Aspect-Based Sentiment Analysis and Emotion Detection for Code-Mixed Review

Andi Suciati[1], Indra Budi[2]
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia

***Abstract*—Review can affect customer decision making because by reading it, people manage to know whether the review is positive, or negative. However, positive, negative, and neutral, without considering the emotion will be not enough because emotion can strengthen the sentiment result. This study explains about the comparison of machine learning and deep learning in sentiment as well as emotion classification with multi-label classification. In machine learning comparison, the problem transformation that we used are Binary Relevance (BR), Classifier Chain (CC), and Label Powerset (LP), with Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Extra Tree Classifier (ET) as algorithms of machine learning. The features we compared are n-gram language model (unigram, bigram, unigram-bigram). For deep learning, algorithms that we applied are Gated Recurrent Unit (GRU) and Bidirectional Long Short-Term Memory (BiLSTM), using self-developed word embedding. The comparison results show RF dominates with 88.4% and 89.54% F1 scores with CC method for food aspect, and LP for price, respectively. For service and ambience aspects, ET leads with 92.65% and 87.1% with LP and CC methods, respectively. On the other hand, in deep learning comparison, GRU and BiLSTM obtained similar F1- score for food aspect, 88.16%. On price aspect, GRU leads with 83.01%. However, for service and ambience, BiLSTM achieved higher F1-score, 89.03% and 84.78%.**

Keywords*—Sentiment analysis; emotion; multi-label classification; machine learning; deep learning

I. Introduction

Review is an evaluation to entities such as product, restaurant, place, etc. that can be used by customers or owner as product input. This review usually contains several aspects such as in laptop [1], the aspects that can be evaluated are hardware, price, etc. This evaluation can affect the decision making from customer. For instance, when people want to go to trip, they will read the review of several places and compare them. One of domain examples that usually get many reviews is restaurant. There are several platforms in internet for restaurant review, such as Zomato[1] and Yelp[2]. In the platform, mostly people only see the ratings of the restaurant, however reading the review is very important because the customers will obtain specific information rather than only seeing the ratings. In addition, sometimes people also give ratings that are very different from the actual review. So, it can be concluded ratings not always give the information about the quality of restaurant. Beside for decision making of customer, review also important for the product owner. Pontiki et al. [2] stated that feedback from customer will help companies measure their customer satisfaction, and for the development of their product and services they provide. For identifying the sentiment of aspect, sentiment analysis can be conducted. However, classifying the sentiment is not enough without considering the emotions from customers. Knowing the emotion can strengthen the sentiment results from a review. Furthermore, mostly a review contains two or more languages, or called code-mixed languages. This kind of review is difficult to understand by computer because computer cannot identify the languages easily like human. This also a big challenge for sentiment analysis and emotion detection. There are several classification methods that can be used, such as machine learning and deep learning. Mohammad et al. [3] used Support Vector Machine when classifying sentiment data from Twitter[3]. In the other hand, Stojanovski et al. [4] applied deep learning algorithm for sentiment analysis and emotion detection for Twitter data.

This research focuses to conduct sentiment analysis an emotion detection in every aspect that appeared in a restaurant review. The data were collected from Indonesian restaurant review platform, named PergiKuliner[4], and this study using 'food', 'price', 'service', and 'ambience' as aspects. The sentiment polarities that were used for emotions are 'positive', 'negative', and 'neutral', while 'happy', 'sad', 'surprised', and 'neutral'. The addition of 'neutral' because there is a possibility that a review contains sentiment polarity, but the emotion is difficult to detect. The method of classification that we applied is multi-label classification while the algorithms that we used are from machine learning and deep learning.

The rest of paper was organized into: in Section 2, we explained about several researches that related to our study. In Section 3, we illustrate the research steps of our experiments. For Section 4, we showed the classification results as well as analyzing them. Then in last part, we concluded the results and future work for this study.

[1] https://www.zomato.com
[2] https://www.yelp.com/
[3] https://twitter.com/home
[4] https://pergikuliner.com/

II. Related Work

There are many studies about sentiment analysis and emotion detection. Mohammad [5] did a literature studies regarding several researches about valence, emotion, and other aspects that can affect the feeling from a person. From that study, the writer describes the challenges for sentiment and emotion detection, such as language complexity, non-standardized language, lack of labeled data, subjectivity, culture differences, etc. Stojanovski et al. [4] did a sentiment analysis research using SemEval 2015[5] and emotion detection using Twitter data. The sentiment polarities that we used are 'positive', 'negative', and 'neutral', while for emotions, we utilized 'love', 'joy', 'surprise', 'anger', 'sadness', 'fear', and 'thankfulness'. After that, the writer applied Deep Convolutional Neural Network for sentiment and emotion detection. However, the sentiment analysis and emotion detection were conducted in separated dataset. Another study about emotion was conducted by Hassan et al. [6]. This study was emotion classification using Skip-thought Vector. Khawaja et al. [7] also did an experiment about emotion which is developing an automatic lexicon for emotion.

