International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Analysis of Models and Dataset used for Predicting Emotion in Text

Adzmer Muhali

Submitted: 26/06/2023 **Revised:** 06/08/2023 **Accepted:** 25/08/2023

Abstract: This paper presents the analysis of two models namely Naïve Bayes and Logistic Regression, and a dataset for predicting emotion in a text. The experiment use emotion dataset from KAGGLE website, containing 21,459 data with two columns labelled as Text and Emotion, emotion class consists of happy, anger, sadness, love, fear, and surprise. This is to evaluate the models and dataset applied in this research if it is good and enough for predicting emotion in text. Specifically, to apply data collection, data preparation, feature engineering, model building, and model evaluation. Based on the results, we conclude that Logistic Regression Model gives the best performance. In classification report, the result shows that the accuracy of Naïve Bayes is 77 percent only while Logistic Regression is 89 percent. The result for the best model performance also has the highest percentage of accuracy obtain rather than the previous research discussed in this paper that uses different models. The result of analysis for the dataset is good when it comes for training purposes but for the real time application, the data for each emotion should be balance since the dataset utilized in this research is an imbalance dataset.

Keywords: *Naïve Bayes, Logistic Regression, Accuracy, Precision, Recall, Classification Report, Confusion Matrix, Dataset, Emotion, Text*

1. Background

Lots of research done on predicting or classifying emotion in a text using machine learning models. The prediction and classification were made by the researchers out there, but how accurate the prediction was? Based on the dataset provided, what model used to provide an accurate prediction, and is the dataset was enough to use for prediction?

Several text-based emotion detections were proposed. Chaffar and Inkpen [2] evaluated three classification methods NB, J48, and SVM-SMO to recognize six basic emotions (anger, disgust, fear, happiness, sadness and surprise). The experiment was carried out using several datasets: Text Affect, Alm's dataset, Aman's dataset and the Global dataset. Based on the experiments, SVM-SMO outperformed to the other classification methods.

Muljono, Winarsih, and Supriyanto [1] presents Indonesian text emotion detection and evaluates the performances of four different classification methods: Naive Bayes (NB), J48, K-Nearest Neighbor (KNN) and Support Vector Machine-Sequential Minimal Optimization (SVM-SMO). They concluded that SVM-SMO classifier gives the best performance. In the 10-fold cross validation, the result shows that the accuracy of NB, J48, KNN and SVM-SMO are 80.2%, 80.8%, 68.1%, and 85.5% respectively. The same conclusion is also demonstrated by the split validation, the highest accuracy of 86% is also achieved by SVM-SMO.

Nivet Chirawichitchai [3] presents Thai text emotion classification by using several machine learning algorithm and various term weighting methods. Support Vector Machine (SVM) with Boolean weighting gave the best performance compared to Naïve Bayes (NB), K-Nearest Neighbor (KNN) and Decision Tree (DT). Thai emotion classification is also studied by Inrak and Sinthupinyo [4]. They proposed to use Singular Value Decomposition (SVD) method to reduce the dimension of the vector. In the experiment, SVM is the best classifier compared to NB and DT.

Li and Xu [5] proposed to use emotion cause extraction to support the emotion classification model. The cause of emotion was considered as an important factor for emotion detection. The model uses chi-square test to select the best features from the corpus. In the classification step, they use Support Vector Regression (SVR) which is variant of SVM.

Another approach is proposed by Jun Li et. al. [6]. They present Chinese text emotion classification based on emotion dictionary. The methods include WordNet for vector construction, SVM and NB for classification. In the comparison of classification methods, SVM gave the best accuracy compared to NB.

Arifin et. al. [7] present tweet emotion detection in Indonesian Language. Non-Negative Matrix Factorization (NMF) is proposed to reduce the number of features. KNN was used to classify 764 tweets from various emotions. Arifin and Ketut Eddy Pumama [8] also present emotion

Department of Computer Studies and Engineering, Sulu State College, Jolo, Sulu, Philippines
asmuhali@mymail.mapua.edu.ph/asmuhali@gmail.com

classification in Indonesian Language. KNN and SVM were used to classify the text corpus. As the result, SVM outperforms KNN in term of accuracy.

