for a recommendation from other people. Some of the users looked for content to enjoy with the help of Twitter. However, there were irrelevant tweets shown in the results, showing sentences not related at all to the content in the streaming services platform. This study addressed the classification of relevant and irrelevant tweets for streaming services' content recommendation using random forests and the Convolutional Neural Network (CNN). The result showed that the CNN performed better in the test set with higher accuracy of 94% but slower in running time compared to the random forest. There were indeed distinctive characteristics between the two categories of the tweets. Finally, based on the resulting classification, users could identify the right words to use and avoid while searching on Twitter.

Keywords: text mining, streaming services, classification, random forest, CNN

1. Introduction

Streaming services were on the rise recently. There were 59% of internet users aged 16 to 64 owning a technology device watching television content via streaming services platform each month [1]. As of September 2020, eleven video streaming services were operating in Indonesia [2]. Two of the most popular streaming services in Indonesia are Netflix and Disney+, offering various content such as movies, series, and animation. However, due to the vast collection offered on the platform, some subscribers cannot decide on what kind of content that they want to enjoy. This leads to them checking out the recommendation from their colleagues, and some even ended up browsing for a recommendation from a stranger on social media, for example, Twitter. In Indonesia, Twitter is one of the most visited platforms, amassing more than 90 million monthly traffic with 56% of internet users actively using the platform for social media activity [1].

While looking for a recommendation, people usually wrote the name of the platform and the type of content that they want to know in the search box on Twitter. They would type some keywords like "Netflix movie", "Netflix series", "Disney movie recommendation", etc., and get the tweet as a result containing the title of movies that is usually popular in their region. However, the result often showed nonrelevant things such as tweets about offering the streaming services platform. It encouraged the users to

subscribe to a certain platform but sometimes people got uncomfortable since they expected to get relevant things (synopsis of a movie or movie title recommendation).

There are various reasons for people subscribing to a certain streaming service, such as option availability, social trends, and subscription fee [3]. As there are not many Indonesians having suitable methods for paying the subscription, there has been a rise in Twitter accounts offering to help Twitter users to subscribe to the services. However, the rise of these kinds of accounts sometimes was distracting for people wanting to find a content recommendation on Twitter, as there were many tweets with the keywords found but there were many nonrelevant tweets instead.

This study aims to identify the tweets relevant for finding content recommendations respective to the streaming services of Disney and Netflix on Twitter. The tweets will then be classified into relevant and nonrelevant tweets. There are several methods often used in classification problems, with random forest and convolutional neural network (CNN) being the most popular methods of all. Random forest is an ensemble comprising of multiple decision trees, while CNN is the improved version of the neural network algorithm. Previously, the random forest is used in the research for movie sales prediction in Korea, resulting in the analysis of the related factors to the success of the movie [4]. Another research showed that random forest is better than naïve Bayes for classifying sentiment analysis of movie recommendations for users [5]. Although CNN is generally more popular for deep learning, the method showed better results for sentiment analysis compared to Backpropagation Neural Network (BNN) for classifying sentiment in Twitter of the government of Surabaya [6].

In this study, both random forests and CNN were used for classifying the relevant and irrelevant tweets for streaming services' content recommendation on Twitter. There was expected to exist a clear distinction between tweets from the two categories. The characteristics of the tweets will be useful for future references in Twitter searches.

2. Literature review

2.1. Streaming service provider. Streaming is a process of listening or watching sound or video from the internet without the need to download the content [7]. The streaming service provider is a system offering online streaming access, usually with a subscription, to content such as movies, series, or animations. The subscriber can play the contents via their media players, such as a computer, phone, or smart TV.

2.2. Text mining. Text mining is a process to extract information from unstructured data in documents. It is essentially similar to data mining, with the main difference being in preprocessing stage, where a transformation from unstructured data to a more familiar format is needed [8]. The general steps of preprocessing: removing lines, links, numbers, username, punctuations, and stopwords, case folding and tokenizing. Extracted words are then weighted using Term Frequency-Inverse Document Frequency (TF-IDF) weighting. The formula for TF-IDF is described in equation (1) to (3) [9].

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

$$TF(t,d) = \frac{n_t}{n} \quad (2)$$

$$IDF(t) = \frac{N_d}{N} \quad (3)$$

where *t, d, n,* and *N* are terms, document, number of terms, and number of documents, respectively.

