SINERGI Vol. 24, No. 2, June 2020: 87-94
http://publikasi.mercubuana.ac.id/index.php/sinergi
http://doi.org/10.22441/sinergi.2020.2.002 .

SENTIMENT ANALYSIS ON TWITTER BY USING MAXIMUM ENTROPY AND SUPPORT VECTOR MACHINE METHOD

Mona Cindo, Dian Palupi Rini*, Ermatita
Graduate Schoool of Computer Sciences, Universitas Sriwijaya,
Bukit Besar, Palembang 30662, Indonesia
*Corresponding Author Email: dprini@unsri.ac.id

Abstract -- *With the advancement of social media and its growth, there is a lot of data that can be presented for research in social mining. Twitter is a microblogging that can be used. In this event, a lot of companies used the data on Twitter to analyze the satisfaction of their customer about product quality. On the other hand, a lot of users use social media to express their daily emotions. The case can be developed into a research study that can be used both to improve product quality, as well as to analyze the opinion on certain events. The research is often called sentiment analysis or opinion mining. While The previous research does a particularly useful feature for sentiment analysis, but it is still a lack of performance. Furthermore, they used Support Vector Machine as a classification method. On the other hand, most researchers found another classification method, which is considered more efficient such as Maximum Entropy. So, this research used two types of a dataset, the general opinion data, and the airline's opinion data. For feature extraction, we employ four feature extraction, such as pragmatic, lexical-grams, pos-grams, and sentiment lexical. For the classification, we use both of Support Vector Machine and Maximum Entropy to find the best result. In the end, the best result is performed by Maximum Entropy with 85,8% accuracy on general opinion data, and 92,6% accuracy on airlines opinion data.*

Keywords: *Microblogging; Twitter; Support Vector Machine; Maximum Entropy; Feature Extraction*

Received: June 28, 2019 *Revised: February 5, 2020* *Accepted: February 10, 2019*

INTRODUCTION

Social media is an important part of the daily routine for people at this time. Social media and the internet can be used for a variety of things, such as advertising, spread the political opinion and financial trends, getting user comments about products review, spread spam, spreading the news [1]. Social media create a virtual bond between users, which people express the opinion and develop the relationships through the posts, comments, messages, and likes. There are hundreds of social media channels operating throughout the world today, with three majors: Facebook, LinkedIn, and Twitter. Twitter is the most popular microblog in Indonesia. This microblog allows users to send and read the message called tweets, in the form of a maximum of 140 characters of text displayed on the user profile page [2].

Twitter allows people to express their emotions, feelings, share their thoughts, opinions with others instantly and easily. To detect the emotional sentences in online media content can be done by analyzing the sentiments of the users conveyed through messages in the social networking site Twitter. Sentiment analysis is the process of understanding, extracting, and automatically processing the textual data to obtain the information [3] so that this sentiment analysis can be used to get someone's emotional information contained in the messages of users of Twitter social networks on the topics discussed by users.

In the previous study, the discussion about the sentiment analysis on an emotion such as [4] has added the features in a set of sentiment analysis using Naïve Bayes (NB), Maximum Entropy (MaxEnt) and Support Vector Machine (SVM). Kouloumpis et al. evaluating the use of the existing lexical resources as well as features that capture information about informal and creative languages used in microblogging by simply utilizing the hashtag on Twitter data to build training data [5].

Emoticons are used for users as an easy way to express emotions briefly. Research [6] discusses sentiment analysis to detect significant emotional changes in extracting information about the polarity of user sentiments (positive or negative). Support Vector Machine results provide the highest accuracy. Adding features to sentiment analysis such as pragmatic features

such as removing URLs and tags ("@user") associated with hate expressions and using them as additional features for hate detection with other features for detecting hate speech in short text messages on Twitter can make it easier to detect expressions of hatred [7]. We use pragmatic features with additions that combine with counting punctuation, capital letters, hashtags, @symbols, and emoticons. We also use Lexical-gram, POS-gram, and Lexical Sentiment features to emphasize one's emotions. by comparing the two Support Vector Machine methods and the Maximum Entropy method to get the best results

MATERIAL AND METHOD

The following are five steps for classifying tweets as sentiment analysis, as shown in Figure 1. The first step is collecting data and followed by preprocessing, feature extraction, classification, and the last step is evaluation.