Based on the previous research, Support Vector Machine (SVM) model was provided the best performance than the other models they utilized and most of them applied also Naïve Bayes (NB) model in their research study. But in the previous research with best model performance which SVM obtained only a maximum of 86 percent of accuracy and below 86 percent for the other models. 86 percent is good but for the purposed of providing a more accurate one there's need an improvement or there should be another model needed to use for prediction to further increase the accuracy of prediction obtain, with that, in this paper we try another model which is Logistic Regression Model and used Naïve Bayes also for comparison of which models is best in performance and with high accuracy that the previous researchers obtain.

The main purpose of this paper is to analyse the models and dataset applied in this research if it good and enough for predicting emotion in text. Specifically, to apply data collection, data preparation, feature engineering, model building, and model evaluation.

Choosing a machine learning model and dataset for predicting emotion in a text needs to be reliable. Researchers, developers, and etc. should apply an appropriated model and dataset when they are conducting research regarding the prediction of emotion in a text. This research help those analyse the model and dataset used in this paper and also how are they going to choose a model and dataset for their future project related to this research.

The models applied in this paper are the Naïve Bayes and Logistic Regression model and the dataset that is available on the KAGGLE website which most of the research was applied in their research. The dataset is composed only of two columns labelled with Text and Emotion with the total of 21,459 data with 6 emotions labelled as happy, anger, sadness, love, fear, and surprise. The evaluation and analysis is focusing only on the accuracy, precision, and recall obtain during prediction with the used of classification report and confusion matrix.

This paper only analyses the results of prediction with the used of Naïve Bayes and Logistic Regression Model with the given dataset, it is not for the implementation of prediction system for emotion in a text. Other machine learning model aside from what are used in this paper is not utilized.

2. Methodology

This chapter centres its brief discussion on the following aspects: Data Collection, Data Preparation, Feature Engineering, Model Building, and Model Evaluation.

2.1. Data Collection

The researcher conducts some data gathering on the dataset being used in this paper. Searching online the available dataset for emotions which most of the researchers applied in their research was done. The dataset applied in this paper is available on the www.kaggle.com website, most of the researchers used this dataset on their machine learning modelling to predict emotion in a text.

The programming tools used in this paper is JUPYTER notebook with python programming language for constructing the prediction and evaluation. Required libraries also being collected and installed such as numpy, pandas, neattext, matplotlib, seaborn, textblob, counters, WordCloud, and sklearn.

Conceptual framework is also made in this part of research. The workflow of the analysing the models and dataset applied in this research that predict emotion in text is constructed to make sure that the evaluation made was reliable and efficient.

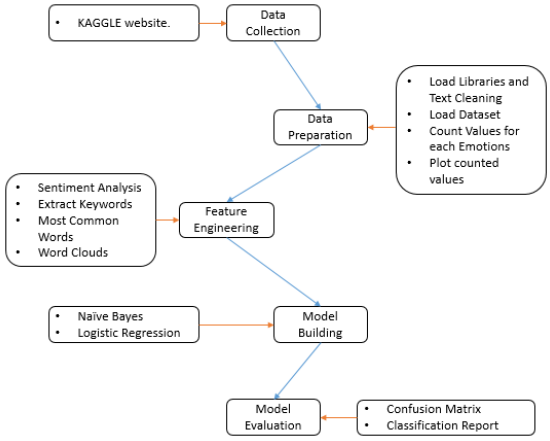

Fig. 1 Conceptual Framework

2.2. Data Preparation

This part the data being collected is prepared for feature extraction, modelling and evaluation. The packages are imported and loaded the text cleaning and also the dataset. Text cleaning was applied to clean those values on a text column to further increase the prediction efficiency. The dataset is composed of two columns labelled with Emotion and Text and a total of 21,459 data for both columns. For Emotion class the values are labelled with happy, sadness, anger, surprise, love, and fear with corresponding text to that emotions written in the text class and those columns are in object type.

	Text	Emotion
0	i didnt feel humiliated	sadness
1	i can go from feeling so hopeless to so damned...	sadness
2	im grabbing a minute to post i feel greedy wrong	anger
3	i am ever feeling nostalgic about the fireplac...	love
4	i am feeling grouchy	anger

Fig. 2 Some Sample of Dataset

Visualizing the dataset and counted each value on the emotion and text column are also created.

