Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 179 (2021) 47–54

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Computer Science and Computational Intelligence 2020 (ICCSCI), 2-4 September 2020

# Multi-class Weather Forecasting from Twitter Using Machine Learning Aprroaches

Kartika Purwandari[a,*], Join W. C. Sigalingging[b], Tjeng Wawan Cenggoro[a,c], Bens Pardamean[a,d]

[a]*Bioinformatics & Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia 11480*
[b]*Database Center Division of BMKG, Meteorological, Climatological, and Geophysical Agency, Jakarta, Indonesia 10720*
[c]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
[d]*Computer Science Department, BINUS Graduate Program - Master of Computer Science Program, Bina Nusantara University, Jakarta, Indonesia 11480*

**Abstract**

Now in the era of big data, many are applying information methods accurately especially by social media. The aims of this study to classify the weather based on Twitter automatically using text mining by using Support Vector Machine (SVM), MultinomialNaive Bayes (MNB), and Logistic Regression (LR) method. The experimental results show that SVM substantially outperforms various other machine learning algorithms for the task of text classification with an accuracy value of 93%. This result proves that SVM is very suitable for text categorization. We use clustering technique to read the pattern in customers' opinion about the restaurant based on some measurement variables.



*Keywords:* weather classification;twitter;machine learning;support vector machine;multinomial naive bayes; logistic regression;

## 1. Introduction

### *1.1. Background*

The combination of high population size and high levels of biodiversity, combined with an unprecedented 80,000 km of coastline and 17,508 islands, makes Indonesia one of the most vulnerable countries in terms of climate change. Indonesia has a tropical climate on both sides of the equator with two seasons; the rainy season and the dry season [1].

* Corresponding author. Tel.: +62-8232-822-5813
*E-mail address:* kartika.purwandari@binus.edu


10.1016/j.procs.2021.12.006

Weather forecasting is always relevant. Users can experience this about four-five times a day by using mobile phone apps, TV, newspaper, tweets, etc. Climate determination is one of the most popular topics that has changed people's lives and activities for a long period of time. The assessment of climate data may be a discreet-continuous option problem. Then, rainfall is difficult to predict precisely because of the energetic precipitation and the diverse physical forms involved. Precipitation is a crucial component of climate change, because it is a crucial source of water for life [2]. The use of weather knowledge also has a major effect on the profitability of everyday human activities, which has an effect on economic and social impacts. As an example in the field of agriculture [3]. Estimates based on temperature and precipitation are critical to horticulture, and thus to dealers inside product markets. The increase in the use of climate information was more prominent among producers with a differentiated agrarian generation. Differentiated producers were more likely to use the information on the timing of planting, growth, and harvesting [4].

One bad thing is that, in some situations, the climate forecast is not reliable. None consider shifts in precipitation to hourly timescales due to insufficient under-daily expectations and inability of climate models to re-create sub-daily precipitation reliably [5]. One bad thing is that, in some situations, the climate forecast is not reliable. None consider shifts in precipitation to hourly timescales due to insufficient under-daily expectations and inability of climate models to re-create sub-daily precipitation reliably. Now in the era of big data, many are applying information methods accurately and in real-time. Including for weather information. One of the data used is data from social media, such as Twitter. In this paper, we will process Twitter data to provide the latest weather information to the public. The information will extract the clue and topic of weather in Indonesia. It means we are doing the text classification using weather data. This like a chance for us to do weather classification based on Twitter from people around Indonesia.

This paper demonstrated machine learning model to process Twitter data in order to provide public with the latest weather information. The purpose of this paper is how to identify the weather based on Twitter. There have been several experiments using various models of machine learning to classify texts. Some of the approaches that have proved to be successful are Support Vector Machine (SVM) [6,7,8], Multinomial Naive Bayes (MNB) [9], and Logistic Regression (LR) [10]. Several Natural Languange Processing (NLP) techniques are used in this study. Natural Language Toolkit (NLTK), which may be a python library for NLP developed by Microsoft, still played a critical role as its ability to perform various basic tokenization exercises. Comparable to NLP in Japanese [11], the correct combination of devices and pre-processing techniques for content knowledge in Bahasa Indonesia is evaluated in this study.