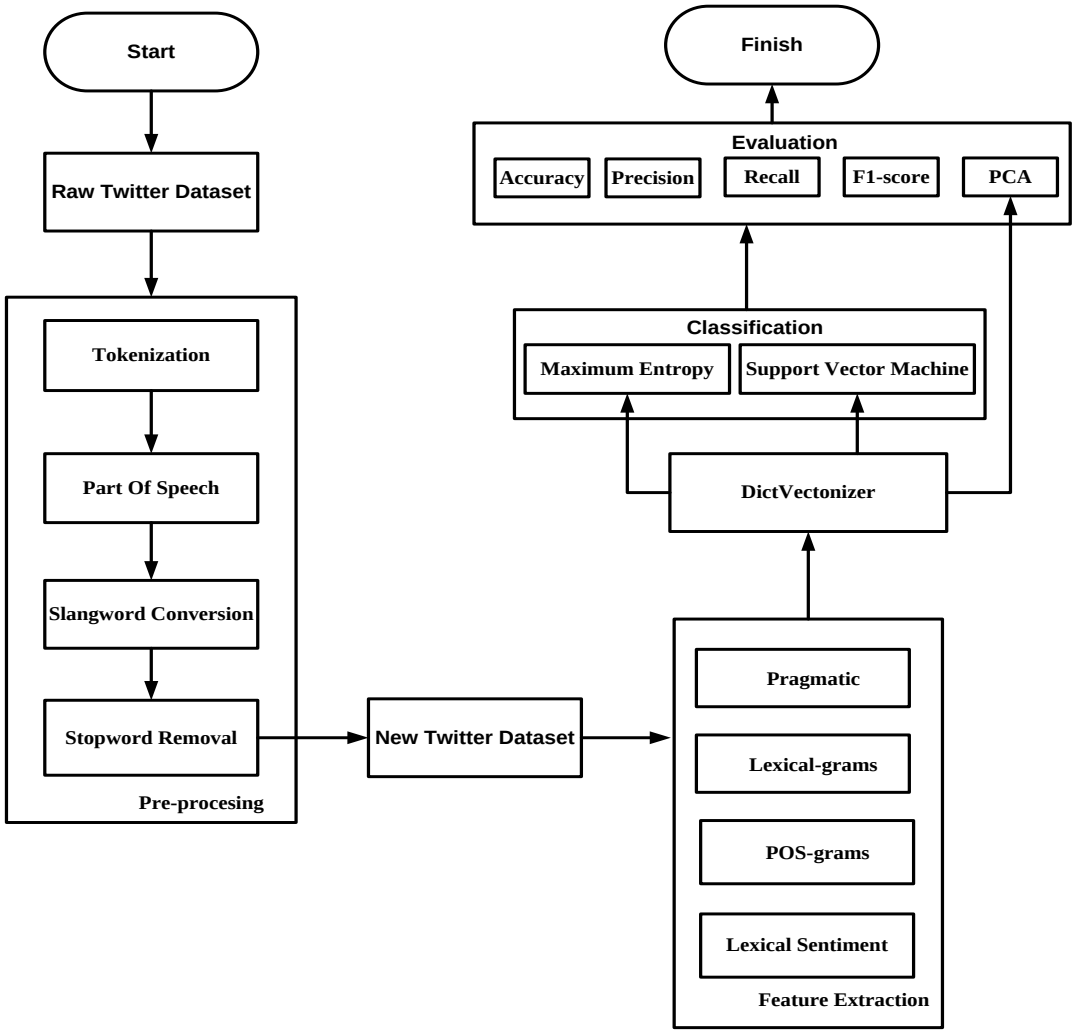

Figure 1. Sentiment Analysis Architecture

Data Collection

Data collection is one of the important keys to the knowledge discovery process. In this research, we used two different datasets. The general dataset we got from collect Twitter by using Twitter API and the open-source airline's opinion Twitter data collected by the internet.