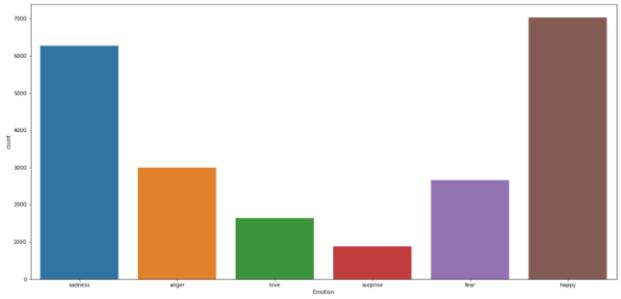

Fig. 3 Emotions value counted.

The text values counted for emotion happy is equal to 7,029 values, sadness is equal to 6,265 values, anger is equal to 2.993 values, fear is equal to 2,652 values, love is equal to 1,641 values, and surprise is equal to 879 values.

Table 1 Value for each Emotions

Dataset	
Emotion	Values
happy	7,029
sadness	6,265
anger	2,993
fear	2,652
love	1,641
surprise	879

2.3. Feature Engineering

Sentiment analysis is applied in this part of the research. It was not the purpose of this paper to identify the sentiment of each text values but to further understand the data, we used it to classify those values on each of the emotions if those are positive, neutral, or negative emotion which also help us understand more the emotions of a text value given on the dataset.

	Text	Emotion	Sentiment
0	i didnt feel humiliated	sadness	Neutral
1	i can go from feeling so hopeless to so damned...	sadness	Neutral
2	im grabbing a minute to post i feel greedy wrong	anger	Negative
3	i am ever feeling nostalgic about the fireplac...	love	Negative
4	i am feeling grouchy	anger	Neutral

Fig. 4 Dataset with Sentiment

Visualizing the dataset with sentiment analysis is also created.

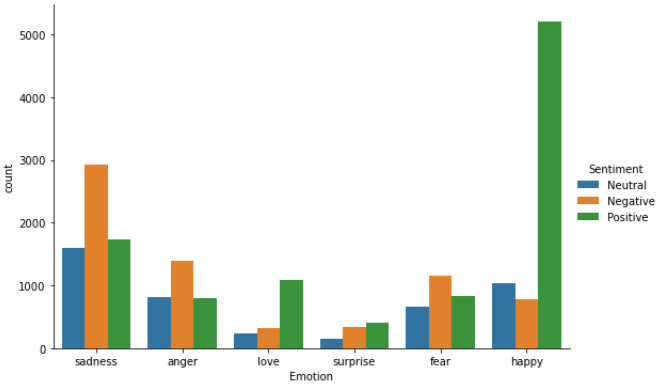

Fig. 5 Visualized Dataset with Sentiment

The emotion happy composed of a sentiment of 781 negative emotions, 1,035 neutral emotions, and 5,213 positive emotions. For anger composed of a sentiment of 1,390 negative emotions, 806 neutral emotions, and 797 positive emotions. For fear composed of a sentiment of 1,153 negative emotions, 665 neutral emotions, and 834 positive emotions. For love composed of a sentiment of 318 negative emotions, 234 neutral emotions, and 1,089 positive emotions. For sadness composed of a sentiment of 2,934 negative emotions, 1,592 neutral emotions, and 1,739 positive emotions. And lastly, for surprise composed of a sentiment of 329 negative emotions, 152 neutral emotions, and 398 positive emotions.

Table 2 Sentiment analysis for each emotions

Dataset		
Emotion	Sentiment	Values
happy	Negative	781
	Neutral	1,035
	Positive	5,213
sadness	Negative	2,934
	Neutral	1,592
	Positive	1,739
anger	Negative	1,390
	Neutral	806
	Positive	797
fear	Negative	1,153
	Neutral	665
	Positive	834
love	Negative	318
	Neutral	234
	Positive	1,089
surprise	Negative	329
	Neutral	152
	Positive	398

Cleaning the text value of each emotion is also applied by using neattext function. In here, we only remove the stop words, user handles, and punctuations to clear the text value which help our model classify the training and text values clearly with reliable results. The clean text was used for the

model training and testing in this research.