2.3. Random forest. Random forests are methods for supervised learning, constructed from ensemble methods of decision trees [10]. The trees in random forests are grown with random inputs and features. It is effective for the problem of prediction and is effective to improve the accuracy of the model. For the problem of classification, the algorithm of random forests is as follows.

1. Repeat the following steps *K* times:
   a. Drawing a bootstrap sample of *n* from *N* training data.
   b. Train a random forest for the bootstrapped data with a random feature $F = \log_2 M + 1$ where *M* is the number of features.
   c. Predict the test set based on the trees in the previous step.
2. Predicting the final dataset by combining the result of classification using majority vote.

2.4. Convolutional neural network (CNN). CNN is a method commonly used for image analysis. It has advantages over the other neural network-based method in transforming the data into an easier processed input. In the text mining problem, the input is in the form of a matrix from a sentence [11]. Generally, the process of CNN is feature extraction and classification process. Feature extraction consists of transforming the complex to a more simplified input. There are layers for convolution and layers for pooling for reducing the dimension of the parameters. Several popular activation functions including sigmoid functions, rectified linear units (RelU), and parametric RelU.

In the classification process, it processed the input from the previous steps in fully connected layers. The outputs are stored in an N-dimensional vector containing the N class probability [12]. The training process is determined by the batch size and epoch, with logistic sigmoid as the activation function in the dense layer.

2.1. Evaluation measure. Accuracy, precision, recall, and F-measure are used to evaluate the overall performance of a classifier. They were calculated based on the confusion matrix showed in Table 1. It contains the amount of data for which the row showed the actual class and the column as predicted class [13] Accuracy evaluates the model by estimating the probability of the true value of the overall class label [14]. Precision is the fraction of correctly classified instances over all the instances available, while recall is the fraction of correctly classified instances over the number of relevant instances. F-measure is a measure of accuracy for the classification problem, the harmonic mean of precision and recall. All the metrics used in this study are written in the equation for $Recall = \frac{TP}{TP+FN}$ (4) [15].

Table 1. Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

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

$$F-measure = \frac{2 \times TP}{2 \times TP+FP+FN} = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

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

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

2.2. Word cloud. A simple and effective way for text visualization. Word cloud is a picture displaying the frequent words sized respectively to their frequency. The picture is usually used to show the important words in the text, with the bigger as the more important word [16].

3. Results and Discussion

The data was collected on December 12th, 2020, filtered using ID (Indonesian language) with a total of 20 keywords from the name of the service providers, limited to Netflix and Disney, and the combination of them with the terms drama, series, film,

movie, streaming, review, *nonton* (watch), *sinopsis* (synopsis), and *rekomendasi* (recommendation). The tweets are labeled manually according to the relevancy of the tweets to a content recommendation on Twitter to three categories:

- Nonrelevant: coded as 0, for tweets having no relation to the content of streaming services, for example, the speed of internet connection or promotion tweets.
- Relevant: coded as 1, for tweets opening a discussion about content, containing the title of series/movies/animation, etc.
- Tweets from other languages: coded as 2 for tweets not written in Indonesian, for example, Malay or Indian. Tweets in this category were removed and not included in the analysis.

Table 2. Category of labels

Category	Number
0 – Nonrelevant	1,802
1 – Relevant	1,603
2 – Tweets from other languages	285
Total	3,690

The number of tweets in each category is shown in Table 2, with a total of 3,690 tweets. After removing tweets written in other languages, there were 3,405 tweets used in the analysis. The first step was checking duplicate values and removing foreign tweets, resulting in 1,423 irrelevant and 1,297 relevant tweets, amounting to 52,3% and 47,7% of all data, respectively. The next step was converting into lower case and splitting. Afterward, word clouds for both categories were shown as an initial comparison between categories.

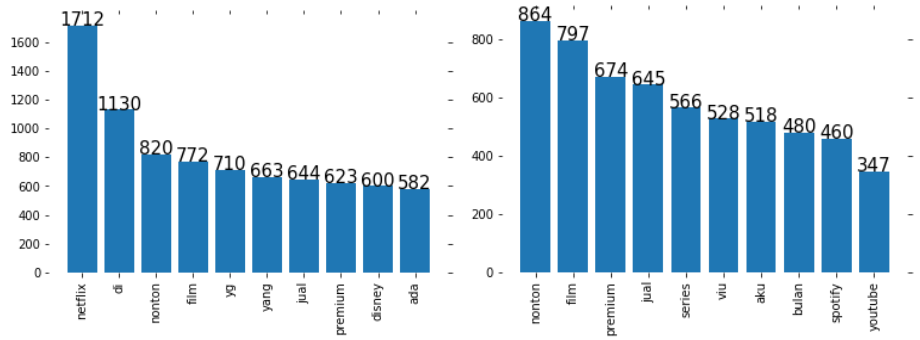
(a) (b)