The Distribution of the Dataset is shown in Table 1.

Table 1. Twitter Dataset Distribution

Dataset	Total	Sumber
General Opinion	10308	Twitter API
Airlines Opinion	2812	https://www.figure-eight.com/data-for-everyone/

Preprocessing Data

The purpose of the preprocessing to simplify retrieving features by deleting and changing the data that does not need to be easy to process data. The stage of preprocessing consists of tokenization, part-of-speech, slang words, and stop word removal.

Feature Extraction

Feature extraction aims to recognize the characteristics of a Twitter sentence and make it as a feature. In general, this study uses four feature extraction, namely pragmatic, lexical-grams, pos-grams, and sentiment lexical.

- **Pragmatic**

Twitter is microblogging. Users are allowed to share their opinions in the form of tweets, using only 140 characters [8]. Thus, causes limited users to express emotion so that some users use some punctuation to show their emotions [9]. This feature calculates punctuation marks such as the use of capital letters, hashtags, tags (@), exclamation marks, emoticons, and negation sentences such as "no," "none," "never," etc. [10].

- **Lexical-grams**

Lexical-gram (n-grams) is the feature that often used for the text mining process. It is a set of the word that appears together in a text [11]. This feature can be used by using tokenization in the python library to be able to solve the sentences.

- **POS-grams**

Part-of-speech is a class giving process to a word by dividing the sentences or paragraphs into words [12]. The first process is obtained during preprocessing by using CMU tagger. A sentence is broken down into classes such as adjective (D), nouns (N), and verbs (V). then the words are breaking down and made into features such as lexical-grams feature.

- **Lexical Sentiment**

This feature utilizes the lexical SentiWordNet dictionary, which has a sentiment value for each word [13]. The tokenization process is needed to get a value on tweets, and then selected some words needed to be weighted sentiment values such as v (verb), n (noun), d (adjective), and r (adverb). After each word is given a sentiment value, the value is then in total to get the final result from the sentiment per tweet.

Classification Methods

This study uses two classification methods for comparison while getting the best classification results.

- **Support Vector Machine**

When analyzing data, SVM determines decision boundaries and uses the kernel to perform input calculation [14]. We use the linear kernel that can be defined as:

$$f(x) = \beta(0) + \text{sum}\,(a_i * (x, x_i)) \quad (1)$$

For linear kernel calculation, the prediction from the new input use product point between input (x) with all supported vector from training data. While for coefficient *β(0)* and a_i for each input must be estimated from training data.

- **Maximum Entropy**

In the Maximum Entropy classification. There are no assumptions used in the relationship between features. The method aims to maximize entropy in the system by predicting the condition distribution of labels in each class. This classification handles overlapping features, such as logistic regression. This distribution is then defined as MaxEnt, which does not make any assumptions on its features [15].

MaxEnt can be defined as:

$$P_{ME}(c \mid d, \lambda) = \frac{\exp[\sum_i \lambda_i f_i(c,d)]}{\sum_c \exp[\sum_i \lambda_i f_i(c,d)]} \quad (2)$$

Where c is a class, b is twitter, and λ is vector weight.

Evaluation

The last step is evaluation. In this part, we use several matrix and graph to analyzing the feature extraction and classification. A confusion matrix is a useful tool for analyzing how well the classifier recognizes tuples from different classes. True Positives (TP) and True Negatives (TN) provide information when the classifier is true. In contrast, False Positives (FP) and False Negatives (FN) tell when the classifier is wrong [16].