In Indonesia, there are also few researches about sentiment and emotion. Wikarsa dan Tahir [8] studied about emotion detection using data from Twitter, but the data were in English. Savigny and Purwarianti [9] also conducted emotion classification using YouTube [6] comments. For sentiment analysis, [10][11] studied it for restaurant review in Indonesia.

Several studies also have conducted for sentiment analysis and emotion detection using code-mixing data. Shalini et al. [12] studied sentiment analysis for Facebook[7] comments with Kannada-English languages. The experiment was done by applying Facebook's fast text, Doc2Vec with SVM, Bidirectional LSTM, and CNN. Lee and Wang [13] experimented using Chinese-English data and proposed multi-learning framework for emotion detection.

III. Research Steps

This section explains the methodology that applied in this research as shown by Fig. 1.

A. Data Collection

The data were collected from PergiKuliner platform by scraping them. The collected data are the reviews for several restaurants in Jakarta, Bogor, Depok, Tangerang, and Bekasi, and the total are 20000 reviews. After filtering the data, such as deleting the duplicate and removing the spam reviews, the final data that annotated are18908 reviews. The data were including reviews that use Indonesian, English, and code-mixed (Indonesian-English). Below are the examples of data:

1) Indonesian: Akhirnya cobain taichan sm martabak tipkernya Dann taichannya enak!! Hehehe Asik jg tmptnya rame. (Finally, can taste its thaichan and martabak tipker and the taichan was delicious!! Hehehe it was fun, the place also crowded.).

2) Mixed: Finally got to try this current happening Korean food! Gyeran Jim (22k) Ini kaya steamed egg, yang rada di bake. Telornya ga tawar, tasty dan pinggirannya agak kering gitu. Menurut gue worth sih 22k buat ini, hehe. Probably gonna try again :) (Finally got to try this current happening Korean food! This Gyeran Jim (22k) was like steamed egg. The egg wasn't blend, tasty and the crust is bit dry. In my opinion 22K was worth for this, hehe. Probably gonna try again :))

3) English: Been here for several times I've been loving this place so much. The ambience is truly Japanese izakaya dining. If you eat with many people (sharing) the price would be reasonable, however if you only eat for two the price might get a little high for izakaya. Though the foods are mostly great. Cool place to hangout!

B. Building Annotation Guidelines

After collecting data, next step is building the annotation guidelines. There are two annotation guidelines that were made. First is annotation guideline for sentiment annotation, and another one is for emotion annotation. The aspects that used *'food', 'price', 'service',* and *'ambience'*. The sentiment polarities that used, following [14], which are *'positive', 'negative'*, and *'neutral'*, while for emotions, we followed [15], that divided emotions into *'happy', 'sad', 'surprised', 'angry', 'disgusted',* and *'fear'*. We also added *'neutral'* for emotion list because the possibility if the emotion is difficult to detect. Below are the definitions of the label that used.

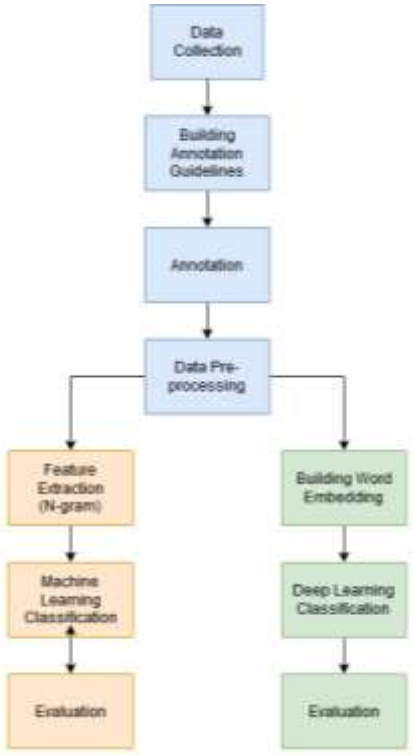

Fig. 1. Research Steps.

[5] http://alt.qcri.org/semeval2015/
[6] https://www.youtube.com
[7] https://www.facebook.com/

1) Sentiment labels:

a) Positive: Positive value can be seen by the appearance of positive terms, such as: "delicious", "recommended", "cheap", "clean", "friendly", etc.

b) Negative: Negative label is given if the negative terms occur, for instance: "bad", "horrible", "not recommended", "pricey", "expensive", "dirty", etc.

c) Neutral: A review is classified as neutral if the terms that appear do not show positive or negative values. Besides, it can be noticed by the appearance of neutral terms, such as: "standard", "so so", "not bad but not good", etc. In addition, the neutral label also given to the aspect that does not appear, because we assumed if an aspect does not mentioned, that means the polarity will be neither positive nor negative.