Figure 1. Word cloud for (a) Nonrelevant (b) Relevant tweets

Figure 1 shows the frequent words in each category. There were many unrelated terms in Figure 1 (a) and many different words were shown in the same size, as there were no obvious key terms shown in the cloud. There are 65,031 words listed from both

categories. The top ten words were shown in Figure 2 and from the figures, it can be inferred that there are words with no significant meaning leading in numbers, such as *di*, appearing 1,130 times, *yg* and *yang* appearing over 600 times, *ada*, written over 500 times, and so on. The existence of these words suggested that data preprocessing is needed for the tweets, and those words were included in the list of removed words.

The first stage of preprocessing step involved: removing links, digits, username, punctuation, emoji, and finally concluded with case-folding. Afterward, Sastrawi library is used to remove the stopwords in the sentences. The stopwords were obtained from the package.



(a) (b)

Figure 2. Top ten words in the tweets (a) before preprocessing (b) after preprocessing

After data preprocessing, there are 52,118 remaining words. There were visible changes on the top ten words, with certain words missing from the top list. Word nonton now leading with 864 words, followed by film and premium, *jual*, and series. Some of the top words had no relation with content recommendation, such as *jual* (sell), premium, Spotify (music platform), and so on, suggesting that those words might be the keywords for the irrelevant tweets.

The next step was constructing bag-of-words from the sentences in each instance, with 1297 words chosen as the selected features. The clean data was split into 80% training and 20% testing. The training set was used for parameter tuning with a 5-fold Stratified CV for both methods. For the remaining steps in the analysis, the random forest required TF-IDF while the CNN required word embedding. The steps of this research were summarized in the flowchart displayed in Figure 3.