We use the confusion matrix to calculate the precision, recall, f1-score, and the accuracy of the classification. Also, we use the principal component analysis (PCA) to see each of the data distribution. PCA determines a smaller set of artificial variables that will represent the variance of a series of observed variables the calculated artificial variable is called the main components. The main component is used as a predictor variable or criterion in another analysis [17],[18].

RESULTS AND DISCUSSION

The data is consisting of 5101 positive tweets and 5207 negative tweets. The total of the general dataset is 10308 tweets. From the preprocessing process, the clean dataset was obtained. Also, it is generally given part-of-speech tags. After that, the data was divided into 9277 training data and 1031 testing data. To do the evaluation, we divided the data into 10 different folds. All the fold data were evaluated and calculated the average score of the accuracy. Figure 2 shown the evaluation data.

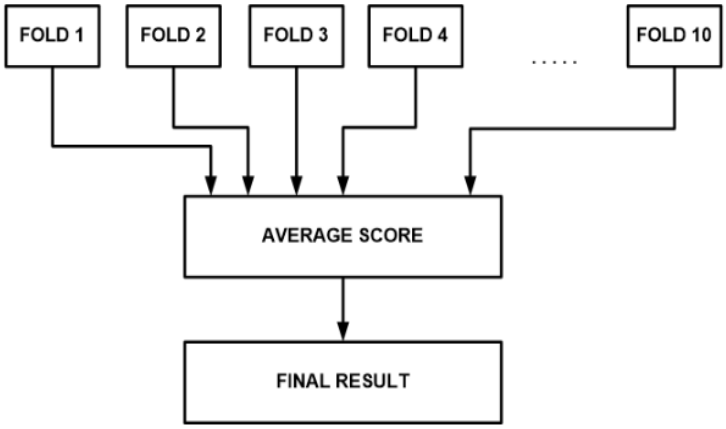

Figure 2. Evaluation of Data Flow

On the other hand, this research also provides an airline opinion dataset for data comparison. The total second data is 2812.

Sentiment Analysis Result

In this study, to get the best result of the classification, we use two classification methods as a comparison method. The result of the evaluation of 10 fold data form the general dataset is shown in Table 2.

Table 2. Accuracy Result from General Dataset

Fold	*SVM*	*MaxEnt*
1	71%	80,1%
2	81,4%	**89,5%**
3	75%	85,2%
4	**82,3%**	89,4%
5	76,2%	87,7%
6	75,9%	84,1%
7	74,8%	82,8%
8	77,2%	87,7%
9	74,3%	84,4%
10	78%	87,1%
Average	76,61%	85,8%

Based on Table 2, we found the best result by using SVM classification is in fold 4 with 82,3% accuracy. On the other hand, Maximum Entropy got the best result in fold 2. Also, based on average score, Maximum Entropy has a higher level of accuracy compare to SVM. Based on the evaluation result, it can be concluded that the best result is in fold 4. In fold 4, each deployment of features in the training data and testing data, can be seen in Figure 3 and Figure 4.