2) Emotion labels:

a) Happy: Happy emotion can be noticed by the appearance of phrases or words like: 'I like it', 'really good', 'happy', 'satisfied', 'cool', 'worth', 'fun', or emoticon ':)', ':D', etc.

b) Sad: Sad emotion shows the sadness or dissapointed, anc can be known by the appearance of terms 'sad', 'dissappointed', 'below expectation', or with emoticon ":(", ":'(".

c) Surprised: Surpised can be noticed by the terms like 'I'm surprised', 'beyond expectation', 'shock', etc.

d) Angry: Few terms that can be considered to label data as angry are 'damn', 'angry', 'annoyed', 'annoying', etc.

e) Disgusted: Disgusted emotion can be classified by the appearance of terms 'dirty', 'disgusted', etc.

f) Fear: Review is classified as fear if the terms like 'afraid', 'worried', etc, appears.

g) Neutral: Neutral label in is given if the emotion in a review difficult to be interpreted. In addition, neutral emotion also will be given even though the aspects are not mentioned, like neutral definition in sentiment.

C. Annotation

The next step is annotating the data. The annotation step consists two stages, which are sentiment annotation and emotion annotation. The method for deciding the annotator is crowdsourcing method, following a study from Sabou et al. [16]. The annotators are not linguistic experts. Besides, every review is annotated by 3 people in every stage. The method for retrieving the final label is major voting. After sentiment annotation, there are 562 data that cannot be used because the major voting results indicated that every annotator has labelled them with different labels. So, the data for the next annotation stage are 18346 reviews. However, because the limited time and number of annotators, the data that annotated for emotion label are only 15046 reviews. After applied major voting, the results of data that used are 14188. But the number of data with 'angry', 'fear', and 'disgusted' labels are very small, so we decided to remove those data, and the final number of data that we used for classification are 14103 reviews. Then, the labels that used are 'positive', 'negative', and 'neutral' for sentiment, while 'happy', 'sad', 'surprised', and 'neutral' for emotion.

D. Data Preprocessing

After the annotation process, the next stage is data preprocessing. This stage adapted the research from [17] and consists few steps, which are:

1) Emoticon Processing: In this step, emoticon characters, such as: :(was changed into 'sad', and :) into 'happy'. This was conducted to avoid losing the information about the emoticon. Furtherore, when removing non alphabetical characters step is applied, the emoticon is not removed.

2) Case Folding: All of strings were changed into lowercase format to match the structures. For example, 'Food' was converted into 'foods'.

3) Abbreviation and Spelling Correction part 1: In this part, the word spelling was corrected into formal form. For illustration, 'I've visted the place, that wasn't too crowd' was corrected into 'i have visted the place, that was not too crowd'. We used the abreviation dictionary that id self-developed by [17], and contains abreviations from indonesian and english.

4) Removing Non-alphabetical Characters: After normalizing the words, then the non alphabetical characters, such as '.', '!', '@', etc, are removed in this step.

5) Abbreviation and Spelling Correction part 2: In this step, the words are checked again whether all of them have been corrected. This step was applied to avoid the words that has the possibilities haven't been corrected in the third step. For instance, the phrase 'tmptnya ga bgs!!' was changed into 'tempatnya tidak bgs!!' after third step, but the word 'bgs' does not change into 'bagus' (good) because there are exclamation marks '!!' that attached after words 'bgs'. So, after the exclamation marks were removed in the fourth step, the phrase 'tempatnya tidak bgs', was corrected again into 'tempatnya tidak bagus' (the place was not good).

6) Removing Stopwords: In this stage, the stopwords that occur, like 'i', 'you', 'always', were removed. This step used dictionary built by [17] by combining NLTK[8] for English and Sastrawi[9] for Indonesian.

7) Removing Repetitive Characters: Sometimes, people like to express their feeling by using many unecessary duplicated characters. These characters should be removed, and to illustrate this step, 'happppyyyy' is changed into 'happy'.

8) Stemming: In this last preprocessing step, we removed the affixes and suffixes from the words to make them back into their base form. The functions that implemented are Snowball Stemmer by NLTK for English, and Sastrawi Stemmer for Indonesian because the data are in Indonesian and English, so, we applied two stemmers.

[8] https://www.nltk.org/
[9] https://github.com/har07/PySastrawi

E. Feature Extraction

This part explains about the feature extractions for machine learning, and the development of word embedding for deep learning.

1) N-gram: The features that used for classification using machine learning is n-gram language model word level. The number of gram that extracted as features are unigram, bigram, and the combination of unigram-bigram. We also applied chi-square method for feature selection.

2) Word embedding: For deep learning, we built our own word embedding using all scraped data from PergiKuliner. The method that implemented to build word embedding is skip-gram with dimension = 300.