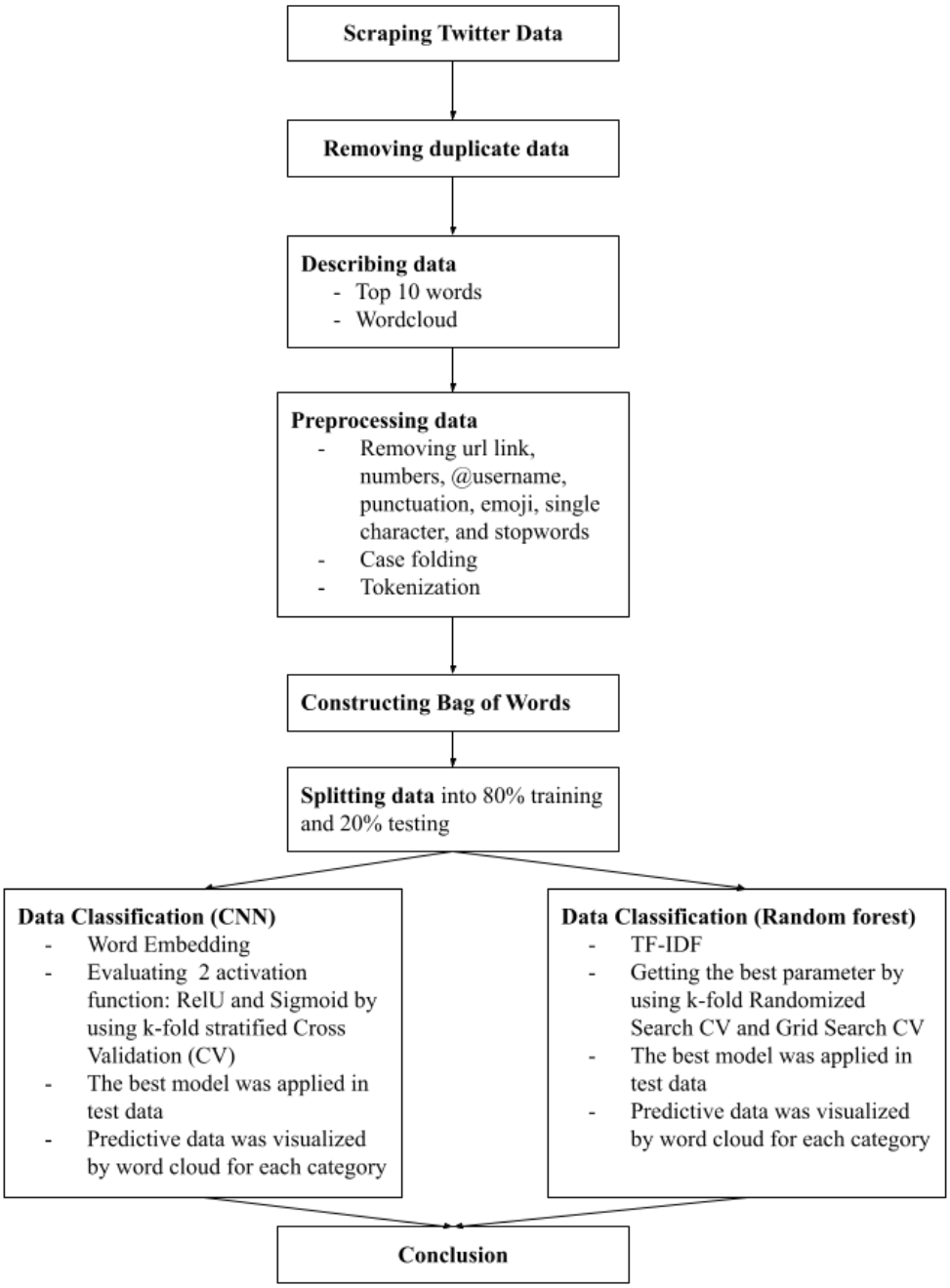


Figure 3. Flowchart of research steps

3.1 Random forest classification. The tweets were classified using random forest classification using scikit.learn package in Python. To get the best result for the data, tuning the hyperparameter to find out the best parameter for the analysis is required [17]. To investigate the best parameter for the random forest, an initial check for the parameters using Random Search was employed. There were several parameters considered in this study: number of trees, number of features, number of levels, the

minimum number before node splitting, and the resampling method for the data. Each parameter was given several choices of value to be executed in Random Search. From the pool of available parameters written in Table 3, half of all possible combinations, 240 parameter combinations, were sampled and evaluated.

Table 3. Parameters for random forest

Parameters	Description	Value (1st tuning)	Value (2nd tuning)
n_estimators	Number of trees	100,200,300,400,500, 600,700,800,900,1000	650,700,750
max_features	Maximum number of features	auto, sqrt	auto
max_depth	Maximum number of levels	10,20,30,40,50,None	None
min_samples_split	Minimum number of data before node splitting	5,10	3,5,7
bootstrap	Data sampling method True: with replacement False: without replacement	True, False	False

From the result of the sampled parameter combination, the best parameter was: bootstrap: False, max_depth: None, max_features: auto, min_samples split: 5, and n_estimators: 700. For the second part, the choices for the number of trees and the number of min sample split were added with the value close to the chosen parameter. The second parameter tuning used Grid Search with the parameters written in the value of 2nd tuning in Table 3.

Table 4. Result of random forest (2nd tuning)

min_samples_split	n_estimator	Accuracy	Precision	Recall	F1 score
3	650	85.43%	86.04%	85.12%	85.26%
	700	85.57%	86.19%	85.26%	85.40%
	750	85.52%	86.06%	85.23%	85.37%
5	650	85.61%	86.36%	85.28%	85.43%
	700	85.62%	86.35%	85.28%	85.43%
	750	85.52%	86.30%	85.18%	85.33%
7	650	85.48%	86.23%	85.14%	85.29%
	700	85.29%	86.08%	84.95%	85.10%
	750	85.48%	86.26%	85.13%	85.28%

Based on the evaluation metric written in Table 4, the best result was obtained using n_estimator of 700 and min samples split of 5 with the accuracy of 85,62%. The F1 score for this combination was also better compared to the others. This parameter was applied to the test set, resulting in 83,823% of accuracy. The random forest was generally able to classify the tweets based on their relevance to the content recommendation of streaming services on Twitter.

3.2 Convolutional Neural Network. After obtaining bag-of-words, the words were transformed into feature vectors. Word2vec technique was used so that each word becomes a vector of weight that represents its characteristics. Afterward, the embedding layer was constructed to implement those vectors. At this point, the data is ready to be trained. Trained data was 80% of overall data. To investigate the best model, two activation functions were compared and evaluated with k-fold Cross-Validation (CV). The results were shown in Table 5.