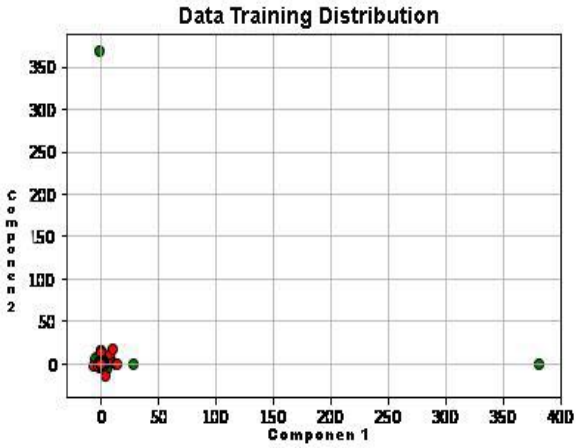

Figure 3. PCA the deployment of training data

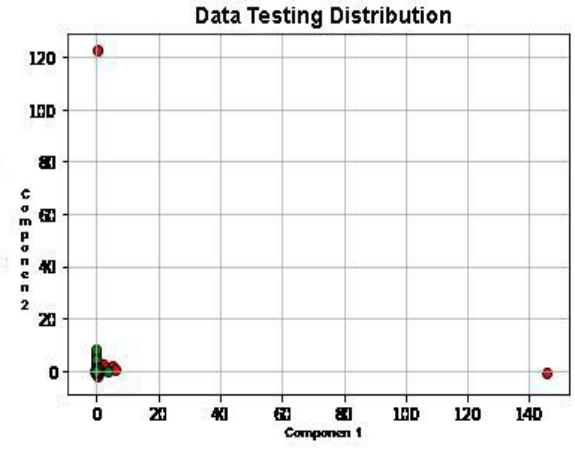

Figure 4. PCA the deployment of testing data

The following picture above shows the distribution of training and testing data. The green point shows positive data, and the red point indicates negative data. Apart from accuracy, the other matrix also needs to be observed, such as the confusion matrix. With the confusion matrix, we can observe how many predicted positive/negative data and how many data that failed to predict.

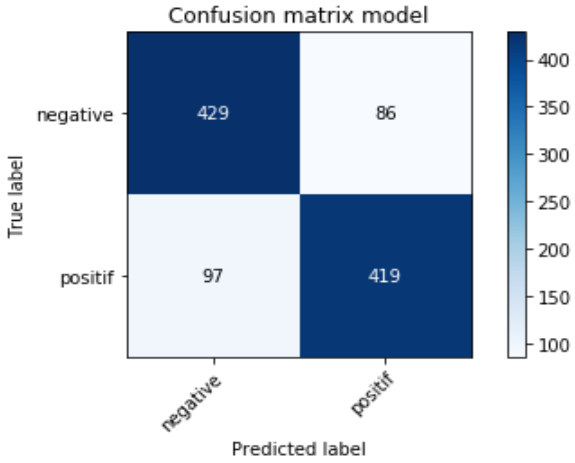

Figure 5. Confusion Matrix Model SVM

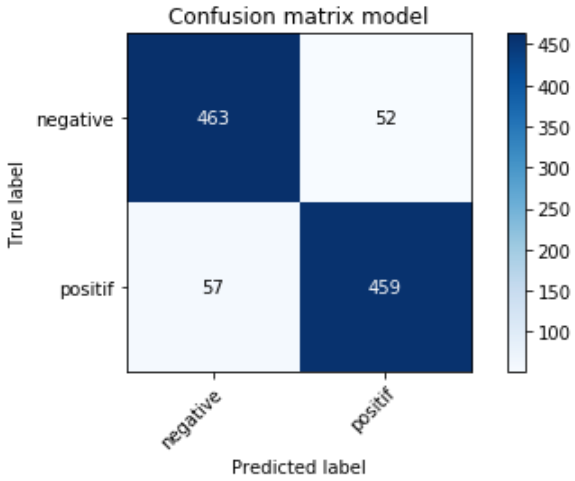

Figure 6. Confusion Matrix Model MaxEnt

Figure 5 and Figure 6 show the model of the confusion matrix. Based on the model above, it can be used to calculating the evaluation matrix. The result can be seen in Table 3.

Table 3. Evaluation Matrix Result

		Precision	Recall	F1-Score
SVM	Negative	82%	83%	82%
	Positive	83%	81%	82%
Max Ent	Negative	89%	90%	89%
	Positive	90%	89%	89%

For comparison, this research also provides the other different datasets. The airline's opinion dataset can get from the open-source sentiment analysis dataset website. The data is consisting of 2812, which divided into 2530 training data and 282 testing data. The feature data deployment model in this dataset can be seen in Figure 7 and Figure 8.