IV. Results and Analysis

This part explains about the experiments, results, and analysis of this research.

A. Experiments

In this study, we utilized the dataset that we made and created two scenarios for multi-label classification. Then, we compared several algorithms from machine learning and deep learning. After that, we evaluated the performances of those algorithms by comparing their F1 scores.

1) Data: This experiment using all data that are retrieved from annotation step. The total of data are 14103 reviews with three sentiment labels and four emotion labels. The distribution of labels for sentiment and emotion can be seen at Fig. 2 and Fig. 3, respectively. By seeing both figures, we noticed that the data have imbalanced labels for both sentiment and emotions. To illustrate, 'food' aspect is dominated by 'positive' sentiment and 'happy' emotion. On the other hand, all aspects beside 'food' is dominated by 'neutral' for both sentiment and emotion.

2) Scenarios:

a) First scenario: In first scenario, we employed problem transformation methods for multi-label classification in machine learning. Transformation methods that we implemented are Binary Relevance (BR), Label Powerset (LP), and Classifier Chain (CC). For machine learning algorithms, we applied are Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Extra Tree Classifier (ET). The features that we used are unigram, bigram, and combination of unigram-bigram.

b) Second scenario: In this scenario, the deep learning algorithms that utilized are Bidirectional Long Short-Term Memory (BiLSTM), and Gated Recurrent Unit (GRU). We do not use problem transformations method like machine learning, but we assigned sigmoid as the activation function and binary cross entropy as loss function for retrieving the labels of data. The word embedding that has developed before is employed in this scenario.

3) Evaluation: Evaluation for both machine learning and deep learning is using kfold cross validation technique, with the number of k = 10. The scores that evaluated is f1-scores.

Fig. 2. Distribution of Sentiment Labels.

Fig. 3. Distribution of Emotion Labels.

B. Results

This section shows the performance of machine learning in first scenario, and deep learning in second scenario in every aspect of review. After that we assessed every performance in both scenarios by comparing their f1-scores.

1) First scenario:

a) Label powerset: This part presents the performance of machine learning algorithms when classified using Label Powerset (LP) as transformation method.

From Table I, it shows that ET achieved highest score, 88.17% for unigram feature. While for bigram, the highest score was acquired by RF with 87.3% for f1-score. This score was higher 0.61% compared to SVM score as second place. In the other hand, RF and ET claimed same f1 scores for unigram-bigram, which is 88.16%. By seeing the scores, it can be concluded that the best feature in this classification results is unigram.

Table II shows the performance of RF that dominated every feature in price aspect. However, for unigram-bigram feature, ET obtained same f1-score with RF, which is 89.54%. For the best feature in classification for price aspect, unigram-bigram achieved highest score compared to other two features.

TABLE I. Classification Results for Food Aspect

	Unigram	**Bigram**	**Unigram-bigram**
SVM	85.04%	86.69%	86.64%
DT	82.21%	81.98%	83.40%
RF	88.16%	**87.30%**	**88.16%**
ET	**88.17%**	86.10%	**88.16%**

TABLE II. CLASSIFICATION RESULTS FOR PRICE ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	83.16%	84.71%	85.24%
DT	84.96%	83.80%	86.21%
RF	**87.17%**	**87.53%**	**89.54%**
ET	86.84%	86.64%	**89.54%**

For service aspect, Table III shows ET monopolized the scores for both unigram and unigram-bigram features. While for bigram, the highest score was led by RF with 90.88%, 0.21% higher than ET. However, the best feature for this classification in service aspect is unigram-bigram with score is 92.65% obtained by ET.

Similar to previous table, Table IV shows ET achieved highest scores for both unigram, and unigram-bigram when classifying 'ambience' aspect. Also, RF obtained highest score for bigram feature with 81.82%. Then, same with service aspect, in this classification results, the best feature is unigram-bigram with score is 86.98% that achieved by ET.

From Table V, we can see the highest scores in every aspect and in every feature that implemented. By seeing the table, it presents that with Label Powerset (LP), 'food' aspect was the only one that has highest score when it was classified using unigram with score 88.17%, while other aspects got their best performances when they were classified with unigram-bigram. Besides, ET obtained highest scores in every aspect except 'price' which its highest score achieved by RF. For bigram feature, all aspects were dominated by RF, but the scores are below unigram-bigram features. In addition, the aspect that has the highest score compared to other aspects is service that attained by ET with score is 92.65%.