### *1.2. Weather Classification*

Just time-series data from previous years have been used for the last weather forecast in Indonesia. Skill (accuracy)prediction system already operational at Meteorology, Climatology, and Geophysical Agency (BMKG) in forecasting bulk Rainy dichotomy (yes/no) is good enough compared to predictions of moderate rainfall and very low rainfall. Several models, such as the Weather Research and Forecasting (WRF) model, are capable of producing predictions that are approaching the results of operational predictions and are also capable of predicting extreme and very extreme rainfall intensity, although they appear to produce over-forecast predictions. Based on Gustari's experiment, it shows that forecasting rainfall daily, particularly for heavy and very heavy categories, still needs more research and study to produce good predictive quality [12].

A lot of weather classifications are performed in the form of images. Martin and Frank use single-color images to describe the weather classification as clear air, light rain, heavy rain using the Support Vector Machine (SVM) model [13].

### *1.3. Twitter Dataset on Weather Condition*

There are several ways of disseminating climate statistics, their themes are created and distributed widely and rapidly by users of Twitter. And the retweet distribution doesn't say a difference between the tweets and the rumors. Subsequently, a clue is needed to identify rumors [14].

Many weather conditions details did not apply to the Twitter dataset, whereas other tests typically used time-series data such as humidity, temperature per day, weekly, monthly, or annual. Yet the Twitter dataset was used to track violence. As of now, the Twitter dataset has been used in a few experiments to predict the time and location in which a specific form of crime is likely to occur. It may discern the relationship between certain points of view with the extremity of view and the actual criminal activity. Chen et al [15] used measured regression for predictive modeling.

Support vector machine or other specialized techniques can be used to validate the possible non-linear effect between extremity and crime.

## 2. Method

The study method is going through a variety of phases. Figure 1 provides a general overview of the structure for a traditional classification scheme. As shown in Figure 1, this work was performed primarily in five phases: data collection, pre-processing, Term Frequency - Inverse Document Frequency (TFIDF), classification, and evaluation, as defined in the following section.


Fig. 1. Research Methods

### 2.1. Data Collection

Twitter information was obtained from the Twitter social organization via the Twitter Application Programming Interface (API) using the "Twitter" package in R. The feature of this is the use of the Twitter message content known as tweets. The keywords used are sunny, cloudy, rainy, heavy rain, and thunderstorms. The region should have been based on the Indonesian nation. This study used the language 'id' code which means to sign it as an Indonesian tweet.

### 2.2. Preprocessing

Twitter data has a noisy text. This was first designed so that it could be used at the next level. The method of storing raw data is also called pre-processing. Preprocessing helps to translate unstructured text data into structured data[16]. The preprocessing step is conducted as follows:

- Case folding has fixed lowercase letters and the cleaning process eliminates URLs, emails, usernames, hashtags, or Twitter data obtained.
- Delete characters other than 'a' to 'z' including the addition of numbers and punctuation marks. Another procedure is to delete the rehashed vocals in the class at least three times.
- Stemming up the words. Sastrawi may be a simple Python library that allows you to reduce the bent terms in Indonesian dialect (Bahasa Indonesia) to their base frame (stem).
- Tokenizing splits tweets into words that use whitespace characters as a cutter.
- Filtering has taken important words out of the tokenization stage. This step would delete a word that does not have a specific meaning by using a stopword dictionary.
- Adjust the KBBI-based sentence. Kamus Besar Bahasa Indonesia (KBBI) is the official word reference for Indonesian, distributed by Badan Pengembangan Pembinaan Bahasa (The Language Improvement and Development Organization) or Badan Bahasa under the Instruction and Culture Service of the Republic of Indonesia.

### 2.3. Term Frequency - Inverse Document Frequency (TFIDF)

TF-IDF is one of feature extraction processes[17]. Once the punctuation is eliminated and the words are reduced, the meaning of a word is determined by its frequency[18]. Binary weighting is the formation of feature vectors by looking at the presence or absence of a term in the document. TF is the formation of feature vectors from the number of terms in a document. IDF is TF multiplied by Log (N / $n_i$) where N is the total number of documents, and $n_i$ is the number of documents containing the term. TF-IDF is a combination of TF and IDF where TF is multiplied by IDF.