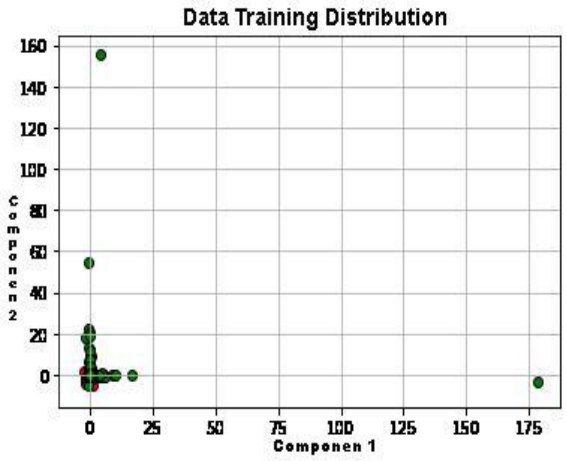

Figure 7. PCA data training deployment for general opinion

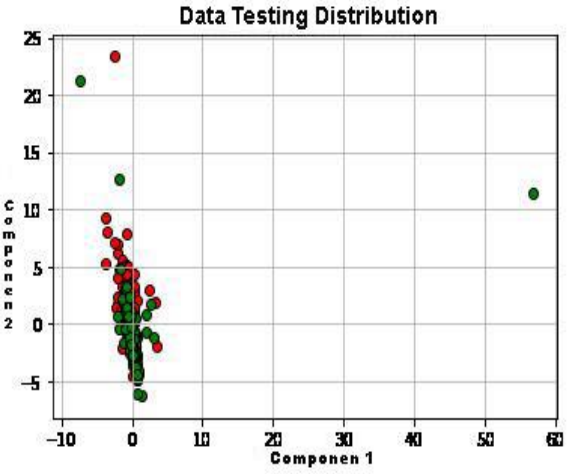

Figure 8. PCA data testing deployment for general opinion

From the data deployment, the classification is done in Figure 9 and Figure 10.

The following is shown in Table 4, the confusion matrix result used as the basis of the evaluation matrix for airline opinion datasets.

Table 4. Accuracy Result from Airline Dataset

	Accuracy	Precision	Recall	F1-Score
SVM	86,9%	87,0%	86,9%	86,9%
Max Ent	92,6%	92,8%	92,6%	92,5%