	Text	Clear_Text
0	i didnt feel humiliated	didnt feel humiliated
1	i can go from feeling so hopeless to so damned...	feeling hopeless damned hopeful cares awake
2	im grabbing a minute to post i feel greedy wrong	im grabbing minute post feel greedy wrong
3	i am ever feeling nostalgic about the fireplac...	feeling nostalgic fireplace know property
4	i am feeling grouchy	feeling grouchy
...	...	...
21454	Melissa stared at her friend in dism	Melissa stared friend dism
21455	Successive state elections have seen the gover...	Successive state elections seen governing part...
21456	Vincent was irritated but not dismay	Vincent irritated dismay
21457	Kendall-Hume turned back to face the dismayed ...	KendallHume turned face dismayed coup
21458	I am dismayed , but not surpris	dismayed surpris

Fig. 6 Sampling Text Data Cleaning

We also applied keywords Extraction and identify the Most Common Words and Word Clouds on each of the text values assigned on each of the emotions. The keywords, most common words, and word clouds are used for our model to predict a given text efficiently by only determining the words on the text and provide the results based on the words assigned on each of the emotions categories.

	Text	Emotion	Sentiment	Clear_Text
8	i have been with petronas for years i feel tha...	happy	Positive	petronas years feel petronas performed huge pr...
11	i do feel that running is a divine experience ...	happy	Neutral	feel running divine experience expect type spi...
14	i have immense sympathy with the general point...	happy	Positive	immense sympathy general point possible proto ...
15	i do not feel reassured anxiety is on each side	happy	Neutral	feel reassured anxiety
22	i have the feeling she was amused and delighted	happy	Positive	feeling amused delighted
...	...	...	...	...
20263	He uttered a short sharp bark , which made Ros...	happy	Negative	uttered short sharp bark Rosie jump assumed ...
20264	He considered this thoughtfully , then a gleam...	happy	Positive	considered thoughtfully gleam amusement came e
20265	A look of intense amusement crossed Catherine ...	happy	Positive	look intense amusement crossed Catherine s f
20266	As a toddler she filled in concentric circles ...	happy	Positive	toddler filled concentric circles provided amu...
20267	A smile of amusement played on his lips as he ...	happy	Positive	smile amusement played lips studied unkempt pr

Fig. 7 Sample Keywords Extraction

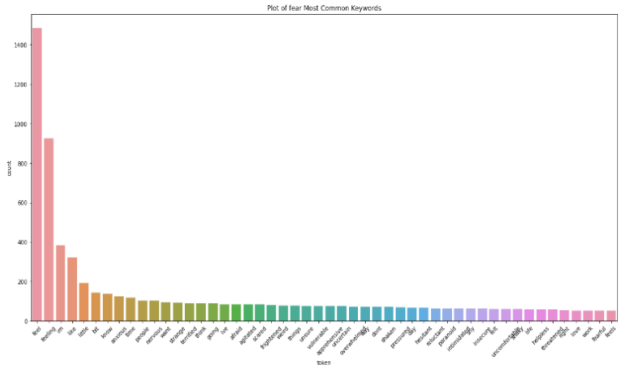

Fig. 8 Visualization sample for Most Common Words

Fig. 9 Word Clouds Visualization Sample

2.4. Model Building

In building the model, we applied the feature extraction to our clean text and emotion of the dataset we prepared on the previous steps. We used Count Vectorizer to fit transform the clean text features, we declared Xfeatures variable for text value column and ylabels for emotion value column. We also split the dataset into training and testing set in which 70 percent are assigned for training and 30 percent are for testing set.

This research applied the Naïve Bayes and Logistic Regression Model in predicting the emotion in text. We train the models using the training set and identify the prediction accuracy of each model using testing set. For Naïve Bayes Model, the accuracy predicted is 77 percent while Logistic Regression Model predicted 89 percent of accuracy.