TABLE III. CLASSIFICATION RESULTS FOR SERVICE ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	88.55%	87.84%	89.80%
DT	88.74%	87.89%	89.54%
RF	90.64%	**90.88%**	91.84%
ET	**90.77%**	90.67%	**92.65%**

TABLE IV. CLASSIFICATION RESULTS FOR AMBIENCE ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	81.61%	79.74%	83.13%
DT	80.41%	75.53%	82.53%
RF	85.82%	**81.82%**	86.48%
ET	**85.96%**	81.48%	**86.98%**

TABLE V. BEST PERFORMANCE OF EVERY ASPECT

	Unigram	Bigram	Unigram-bigram
Food	**88.17% (ET)**	87.30% (RF)	88.16% (RF)
Price	87.17% (RF)	87.52% (RF)	**89.54% (RF)**
Service	90.77% (ET)	90.88% (RF)	**92.65% (ET)**
Ambience	85.96% (ET)	81.82% (RF)	**86.98% (ET)**

b) Binary relevance: This part presents the performances of machine learning algorithms when classified using Binary Relevance (BR) as transformation method.

Table VI shows the performance of RF that attained highest scores in every feature in 'food' aspect. ET follows it by obtaining scores that not really far from RF scores. The table also shows that classification result using unigram-bigram feature is higher than other features, even though the score is only 0.01% higher than score that retrieved by using unigram feature only.

By seeing the Table VII, for the first time DT attained highest score comparing to other algorithms, with unigram feature. DT achieved 83.60%, followed by ET that got score which was 1.27% lower than DT. For bigram feature, RF achieved highest score when classifying 'price' aspect. However, unigram-bigram, once again, become the feature that helped ET to attain highest score for 'price' aspect with score 87.56%.

Similar to 'price' aspect results, Table VIII shows DT achieved highest score again for classifying 'price' aspect using unigram feature, but for this time, DT was followed by RF that was 0.39% lower than DT. RF also leads the score by classifying using bigram, and its score is 90.07%. Best feature for this aspect also obtained by unigram-bigram, with ET as classification algorithm. The score ET obtained was 91.28%, 1.21% higher compared to bigram and RF pair.

From Table IX, it can be seen that ET leads in both unigram and unigram-bigram features while classifying the 'ambience' aspect. While RF achieved best score when classifying using bigram feature with score id 80.12%. In addition, similar to three previous aspects, best classification score was obtained when using unigram-bigram feature by ET.

From the comparison of all machine learning algorithms that shown in Table X, we can see all best performances were attained by using unigram-bigram as feature. By applying BR method, and unigram-bigram as feature, ET successfully obtained highest scores in three aspects, which are 'price', 'service', and 'ambience'. In other hand, RF dominates all 'food' aspect scores by using all features, including unigram-bigram.

TABLE VI. CLASSIFICATION RESULTS FOR FOOD ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	83.24%	85.53%	84.89%
DT	81.21%	80.67%	82.24%
RF	**88.17%**	**86.78%**	**88.18%**
ET	88.10%	85.55%	88.04%

TABLE VII. CLASSIFICATION RESULTS FOR PRICE ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	80.97%	84.16%	84.03%
DT	**83.60%**	81.86%	85.12%
RF	81.88%	**86.25%**	87.05%
ET	82.33%	85.42%	**87.56%**

TABLE VIII. CLASSIFICATION RESULTS FOR SERVICE ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	86.34%	87.27%	87.83%
DT	**88.08%**	86.43%	88.85%
RF	87.69%	**90.07%**	90.45%
ET	87.61%	89.82%	**91.28%**

TABLE IX. CLASSIFICATION RESULTS FOR AMBIENCE ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	79.66%	79.29%	82.51%
DT	79.43%	73.87%	80.83%
RF	83.75%	**80.12%**	84.72%
ET	**83.85%**	79.48%	**85.51%**

TABLE X. BEST PERFORMANCE OF EVERY ASPECT

	Unigram	Bigram	Unigram-bigram
Food	88.17% (RF)	86.78% (RF)	**88.18% (RF)**
Price	83.60% (DT)	86.25% (RF)	**87.56% (ET)**
Service	88.08% (DT)	90.07% (RF)	**91.28% (ET)**
Ambience	83.85% (ET)	80.12% (RF)	**85.51% (ET)**

Furthermore, like LP, 'service' becomes the aspect that got highest score in Table X, which is 91.28%, compared to other aspects. Then it is followed by 'food', then 'price', and 'ambience' aspect, respectively.

c) Classifier chain: This part shows the performances of machine learning algorithms when classified using Classifier Chain (CC) as transformation method.

In Table XI, the classification results of 'food' aspect were dominated by RF in every feature that was used. However, in unigram-bigram feature, ET successfully gained same score with RF, which is 88.40%. Moreover, similar to LP and BR methods, by using CC, unigram-bigram still becomes the best feature of multi-label classification for 'food' aspect, following by unigram.

For classification of 'price' aspect, Table XII shows that RF attained best score in unigram, and also bigram feature. While for unigram-bigram feature, ET obtained the highest score with 89.24%, 1.62% and 3.93% higher compared to results from RF with bigram and unigram, respectively. This also means that once again, unigram-bigram is the best feature for classifying the 'price' aspect, similar to previous aspect.