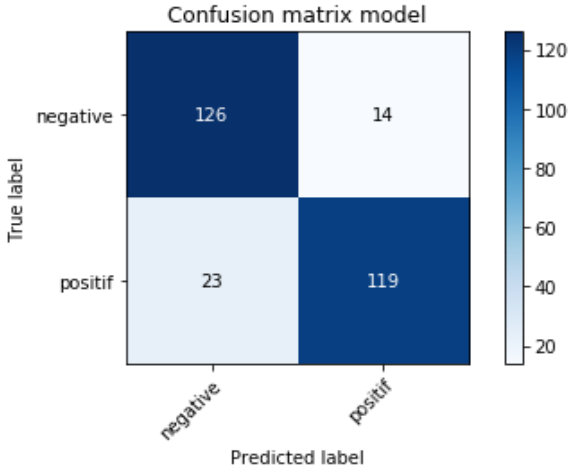

Figure 9. Confusion Matrix Model SVM airline dataset

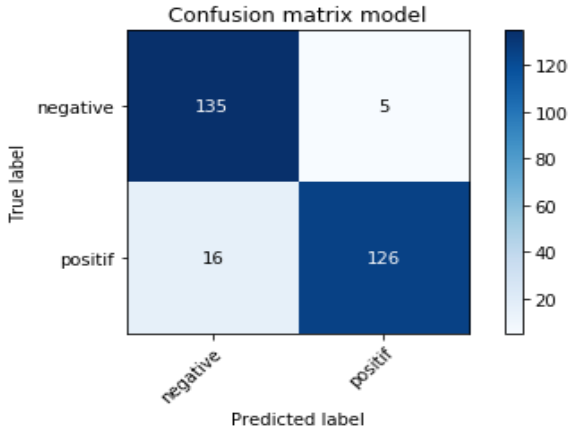

Figure 10. Confusion Matrix Model MaxEnt airline dataset

Based on Table 4, it shows the airline opinion dataset for the SVM method obtains 86,9% accuracy, and for the Maximum Entropy, the method is 92,2%. The following strengthens the maximum entropy getting the best results in this study.

Feature Extraction Result

This sentiment analysis using four different feature extraction. Each feature was tested to get the most important feature in this experiment. The result of each of feature extraction is shown in Table 5.

Table 5. Feature Extraction Accuracy from Fold 4

Feature Extraction	***SVM***	***MaxEnt***
Sentiment lexical	71%	71.5%
Lexical N-gram	**77.2%**	**86%**
POS-gram	66.2%	66.4%
Pragmatic	62.5%	62.3%

Based on Table 5 it can be said that the best performance is given by the lexical n-grams feature with the best result in the Maximum Entropy Method. The score is 86% accuracy compared to the SVM method, with only 77.2% accuracy.

Table 6. Feature Extraction Accuracy Result from Airline Dataset

Feature Extraction	***SVM***	***MaxEnt***
Sentiment lexical	83.7%	84.0%
Lexical N-gram	**80.9%**	**86.2%**
POS-gram	73%	73.8%
Pragmatic	71.6%	69.9%

On the airline's dataset, the best performance is in the lexical-grams feature, as listed in Table 6. It proves that the lexical-gram feature has a considerable influence on these two datasets. With the lexical-grams feature, the classification can easily recognize the words that are often used to show the emotion seen in Table 7.

Table 7. Sample of Word that Signifies Emotion

UNIGRAM	**LABEL**
LOVE	Positive
HAPPY	Positive
HATE	Negative
SAD	Negative

From the table results and our manual observed, we found the predicted and unpredicted data sample. The following is shown in Table 8, which shows the sample of predicted and unpredicted data.

Table 8. Sample of Predicted and Unpredicted Tweet

Predicted Data	
Positive	happy today Monday went well smiles
Negative	can not sleep! I hate this
Unpredicted Data	
Positive	Going out won't be home until probably late
Negative	I can either watch WNBA or rain delay of Texas baseball. Thank god for the Simpsons.

CONCLUSION

In this paper, the research compares both of Maximum Entropy and Support Vector Machine method. The result shows the Maximum Entropy gives the best result with an average 85,8% accuracy. This result exceeds the Support Vector Machine method with an average 76,6% accuracy. The best performance was given by testing on fold 4 results with 89,4% accuracy. The second dataset strengthens the result with the Maximum Entropy result 92,6% accuracy. Also, this research observed all the feature

extraction that present in this research. This result proves the Maximum Entropy can represent the best number of the different probability distribution for this dataset.

The result showing the lexical-grams is the most important feature. This feature gives the best result compared to the other feature set. It is related to the identification of the words that are clearly showing the speaker's emotion, such as "love," "happy," "hate," and "never," who clearly showing a speaker's emotion. For future work, we need to add a more useful feature like sarcasm detection for better results.

REFERENCES

[1] A. Alarifi, M. Alsaleh, and A. M. Al-Salman, "Twitter Turing test: Identifying social machines," *Information Science*, vol. 372, pp. 332–346, December 2016. DOI: 10.1016/j.ins.2016.08.036

[2] M. Badri, *Komunikasi Pemasaran UMKM Di Era Media Sosial. Corporate and Marketing Communication*, pp. 127-147, Jakarta: Pusat Studi Komunikasi dan Bisnis Program, January 2011.