We test each model's prediction accuracy and probability also by simply predicting a given text such as "I love coding so much" and "I hates running all day". For Naïve Bayes Model, we have a prediction of happy for "I love coding so much" text and sadness for "I hates running all day" with a probability of 0.45 and 0.46 respectively. For Logistic Regression Model, we have a prediction of happy for "I love coding so much" text and sadness for "I hates running all day" with a probability of 0.31 and 0.36respectively.

Table 3 Naïve Bayes Model Prediction Probability

Text "I love coding so much"	
Emotion	Probability
happy	0.4508887894621135
sadness	0.2598462520721136
anger	0.09148145056640933
fear	0.07261883660842203
love	0.11332292297668344
surprise	0.011841748314257881

Table 4 Naïve Bayes Model Prediction Probability

Text "I hate running all day"	
Emotion	Probability
happy	0.3657676231329899
sadness	0.4642699697224394
anger	0.05851818494093467
fear	0.06264856620407458
love	0.044350427048113435
surprise	0.004445228951447268

Table 5 Logistic Regression Model Prediction Probability

Text "I love coding so much"	
Emotion	Probability
happy	0.3114429622929565
sadness	0.2601709369108662
anger	0.19118232483859768
fear	0.13391513402359687
love	0.04702496986946897
surprise	0.05626367206451396

Table 6 Logistic Regression Model Prediction Probability

Text "I hate running all day"	
Emotion	Probability
happy	0.35795139454597946
sadness	0.2895873083603611
anger	0.1826485677710679
fear	0.1184829889204575
love	0.016416841638086658
surprise	0.03491289876404745

Prediction of both models are good and accurate. If see the accuracy of prediction, Logistic Regression is higher in accuracy in prediction than Naïve Bayes Model, but you can see it's confusing because when we look at the probability of each text given, Naïve Bayes Model is much higher probability than Logistic Regression Model. To further understand the results of both models, model evaluation is applied on the next part.

Model Evaluation

To further understand more the results made during the model building prediction with probability, we expand the analysis using classification report and confusion matrix.

In classification report, we analyse the prediction thru precision and recall of each of the emotions obtain. But before that let's look at the confusion matrix first as it is the way we can clearly understand and analyse more the classification of the data distributed to each of the emotions during the prediction.