Table XIII presents the performances of algorithms for classifying 'service' aspect. We can see that ET leads the score for classification using unigram and unigram-bigram, while RF achieved highest score for bigram. However, unigram-bigram still becomes the best feature for this aspect while it was classified using ET, and the f1-score is 92.09%.

Identical to previous aspect, as shown by Table XIV, ET obtained highest score for 'ambience' aspect in both unigram and unigram-bigram features. Best score in bigram also obtained by RF with 81.84%. Despite of it, it is still 5.26% lower than score attained by ET with unigram-bigram feature. Again, unigram-bigram becomes the best feature for 'ambience' aspect.

In Table XV, it can be noticed that unigram-bigram becomes the best feature when Classifier Chain (CC) transformation method was applied. Unigram-bigram dominates all aspects, like Binary Relevance (BR). Besides, ET also attained the highest scores almost in all aspects, except 'food' aspect that was dominated by RF, also same with BR.

In addition, like both LP and BR results, the best score between all aspects was obtained by 'service' aspect when it was classified by ET using unigram-bigram. The score that ET achieved for 'service' aspect is 92.09%, 2.85% higher than 'price' aspect which was the second highest after 'service' aspect.

TABLE XI. CLASSIFICATION RESULTS FOR FOOD ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	83.44%	85.74%	85.01%
DT	82.34%	81.54%	83.15%
RF	**88.20%**	**87.21%**	**88.40%**
ET	88.17%	86.10%	**88.40%**

TABLE XII. CLASSIFICATION RESULTS FOR PRICE ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	81.15%	84.16%	84.51%
DT	84.10%	83.16%	85.55%
RF	**85.31%**	**87.62%**	88.74%
ET	84.93%	86.84%	**89.24%**

TABLE XIII. CLASSIFICATION RESULTS FOR SERVICE ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	86.33%	87.21%	88.44%
DT	88.20%	87.05%	89.22%
RF	88.13%	**90.57%**	91.02%
ET	**88.54%**	90.39%	**92.09%**

TABLE XIV. CLASSIFICATION RESULTS FOR AMBIENCE ASPECT

	Unigram	Bigram	Unigram-bigram
SVM	79.89%	79.84%	82.80%
DT	79.81%	74.77%	80.78%
RF	85.61%	**81.84%**	86.32%
ET	**85.74%**	81.37%	**87.10%**

TABLE XV. BEST PERFORMANCE OF EVERY ASPECT

	Unigram	Bigram	Unigram-bigram
Food	88.20% (RF)	87.21% (RF)	**88.40% (RF)**
Price	85.31% (RF)	87.62% (RF)	**89.24% (ET)**
Service	88.54% (ET)	90.57% (RF)	**92.09% (ET)**
Ambience	85.74% (ET)	81.84% (RF)	**87.10% (ET)**

Table XVI shows the comparison of best performances from Binary Relevance (BR), Label Powerset (LP), and Classifier Chain (CC) with unigram-bigram as feature. We can see from the table that 'food' aspect got the highest score when it was classified by RF with CC as problem transformation method. Followed by BR, and LP, respectively. For 'price' and 'service' aspects, LP is better than other transformation methods when classifying both aspects, followed by CC, then BR. For 'price' aspect, the algorithm that obtained the highest score, which is 89.54%, was RF. While for 'service' aspect, the best score was achieved by ET with 92.65%. However, in case of 'ambience' aspect, ET attained the highest score with CC as transformation method for multi-label classification. The score that was achieved by ET in 'ambience' aspect is 87.1%, 0.12% higher than the score it obtained by using LP as problem transformation method.

Furthermore, it also can be noticed that BR cannot surpass both LP and CC, except in 'food' aspect where BR score is 0.02% higher than LP. This maybe happened because as transformation method, BR treats the labels independently before they are classified by machine learning. This means, BR does not consider the relationship between the labels. For instance, the sentiment label 'positive' is considered does not have relation with the emotion label 'happy', because both labels were classified separately. In the other hand, LP transforms the label combinations into new classes before machine learning classified them as multiclass problem. While CC transforms the labels by using the first label that obtained from first classification as a feature for classifying the next label in next classification. Thus, by seeing the way the three transformation methods work, we can conclude that LP and CC consider the relation between labels, while BR does not consider it.

Moreover, both ET and RF always obtain best score than DT and SVM in all aspects inn all transformation methods that were used in this research. It should be remembered that both ET and RF are tree-based ensemble algorithms, which means the way they work is almost similar, except the way they split the nodes and use the samples. However, by seeing Table XVI, we can see that ET dominates 'price', 'service', and 'ambience' aspects for all transformation methods, except for LP in 'price' aspect which its best score was obtained by RF. For 'food' aspect, all highest scores for all transformation method were attained by RF.