### 2.4. Classification

The goal from the classification stage is grouping Twitter data into class sunny, cloudy, rainy, moderate rain, and thunderstorm by implemented Support Vector Machine (SVM), Multinomial Navie Bayes (MNB), and Linear Regression (LR) methods.

- SVM takes the set of preparing information and checking it as a portion of a category at that point predicts whether the test archive may be a part of an existing class. SVM models speak to the information as a point in space separated by a hyperplane line [19]. The hyperplane line equation, which is defined [20]:

$$w.x + b = 0 \tag{1}$$

where, w is the normal plane and b is the bias or position of the field relative to the coordinate center. Whereas the vector which has the closest distance to a hyperplane is called a support vector. It is assumed that both class -1 and +1 can be completely separated by hyperplane. x patterns that belong to class –1 (sample i negative) can be formulated as a pattern that satisfies inequality:

$$w.x + b <= 1 \tag{2}$$

while the x pattern is included as class +1 (positive sample):

$$w.x + b >= +1 \tag{3}$$

- MNB could be a variety of the Naive Bayes outlined to illuminate the classification of content records. MNB employments multinomial dispersion with the number of events of a word or the weight of the word as a classification include [19]. MNB model assigns the test document $t_i$ to the class with the highest probability $Pr(c|t_i)$ which is given by Bayes' rule like equation 4.

$$Pr(c|t_i) = \frac{Pr(c)Pr(t_i|c)}{Pr(t_i)}, c \in C \tag{4}$$

The class prior $Pr(c)$ can be estimated by dividing the number of documents belonging to class $c$ by the total number of documents. $Pr(t_i|c)$ is the probability of obtaining a document like $t_i$ in class $c$ [9].

- LR is an alternative test if the assumption of the multivariate normal distribution cannot be fulfilled when the discriminant analysis is conducted. This assumption is not fulfilled because the independent variable is a mixture of continuous (metric) and categorical (non-metric) variables [10,21]. Consider the LR for multiple classification model with response variable $Y_i$ and $p$ predictor variables $(X_{i1}, ..., X_{ip})$ for i = 1, ..., N:

$$Y = X * \beta + \epsilon \tag{5}$$

where $Y = (Y_1, ..., Y_N)^T$ is an $Nx1$ vector with $Y_i$ being the observed measurement of the $i$th words; $X = (1_N, X_S)$ with $1_N$ is the $Nx1$ vector of all 1's, $X_S = (X_{S1}, ..., X_{SN})^T$ is an $Nxp$ matrix, $X_{Si} = (X_{i1}, ..., X_{ip})^T$, where $X_{i1}, ..., X_{ip}$ are the observed values of the $p$ predictor variables of the $i$th subject; $\beta = (\beta_I, \beta_s^T)^T$ is a $(p + 1)x1$ vector with $\beta_S = (\beta_1, ..., \beta_p)^T$, where $\beta_1, ..., \beta_p$ are unknown coefficient parameters; and $\epsilon = (\epsilon_1, ..., \epsilon_N)^T$ is an $Nx1$ vector with $\epsilon_i, ..., \epsilon_N$ are random variables.

### 2.5. Classification Performance Evaluation

For problems in classifications involving multi-class classifications, the most commonly used performance measures are the accuracy, precision, recall and F1-score [22,23]. These four measures distinguish between the correct grouping of labels in various categories. Recall is a mixture of its correctly classified examples (true positive) and its wrongly

classified examples (false negatives). Precision is a mixture of true positive and misclassified (false positive) cases. The F1-score is evenly distributed when $\beta = 1$. Precision is favoured when $\beta > 1$, and otherwise recalls.

## 3. RESULT AND DISCUSSION

### 3.1. Data Preparation

Tweet data was collected from the crawling process for 1 month from 1 January to 31 January 2020. The data was obtained as much as 412, followed by data cleaning from redundant tweets to 368 data. The data were classified manually by BMKG expert judgment. The data were divided into two parts, namely training and testing. For this analysis, 268 data were randomly distributed as training and 100 data as testing. Depending on the data already given by the labeling, the number of categories can be counted. Figure 2 shows the distribution of training results. Figure 2 shows that the data in the thunderstorm class was collected with the most amount of data and data that was in the rainy class was collected with the least amount of data. This can be seen that in January thunderstorm often occurs in the territory of Indonesia.