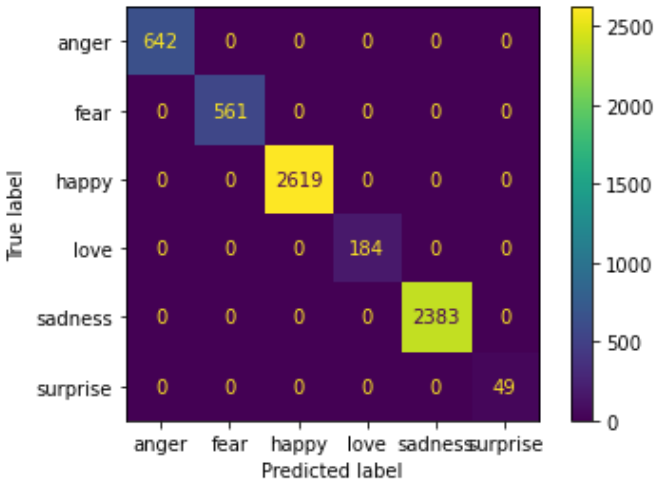

Fig. 10 Naïve Bayes Confusion Matrix

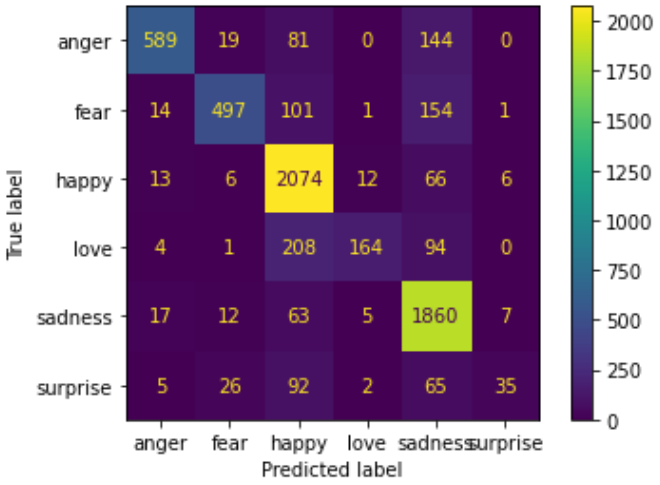

Fig. 11 Logistic Regression Confusion Matrix

When we compared those two matrices, we can see the difference. For Naïve Bayes, the data point distributed for each of the emotions are in freeze, the true label and predicted label is equal and there's no true value predicted to be in the other true value. While in Logistic Regression we can see that for each true value data point distributed there are predicted on the other true values like for example the data point for true label anger is 883 those predicted as anger is 589, other values of anger predicted as fear which is 19, happy is 81, and sadness is 144 values. By looking at the keywords and most common words on a text, logistic regression predicted those words on the other emotions label, which is good because, in a text there are words that used to be anger but based on the combination of words or let's say sentence the text become happy depending on the situation. That is the reason why Logistic Regression provided a highest accuracy in prediction than Naïve Bayes even though the probability is lower than Naïve Bayes, but it accurately predicted the emotions by that way.

Since we know now the difference of both models, let's try to see the classification report to further understand the confusion matrix we have.

Table 7 Naïve Bayes Classification Report

Naïve Bayes (Accuracy = 0.77)			
Emotion	Precision	Recall	F1-score
happy	0.74	0.94	0.83
sadness	0.75	0.93	0.83
anger	0.90	0.64	0.75
fear	0.84	0.59	0.70
love	0.84	0.32	0.46
surprise	0.82	0.16	0.26

Table 8 Logistic Regression Classification Report

Logistic Regression (Accuracy = 89)			
Emotion	Precision	Recall	F1-score
happy	0.89	0.94	0.91
sadness	0.91	0.92	0.92
anger	0.92	0.84	0.88
fear	0.87	0.84	0.85
love	0.82	0.79	0.80
surprise	0.81	0.72	0.77

As expected, when we evaluated the confusion matrix of both models, Logistic Regression model is good for predicting emotions in a text since it shows in the table that it has a highest precision and recall obtain than the Naïve Bayes model.

The maximum score obtained using Naïve Bayes is only 83 percent and a minimum of 26 percent while in Logistic Regression we have a maximum of 92 percent and a minimum of 77 percent which give the Logistic Regression as best model to utilized in prediction of emotion in text than Naïve Bayes.

There are different percentage for each emotion labelled because the dataset is imbalance, we have thousands of data under happy and others while hundreds for surprise as shown in table 1, that is why we obtain a minimum percentage on surprise emotion values and also a maximum percentage for happy and sadness because the values for those two emotions is higher than the others.

3. Results

For predictive model, Logistic Regression provided more accurate prediction than Naïve Bayes as visible in the methodology of this paper. From the sample text prediction given, Naïve Bayes provided a highest probability of 46 percent in prediction that Logistic Regression that has a highest probability only of 36 percent, but Logistic Regression got an accuracy of 89 percent while Naïve Bayes is only 77 percent when we talk about the accuracy of prediction. We have doubt about which models is best since we confused on the results of probability and accuracy of both model, but we do the evaluation to further address that confusion in which we applied the evaluation using confusion matrix and classification report.

For confusion matrix, Naïve Bayes True Positive value predicted as ease True Positive value in prediction, there is no False Positive value as shown in figure 10 while Logistic Regression there is a False Positive value in prediction as shown in figure 11. Naïve Bayes prediction is accurately when we see the confusion matrix results, but we are dealing with the emotion in text in which there is word that can be classify into a different emotion depending on what form of sentences it was and Logistic Regression is classifying that words in different emotions that is why it got a highest accuracy than Naïve Bayes. Classification report as shown in figure also tells that Logistic Regression is much more accurate than Naïve Bayes in prediction of emotion in a text based on the accuracy, precision, and recall obtain of that model.

The dataset for this paper is good for the purposed training a model to find the best for prediction of emotion in a text. In reality, it should be added some text for the other emotions as there is an imbalance of the data for each emotion. We can have a highest prediction for happy and sadness when using this dataset because happy and sadness has more values than the other emotions class and there are times that when we know that the emotion in a text is surprise or love sometimes it predicted as happy because of the imbalance data.