Table 5. Result of Activation Function Evaluation on CNN

No	Activation function	k-fold CV	Accuracy	Recall	Precision	F1-Score
1	RelU	3	71.05%	68.41%	78.40%	72.36%
		5	70.86%	70.96%	71.88%	72.12%
		10	72.95%	73.44%	73.00%	73.23%
2	Sigmoid	3	66.36%	72.05%	64.58%	66.53%
		5	65.63%	76.09%	62.93%	68.17%
		10	64.88%	68.57%	64.09%	67.20%

Table 4 showed that the best model used RelU as its activation function and evaluated using 10-fold CV. It can be known from its accuracy (72.95%) which is higher than the others. The model summary is shown in Table 6.

Table 6. Best Model Summary of CNN on Training Data

Layer (type)	Output shape	Parameter
embedding (Embedding)	(None, 7924, 100)	796000
dropout_18 (Dropout)	(None, 7924, 100)	0
conv1d_18 (Conv1D)	(None, 3961, 300)	90300
conv1d_19 (Conv1D)	(None, 1980, 150)	135150
conv1d_20 (Conv1D)	(None, 989, 75)	33825
flatten_6 (Flatten)	(None, 74175)	0
dropout_19 (Dropout)	(None, 74175)	0
dense_12 (Dense)	(None, 150)	11126400
dropout_20 (Dropout)	(None, 150)	0
dense_13 (Dense)	(None, 2)	302
Total params: 12,181,977		
Trainable params: 11,385,977		
Non-trainable params: 796,000		

The best model then implemented to test data and get 94% of accuracy. It was higher than the accuracy in the training set. It was understandable that CNN was able to classify the tweet, but it indicates that the model was underfitting. The model summary of CNN on testing data is presented in Table 7.

Table 7. Model Summary of CNN on Testing Data

Layer (type)	Output shape	Parameter
embedding (Embedding)	(None, 2882, 100)	290900
dropout_9 (Dropout)	(None, 2882, 100)	0
conv1d_9 (Conv1D)	(None, 1440, 300)	90300
conv1d_10 (Conv1D)	(None, 719, 150)	135150
conv1d_11 (Conv1D)	(None, 359, 75)	33825
flatten_3 (Flatten)	(None, 26925)	0
dropout_10 (Dropout)	(None, 26925)	0
dense_6 (Dense)	(None, 150)	4038900
dropout_11 (Dropout)	(None, 150)	0
dense_7 (Dense)	(None, 2)	302
Total params: 4,589,377		
Trainable params: 4,298,477		
Non-trainable params: 290,900		

3.3 Comparison of the results. The result of both methods suggested that there were distinctive characteristics between the two categories, as the accuracy for both exceeded 70% for the training set. To determine the method that works better for the tweets, the evaluation metrics for the test set were compared in Table 8. The CNN model showed a higher accuracy but slower in running time, indicating that the method was good but inefficient for classifying the relevant and nonrelevant tweets about the content recommendation in streaming services.

Table 8. Comparison of random forest and CNN

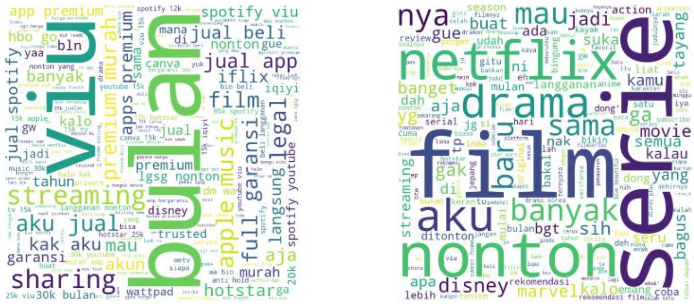
Methods	Accuracy	Precision	Recall	F1 score	Running time
Random Forest	84.38%	85%	84%	84%	2.06 s
CNN	94%	89%	100%	94%	102.33 s

The two word clouds generated from the test set of random forests are shown in Figure 4, while the results from CNN are shown in Figure 5. All the figures suggested that different words appeared often in both categories. For the nonrelevant tweets, there are many words not related to the content of video streaming services.