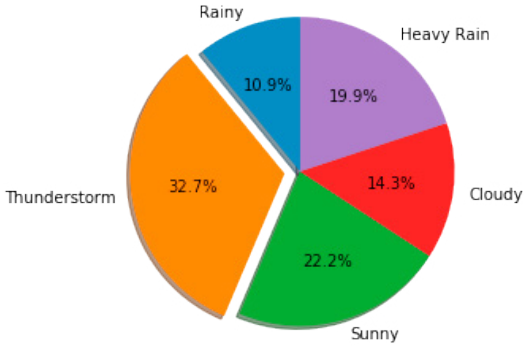


Fig. 2. Data Training Distribution

### 3.2. Word Cloud

A WordCloud that visualized as Figure 3, was developed as a preliminary analysis to see the graphic representation of word frequency from the whole collected tweets. It depicts the 100 most frequently utilized words, with each word's frequency correlated with font size. In plain view, a series of words: cuaca, hujan, panas, angin kencang (weather, rainy, hot, strong wind) is dominating in the word cloud.


Fig. 3. WordCloud

### 3.3. Experiment Result

This study carried out all experiments using SVM, MNB, and LR processes. After all, documents have been classified, testing is done using a confusion matrix consisting of four methods, namely accuracy, precision, recall, and f-measure. Based on the results of the experiments, the values of precision, recall, and f1-score were obtained as shown in Table 1. Table 1 results have used the same pre-processing and TFIDF results as the model input feature.

Table 1. Evaluation Result

| Class | SVM (Accuracy = 93%) | | | MNB (Accuracy = 53%) | | | LR (Accuracy = 88%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Precission** | **Recall** | **F1-score** | **Precission** | **Recall** | **F1-score** | **Precission** | **Recall** | **F1-score** |
| Cloudy | 0.86 | 1 | 0.92 | 1 | 0.33 | 0.5 | 0.92 | 1 | 0.96 |
| Sunny | 0.93 | 0.96 | 0.95 | 1 | 0.86 | 0.92 | 0.96 | 0.89 | 0.93 |
| Rainy | 0.97 | 0.88 | 0.92 | 0 | 0 | 0 | 1 | 0.76 | 0.87 |
| Thunderstorm | 0.95 | 0.95 | 0.95 | 0.33 | 1 | 0.5 | 0.66 | 1 | 0.79 |
| Heavy Rain | 0.86 | 0.86 | 0.86 | 0.4 | 0.86 | 0.55 | 1 | 0.86 | 0.92 |
| **AVG** | **0.914** | **0.93** | **0.92** | **0.546** | **0.61** | **0.494** | **0.908** | **0.902** | **0.894** |

Based on Table 1, the average results of precision, recall, and f1-score in the SVM method show higher results than other methods. The average performance, recall, and f1-score results for the SVM system were 0.914, 0.93, and 0.92, respectively. SVM substantially outperformed other machine learning algorithms for text classification tasks, including the Twitter weather classification with 93% accuracy. Second, LR was able to achieve an accuracy of 88% and, subsequently, the MNB only able to achieve an accuracy of 53%.

In the table obtained perfect precision values for the cloudy and sunny classes and perfect recall values for the thunderstorm class in the MNB method. But the precision is perfect for the rainy and heavy rain classes and the perfect recall value for the Cloudy and Thunderstorm classes in the LR method. In the SVM method, there is no perfect precision value for each class, but it produces a perfect recall value in the cloudy class.

Table 2. Confusion Matrix SVM

| **Actual** | | | | | | |
|---|---|---|---|---|---|---|
| | **Cloudy** | 11 | 0 | 0 | 0 | 1 |
| | **Sunny** | 1 | 27 | 0 | 0 | 0 |
| | **Rainy** | 0 | 1 | 31 | 2 | 0 |
| | **Thunderstorm** | 0 | 0 | 0 | 18 | 1 |
| | **Heavy Rain** | 0 | 0 | 0 | 1 | 6 |
| | | **Cloudy** | **Sunny** | **Rainy** | **Thunderstorm** | **Heavy Rain** |
| | **Predicted** | | | | | |

The results of the confusion matrix using the SVM model are shown in Table 2. There are some incorrect prediction results for each of the data, namely in the sunny, cloudy, heavy rain and thunderstorm categories. It means many data were correctly classified by the SVM model. Whereas there is a mismatch of more forecasts in the category of rainy, namely a number of two data included in the category of heavy rain. The results obtained can be due to very significant weather changes written in a person's twitter. For example after sunny sometimes there will be a sudden cloudy, so also after the rain there will also be sudden thunderstorms.