[3] D. Bandorski et al., "Contraindications for video capsule endoscopy," *World Journal of Gastroenterol*, vol. 22, no. 45, pp. 9898–9908, December 2016. DOI: 10.3748/wjg.v22.i45.9898

[4] H. Saif, Y. He, and H. Alani, "Semantic Sentiment Analysis of Twitter," *CEUR Workshop Proceedings*, vol. 917, pp. 56–66, 2012. DOI: 10.1007/978-3-642-35176-1_32

[5] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," *Proceedings of the Fifth International Conference on Weblogs and Social Media*. AAAI Press, pp. 538–541, July 2011.

[6] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," *Computers in Human Behavior*, vol. 31, no. 1, pp. 527–541, February 2014. DOI:10.1016/j.chb.2013.05.024

[7] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*, vol. 6, pp. 13825–13835, 2018. DOI: 10.1109/ACCESS.2018.2806394

[8] K. Kaviya, K. K. Shanthini, and S. Muthuswamy, "Micro-blogging Sentimental Analysis on Twitter Data Using Naïve Bayes Machine Learning Algorithm in Python-processing Feature Selection Naïve Bayes classification Positive," *International Journal on Future Revolution in Computer Science & Communication Engineering*, vol. 4, no. 4, pp. 46-51, April 2018.

[9] T. Jain, N. Agrawal, G. Goyal, and N. Aggrawal, "Sarcasm detection of tweets: A comparative study," *2017 10th International Conference on Contemporary Computing. IC3*, Noida, 2017, pp. 1–6. DOI: 10.1109/IC3.2017.8284317

[10] M. S. M. Suhaimin, M. H. A. Hijazi, R. Alfred, and F. Coenen, "*Natural language processing based features for sarcasm detection: An investigation using bilingual social media texts*," 2017 8th International Conference on Information Technology (ICIT), Amman, 2017, pp. 703-709. DOI: 10.1109/ICITECH.2017.8079931

[11] P. B. Awachate and V. P. Kshirsagar, "Improved Twitter Sentiment Analysis Using N-Gram Feature Selection and Combinations," *International Journal of Advanced Research in Computer Communication Engineering*, vol.5, no. 9, pp. 154–157, September 2007. DOI: 10.17148/IJARCCE.2016.5935

[12] S. K. Bharti, B. Vachha, R. K. Pradhan, K. S. Babu, and S. K. Jena, "Sarcastic sentiment detection in tweets streamed in real-time: a big data approach," *Digital Communications Networks*, vol. 2, no. 3, pp. 108-121, August 2016. DOI: 10.1016/j.dcan.2016.06.002

[13] C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexicon-based approaches for sentiment analysis of microblog posts," *CEUR Workshop Proceedings*, vol. 1314, pp. 59–68, January 2014.

[14] X. Cheng and H. Shen, "Adaptive Co-Training SVM for Sentiment Classification on Tweets," *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 2079–2088, October 2013. DOI: 10.1145/2505515.2505569

[15] V. A. Kharde and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," *International Journal of Computer Applications*, vol. 139, no. 11, pp. 975–8887, 2016.

[16] A. Khan, B. Baharudin, K. Khan, "Sentiment Classification using Sentence-Level Lexical Based Semantic Orientation of Online Reviews," *Trends Application Scienci Research,* vol. 6, no. 10, pp. 1141-1157, October 2011. DOI: 10.3923/tasr.2011.1141.1157

[17] J. Jotheeswaran, R. Loganathan, and B. MadhuSudhanan, "Feature reduction using

principal component analysis for opinion mining," *International Journal Computer Science Telecommunication*, vol. 3, no. 5, pp. 118–121, 2012.

[18] F. Kherif and A. Latypova, " Chapter 12- Principal componen analysis," *Machine Learning: Methods and Application to Brain Disorder,* 2020, pp. 209-225. DOI: 10.1016/B978-0-12-815739-9.00012-2