2) Second scenario: This part shows the performances of deep learning algorithms, which are BiLSTM and GRU.

TABLE XVI. COMPARISON OF PERFORMACES FROM LABEL POWERSET (LP), BINARY RELEVANCE (BR), CLASSIFIER CHAIN (CC) WITH UNIGRAM-BIGRAM

	LP	**BR**	**CC**
Food	88.16% (RF)	88.18% (RF)	**88.4% (RF)**
Price	**89.54% (RF)**	87.56% (ET)	89.24% (ET)
Service	**92.65% (ET)**	91.28% (ET)	92.09% (ET)
Ambience	86.98% (ET)	85.51% (ET)	**87.1% (ET)**

From the classification results of both deep learning algorithms, Table XVII shows that GRU and BiLSTM attained same scores for 'food' aspect. However, BiLSTM leads the scores for 'service' and 'ambience' aspects. For GRU, it obtained higher score compared to BiLSTM in 'price' aspect, which its score peaks on 83.01%, 0.92% higher than BiLSTM. Nonetheless, the scores from GRU in 'service' and 'ambience' are not very far from BiLSTM scores. The scores achieved by GRU are 0.33% and 0.86% lower than BiLSTM scores in 'service' and 'ambience' aspects, respectively. From this experiment, it can be concluded that GRU can compete with performances from BiLSTM, even though BiLSTM already uses future context that can help it to solve more complex classification problems.

TABLE XVII. COMPARISON OF PERFORMANCES FROM GRU AND BILSTM

	GRU	**BiLSTM**
Food	**88.16%**	**88.16%**
Price	**83.01%**	82.09%
Service	88.70%	**89.03%**
Ambience	83.92%	**84.78%**

In addition, like machine learning, 'service' aspect becomes the aspect that gotten highest score when it was classified by BiLSTM and GRU. Then, the aspect that becomes the second highest is 'food', followed by 'ambience' and 'price', respectively. Furthermore, it can be concluded that self-developed word embedding can work well with deep learning. Hence, the scores that obtained by deep learning algorithms are quite similar to machine learning.

C. Analysis

Overall, the results of both scenarios show that 'service' aspect becomes the aspect that can be classified better than other aspects. After 'service' aspect, it was followed by 'price', 'food', and 'ambience', respectively, for machine learning. For deep learning, the second highest score was obtained when algorithms classified 'food', followed by 'ambience', then 'price' aspect, respectively. This may be affected by the way people express their comments towards the aspects. Usually, whenever people comment about 'service' aspect, people tend to use words like 'service' or 'waitress' directly in the comments, same goes with 'price' aspects. This kind of writing is different when people talk about 'food' and 'ambience' aspect, which can be written more creative by customers. To illustrate, people often write all the food names they ordered, and explain them in detail one by one. This can lead to misclassification by the classification program if there is a conflict occurs in an 'aspect'. For example, the comment 'the noodles were very good but too oily, I don't like it', or 'the fried rice was delicious but the orange juice too blend'. Those kinds of reviews can create a conflict and affects the classification results. Same goes with 'ambience' aspect, people can explain it variatively. For instance, 'it has beautiful decoration, but the room was full of smoke'.

For second scenario results, 'price' aspect become the aspect with lowest score after classification. While 'food' and

'ambience' aspects become two and third place after 'service' aspect that has higher score. This may be caused by the label distribution in dataset, which 'positive' sentiment and 'happy' emotions are dominant in 'food' aspect, followed by 'ambience', 'service', and 'ambience' aspect. Thus, the deep learning models learned 'positive' and 'happy' labels well, compared to other labels.

Furthermore, the features that used also affect the classification results. In first scenario, unigram-bigram feature gave more information compared to apply only unigram, or only bigram independently. When classifying, unigram can work well because in unigram, words are treated individually, and those words often appear in the dataset. To illustrate, the sentence 'I like the food but it was too pricey'. In unigram, it will be 'I', 'like', 'the', 'food', 'but', 'it', 'was', 'too', 'pricey', and for bigram, it will be 'I like', 'like the', 'the food', 'food but', 'but it', 'it was', 'was too', 'too pricey'. When classifying using bigram, the models work well but not always good compared to unigram because the combination of words in bigram are not often appear in reviews compared to unigram. Thus, if unigram and bigram are combined, the models obtain more information about word when they appear individually and when they appear as pairs. Then, for second scenario, classification with self-developed word embedding can give good results with the information especially information about semantic relations between words. Hence, it should be considered to add other features, such as POS tagging, for machine learning and deep learning to enhance their performances.

Label distribution also contributes to affect the classification results. This research has imbalanced dataset, so, it will be good to use data augmentation or apply oversampling/undersampling methods to balance the data.