Table 3. Confusion Matrix MNB

| **Actual** | | | | | | |
|---|---|---|---|---|---|---|
| | **Cloudy** | 4 | 0 | 0 | 3 | 5 |
| | **Sunny** | 0 | 24 | 0 | 3 | 1 |
| | **Rainy** | 0 | 0 | 0 | 31 | 3 |
| | **Thunderstorm** | 0 | 0 | 0 | 19 | 0 |
| | **Heavy Rain** | 0 | 0 | 0 | 1 | 6 |
| | | **Cloudy** | **Sunny** | **Rainy** | **Thunderstorm** | **Heavy Rain** |
| | **Predicted** | | | | | |

Table 3 is displays the results of the confusion matrix using the MNB method. It can be shown that the category of thunderstorms does not have a misclassification in that category, but there are many mistakes in forecasts in other

categories, such as cloudy, sunny, rainy, and heavy rain. That's what makes this approach the lowest degree of accuracy possible.

Table 4. Confusion Matrix LR

| | | | | | | |
|---|---|---|---|---|---|---|
| **Actual** | **Cloudy** | 12 | 0 | 0 | 0 | 0 |
| | **Sunny** | 1 | 25 | 0 | 2 | 0 |
| | **Rainy** | 0 | 1 | 26 | 7 | 0 |
| | **Thunderstorm** | 0 | 0 | 0 | 19 | 0 |
| | **Heavy Rain** | 0 | 0 | 0 | 1 | 6 |
| | | **Cloudy** | **Sunny** | **Rainy** | **Thunderstorm** | **Heavy Rain** |
| | **Predicted** | | | | | |

The results of the confusion matrix using the LR method are shown in Table 4. It can be seen that there is no misclassification of the cloudy and thunderstorm categories in that category. In the category of heavy rain, one data is categorized into the category of thunderstorm class. Whereas there are more forecast errors in the rainy and sunny areas.

### *3.4. Discussion*

These experiments result to prove that rain, light rain, and heavy rain have some keywords in common for example the keyword "rain". This is caused by the similarity of those keywords which affect the TF-IDF value calculation process where this value is used as input for the learning process in the model. As a result, the SVM model sometimes experiences ambiguity in the results of the analysis, likewise with the MNB and LR methods.

Based on the experiment result, SVM has proven that this method is very effective in classifying text. Most of the use of SVM in information retrieval is in the classification and categorization of documents, as well as in the searching process that applies relevance feedback. The ability of fast SVM even for high-dimensional data is sometimes the right solution for information retrieval needs that require speed [17]. Based on the results in table 1 can be seen the superiority of SVM compared to other methods. This proves the statement from Joachims that theoretically and experimentally that SVM is very suitable for text categorization. Text categorization will find a very large number of features (more than 10000), and SVM tends not to depend on the size of the data dimensions [17]. SVM runs optimally for multi-class classification using kernel designed for multidimensional data.

## 4. CONCLUSION

This study has succeeded in proving the good application of machine learning on Twitter data about the weather. In the three methods used (SVM, MNB, and LR), the SVM substantially outperformed the other methods by 93%. Experiment success in improving the precision of previous studies [24]. Further changes can be made by using more reliable and more precise datasets to improve precision by using the deep learning or transfer learning model. Potential research may also conduct an in-depth analysis to identify weather conditions based on all social media.