For the model, we can apply Logistic Regression rather than Naïve Bayes as we already know the difference shown in this paper. For the dataset, for the training purposes it is enough but when apply into reality the dataset is not enough for prediction of emotion in text.

4. Conclusion

Predictive model for emotion in text is accurately using Logistic Regression than Naïve Bayes. In this paper, we suggested to used Logistic Regression for predicting emotion in a text, but you can try another model that you think is best with high accuracy and prediction than Logistic Regression. You can see the difference of those two models in the methodology under model evaluation, so it means that different models applied also has a different output which you need to evaluate clearly so that you provided the results as expected to what you want to do.

The dataset is not totally enough because there is an imbalance data for each of the emotions. We obtain a minimum of 26 percent and 77 percent which not good when we are going to provide a more accurate and efficient emotion in a text prediction. To further increase the accuracy of prediction on the other emotions, the dataset on label emotion like surprise should be added some more data and also the others to level the highest data on the dataset for it to be balance and the prediction is much more accurate and efficient.

Overall, dataset was good if it is used only for training purposes and Logistic Regression is the best model in this

paper with an accuracy of 89 percent and we recommended that model rather than Naïve Bayes model.

References

[1] Singhal P, Muljono, N. A. S. Winarsih and C. Supriyanto, "Evaluation of classification methods for Indonesian text emotion detection," 2016 International Seminar on Application for Technology of Information and Communication (ISemantic), 2016, pp. 130-133, doi: 10.1109/ISEMANTIC.2016.7873824.

[2] S. Chaffar and D. Inkpen, "Using a Heterogeneous Dataset for Emotion Analysis in Text," in *Advances in Artificial Intelligence*, Springer Berlin Heidelberg, 2011, pp. 62-67

[3] N. Chirawichitchai, "Emotion Classification of Thai Text based Using Term weighting and Machine Learning Techniques," in *11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2014

[4] P. Inrak and S. Sinthupinyo, "Applying Latent Semantic Analysis To Classify Emotions In Thai Text," in *2010 2nd International Conference on Computer Engineering and Technology (ICCET)*, 2010

[5] W. Li and H. Xu, "Text-based emotion classification using emotion cause extraction," *Expert Systems with Applications journal*, vol. 41, p. 1742–1749 , 2014

[6] J. Li, Y. Xu, H. Xiong and Y. Wang, "Chinese Text Emotion Classification Based On Emotion Dictionary," in *2010 IEEE 2nd Symposium on Web Society*, 2010

[7] A. Z. Arifin, Y. A. Sari and E. K. Ratnasari, "Emotion Detection of Tweets in Indonesian Language using Non-Negative Matrix Factorization," International Journal of Intelligent Systems and Applications, 2014.

[8] Arifin and K. E. Purnama, "Classification of Emotions in Indonesia Texts using K-NN Method," International Journal of Information and Electronics Engineering, vol. 2, no. 6, 2012

[9] M. Sunghwan, "Recognising Emotions and Sentiments in Text," Thesis, Dept. Elect. and Inform. Eng., University of Sydney, Sydney, Australia, 2011.

[10] Acheampong, FA, Wenyu, C, Nunoo-Mensah, H. Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports. 2020; 2:e12189. https://doi.org/10.1002/eng2.12189

[11] Shanmugam, S. P. ., Vadivu, M. S. ., Anitha, D., Varun, M., & Saranya, N. N. . (2023). A Internet of Things Improvng Deep Neural Network Based Particle Swarm Optimization Computation Prediction Approach for Healthcare System. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4s), 92–99. https://doi.org/10.17762/ijritcc.v11i4s.6311

[12] Mwangi , J., Cohen, D., Silva, C., Min-ji, K., & Suzuki, H. Improving Fraud Detection in Financial Transactions with Machine Learning. Kuwait Journal of Machine Learning, 1(4). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/148

[13] Beemkumar, N., Gupta, S., Bhardwaj, S., Dhabliya, D., Rai, M., Pandey, J.K., Gupta, A. Activity recognition and IoT-based analysis using time series and CNN (2023) Handbook of Research on Machine Learning-Enabled IoT for Smart Applications Across Industries, pp. 350-364.