(a) (b)

Figure 4. Random forests test-set: Word cloud for (a) Nonrelevant (b) Relevant tweet

The top 10 words that can be avoided for getting movie content recommendation for Netflix and Disney on Twitter are *aku jual* (I sell), *apple music* (music platform), *viu* (different streaming platform), *bulan* (month), legal, *murah* (affordable), streaming, sharing, *full garansi* (full guarantee), and premium. The word cloud also displayed the name of streaming service platforms in the relevant tweets (Netflix and Disney), suggesting that people should write the keywords with the specific platforms in mind while searching for a content recommendation on Twitter. The top 5 words that can be used to get any recommendation about movie content are film, serie, Netflix, *nonton* (watch), drama, and *baru* (new).



(a) (b)

Figure 5. CNN test-set: Word cloud for (a) Nonrelevant (b) Relevant tweets

4. Conclusion

The result of this research concluded that CNN performed higher accuracy but slower in running time to classify the tweets compared to Random Forest in the test set. The results suggested that there was indeed a distinctive category between relevant and nonrelevant tweets about streaming services' content recommendation in Twitter. By observing the resulting word cloud, Twitter users could obtain a general idea of what

words they should write and the words to avoid in the search query if they were going to look for a content recommendation in Twitter in the future.

Future research should consider the other potential CNN parameter combination such as learning rate, epoch, batch size, and so on to prevent underfit or overfit model. It is also important to examine other random forest parameter combinations to get more optimum results.

References

[1] We Are Social & Hootsuite, Indonesia Digital report 2020. *Glob. Digit. Insights*, p. 247, 2020, [Online]. Available: https://datareportal.com/reports/digital-2020-global-digital-overview.

[2] Lidwina, A. Persaingan streaming video di Indonesia. 2020. https://katadata.co.id/ariayudhistira/infografik/5f3c7825c56e5/persaingan-streaming-video-di-indonesia?

[3] Lee, C. C., Nagpal, P., Ruane, S. G., and Lim, H. S. Factors affecting online streaming subscriptions. *Commun. IIMA*, vol. 16, no. 1, Jan. 2018.

[4] Lee, H. K., Lee, H. J., Park, J., Choi, J., and Kim, J. B. A Study of Predict Sales Based on Random Forest Classification, *Int. J. u-and e-Service*. 10(7): 25–34, 2017.

[5] Untawale T. M. and Choudhari, G. Implementation of sentiment classification of movie reviews by supervised machine learning approaches. *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019*, 1197–1200. 2019.

[6] Fithriasari, K., Jannah, S. Z., and Reyhana, Z. Deep Learning for Social Media Sentiment Analysis, *Mat. MJIAM*, vol. 36, no. 2, pp. 99–111, 2020.

[7] Cambridge Dictionary, STREAMING | meaning in the Cambridge English Dictionary. https://dictionary.cambridge.org/dictionary/english/streaming (accessed Jan. 23, 2021).

[8] Feldman, R. and Sanger, J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, 2006.

[9] Alessa, A. and Faezipour, M. Tweet Classification Using Sentiment Analysis Features and TF-IDF Weighting for Improved Flu Trend Detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2018.

[10] Breiman, L., Random forests. *Mach. Learn.*, 45(1): 5–32. 2001.

[11] Kim Y. Convolutional Neural Networks for Sentence Classification. Accessed: Feb. 26, 2021. [Online]. Available: http://nlp.stanford.edu/sentiment/.

[12] Patil, S., Gune, A., and Nene, M. Convolutional neural networks for text categorization with latent semantic analysis. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017*, pp. 499-503. 2018.

[13] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. *Data Mining: Practical Machine Learning Tools and Techniques*. 2016.

[14] Bekkar, M., Djemaa, H. K., and Alitouche, T. A. Evaluation Measures for Models Assessment over Imbalanced Data Sets. *J. Inf. Eng. Appl.*, 3(10): 27–38. 2013.

[15] Sasaki, Y. The truth of the F-measure. *Teach Tutor Mater*, pp. 1–5. 2007.

[16] Mueller, A., *et al.*, amueller/word_cloud: WordCloud 1.5.0. 2018.
[17] Koehrsen, W. Machine-Learning-Projects/random_forest_explained at master · WillKoehrsen/Machine-Learning-Projects · GitHub, *Hyperparameter Tuning the Random Forest in Python*, Jan. 10, 2018. https://github.com/WillKoehrsen/Machine-Learning-Projects/tree/master/random_forest_explained? (accessed Feb. 17, 2021).