## References


1. Case, M., Ardiansyah, F., Spector, E., et al. Climate change in indonesia: implications for humans and nature. *Climate change in Indonesia: implications for humans and nature* 2007;.
2. Caraka, R.E., Bakar, S.A., Tahmid, M., Yasin, H., Kurniawan, I.D.. Neurocomputing fundamental climate analysis. *Telkomnika* 2019; **17**(4):1818–1827.
3. Caraka, R.E., Ulhusna, M., Supatmanto, B.D., Goldameir, N.E., Hutapea, B., Darmawan, G., et al. Generalized spatio temporal autoregressive rainfall-enso pattern in east java indonesia. In: *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*. IEEE; 2018, p. 75–79.
4. Frisvold, G.B., Murugesan, A.. Use of weather information for agricultural decision making. *Weather, Climate, and Society* 2013;**5**(1):55–69.
5. Kendon, E.J., Roberts, N.M., Fowler, H.J., Roberts, M.J., Chan, S.C., Senior, C.A.. Heavier summer downpours with climate change revealed by weather forecast resolution model. *Nature Climate Change* 2014;**4**(7):570.
6. Forman, G.. Bns feature scaling: an improved representation over tf-idf for svm text classification. In: *Proceedings of the 17th ACM conference on Information and knowledge management*. 2008, p. 263–270.

7. Colas, F., Brazdil, P.. Comparison of svm and some older classification algorithms in text classification tasks. In: *IFIP International Conference on Artificial Intelligence in Theory and Practice*. Springer; 2006, p. 169–178.
8. Goudjil, M., Koudil, M., Bedda, M., Ghoggali, N.. A novel active learning method using svm for text classification. *International Journal of Automation and Computing* 2018;**15**(3):290–298.
9. Kibriya, A.M., Frank, E., Pfahringer, B., Holmes, G.. Multinomial naive bayes for text categorization revisited. In: *Australasian Joint Conference on Artificial Intelligence*. Springer; 2004, p. 488–499.
10. Pranckevičius, T., Marcinkevičius, V.. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing* 2017;**5**(2):221.
11. Rahutomo, R., Lubis, F., Muljo, H.H., Pardamean, B.. Preprocessing methods and tools in modelling japanese for text classification. In: *2019 International Conference on Information Management and Technology (ICIMTech)*; vol. 1. IEEE; 2019, p. 472–476.
12. Gustari, I., Hadi, T.W., Hadi, S., Renggono, F.. Akurasi prediksi curah hujan harian operasional di jabodetabek: Perbandingan dengan model wrf. *Jurnal Meteorologi dan Geofisika* 2012;**13**(2).
13. Roser, M., Moosmann, F.. Classification of weather situations on single color images. In: *2008 IEEE Intelligent Vehicles Symposium*. IEEE; 2008, p. 798–803.
14. Takahashi, T., Igata, N.. Rumor detection on twitter. In: *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*. IEEE; 2012, p. 452–457.
15. Chen, X., Cho, Y., Jang, S.Y.. Crime prediction using twitter sentiment and weather. In: *2015 Systems and Information Engineering Design Symposium*. IEEE; 2015, p. 63–68.
16. Samudrala, S.. *Machine Intelligence: Demystifying Machine Learning, Neural Networks and Deep Learning*. Notion Press; 2019. ISBN 9781684660834. URL https://books.google.co.id/books?id=LC2DDwAAQBAJ.
17. Purnamawan, I.K.. Support vector machine pada information retrieval. *Jurnal Pendidikan Teknologi dan Kejuruan* 2015;**12**(2):139–146.
18. Yun-tao, Z., Ling, G., Yong-cheng, W.. An improved tf-idf approach for text classification. *Journal of Zhejiang University-Science A* 2005; **6**(1):49–55.
19. Pratama, B.Y., Sarno, R.. Personality classification based on twitter text using naive bayes, knn and svm. In: *2015 International Conference on Data and Software Engineering (ICoDSE)*. IEEE; 2015, p. 170–174.
20. Nugroho, A.S., Witarto, A.B., Handoko, D.. Support vector machine. *Proceeding Indones Sci Meeiting Cent Japan* 2003;.
21. Mohandes, M.A., Halawani, T.O., Rehman, S., Hussain, A.A.. Support vector machines for wind speed prediction. *Renewable Energy* 2004; **29**(6):939–947.
22. Sokolova, M., Japkowicz, N., Szpakowicz, S.. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: *Australasian joint conference on artificial intelligence*. Springer; 2006, p. 1015–1021.
23. Powers, D.M.. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation 2011;.
24. Dilrukshi, I., De Zoysa, K., Caldera, A.. Twitter news classification using svm. In: *2013 8th International Conference on Computer Science & Education*. IEEE; 2013, p. 287–291.