V. Conclusions and Future Work

For this research, we made experiments and evaluated the performances of machine learning algorithms, which are Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Extra Tree Classifier (ET) , as well as deep learning, (Bidirectional LSTM (BiLSTM), and Gated Recurrent Unit (GRU)). We made two scenarios, which in first scenario, we applied transformation methods such as Binary Relevance (BR), Label Powerset (LP), and Classifier Chain (CC) for multi-label classification in machine learning. Then the features that used are unigram, bigram, and combination of unigram-bigram. For second scenario, we utilized sigmoid as the activation function and binary cross entropy as loss function for retrieving the labels of data in deep learning. Then, self-developed word embedding is employed in this scenario for deep learning classification. The results show RF dominates with 88.4% and 89.54% F1 scores with CC method for food aspect, and LP for price, respectively. For service and ambience aspects, ET leads with 92.65% and 87.1% with LP and CC methods, respectively. On the other hand, in deep learning comparison, GRU and BiLSTM obtained similar F1- score for food aspect, 88.16%. On price aspect, GRU leads with 83.01%. However, for service and ambience, BiLSTM achieved higher F1-score, 89.03% and 84.78%.

Since the distribution of label in our data is imbalanced, for the future, it should be considered to use balancing methods such as oversampling or undersampling. We also can apply data augmentation to retrieve new data for labels that have small numbers. Besides, we need to add more features to enhance the performance of both machine learning and deep learning.

Acknowledgment

The authors would like to thank the PUTI research team for the supports and helps that provided during this research.

References

[1] M. Pontiki and J. Pavlopoulos, "SemEval-2014 Task 4 : Aspect Based Sentiment Analysis," no. SemEval, pp. 27–35, 2014.

[2] M. Pontiki et al., "SemEval-2016 Task 5 : Aspect Based Sentiment Analysis," pp. 19–30, 2016.

[3] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," *SEM 2013 - 2nd Jt. Conf. Lex. Comput. Semant., vol. 2, no. SemEval, pp. 321–327, 2013.

[4] D. Stojanovski, G. Strezoski, G. Madjarov, I. Dimitrovski, and I. Chorbev, "Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages," Multimed. Tools Appl., vol. 77, no. 24, pp. 32213–32242, 2018.

[5] S. M. Mohammad, "Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text," Emot. Meas., pp. 201–237, 2016.

[6] M. Hassan, M. S. Bin Alam, and T. Ahsan, "Emotion Detection from Text Using Skip-thought Vectors," in 2018 International Conference on Innovations in Science, Engineering and Technology, ICISET 2018, 2018, pp. 501–506.

[7] H. S. Khawaja, M. O. Beg, and S. Qamar, "Domain specific emotion lexicon expansion," in 2018 14th International Conference on Emerging Technologies, ICET 2018, 2019, pp. 1–5.

[8] L. Wikarsa and S. N. Thahir, "A text mining application of emotion classifications of Twitter's users using Naïve Bayes method," in Proceeding of 2015 1st International Conference on Wireless and Telematics, ICWT 2015, 2016, pp. 1–6.

[9] J. Savigny and A. Purwarianti, "Emotion classification on youtube comments using word embedding," Proc. - 2017 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2017, pp. 1–5, 2017.

[10] A. Cahyadi and M. L. Khodra, "Aspect-Based Sentiment Analysis Using Convolutional Neural Network and Bidirectional Long Short-Term Memory," 2018 5th Int. Conf. Adv. Informatics Concept Theory Appl., pp. 124–129, 2018.

[11] D. Ekawati, "Aspect-based Sentiment Analysis for Indonesian Restaurant Reviews," 2017.

[12] K. Shalini, B. Ganesh, A. K. M, and K. P. Soman, "Sentiment Analysis for Code-Mixed Indian Social Media Text With Distributed Representation," 2018 Int. Conf. Adv. Comput. Commun. Informatics, pp. 1126–1131, 2018.

[13] S. Y. M. Lee and Z. Wang, "Multi-view learning for emotion detection in code-switching texts," in Proceedings of 2015 International Conference on Asian Language Processing, IALP 2015, 2016, pp. 90–93.

[14] M. Pontiki, D. Galanis, and H. Papageorgiou, "SemEval-2015 Task 12 : Aspect Based Sentiment Analysis," 2015.

[15] P. Ekman, W. V. Friesen, and P. Ellsworth, "What Emotion Categories or Dimensions Can Observers Judge from Facial Behavior? In Emotion in the Human Face," no. Cambridge University Press, pp. 98–110, 1982.

[16] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines," Proc. 9th Int. Conf. Lang. Resour. Eval. Lr. 2014, no. 2010, pp. 859–866, 2014.

[17] A. Suciati and I. Budi, "Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia," in 2019 International Conference on Asian Language Processing (IALP), 2019, pp. 59–64.