CBE : Corpus-Based of Emotion for Emotion Detection in Text Document

Fika Hastarita Rachman
Informatics Department
University of Trunojoyo Madura
Bangkalan-Madura, Indonesia
fika@if.trunojoyo.ac.id

Riyanarto Sarno
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Chastine Fatichah
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
chastine.fatichah@gmail.com

***Abstract*—Emotion Detection is a part of Natural Language Processing (NLP) that still evolve. Emotional Corpus that had been widely used are Wordnet Affect Emotion (WNA) and ANEW (Affective Norms for English Words). There are two ways to analyze the text based Emotion Detection: Categorical and Dimensional Model. Each model has different advantages and disadvantages. And each model has a different concept to predict emotion. The contribution of this research is forming automatic emotional corpus with merging two computational model. It called Corpus-Based of Emotion (CBE). CBE developed from ANEW and WNA with term similarity measure and distance of node approach. Latent Dirichlet Allocation (LDA) is used too for automatically expand CBE. The CBE attributes are a score of Valence (V), Arousal (A), Dominance (D) and categorical label emotion. Categorical label emotion based on six basic emotion of Ekman. Based on experiment results, it is known that CBE is able to improve the accuracy in detection of emotions. F-Measure using WNA+ANEW is 0.50 and F-Measure using CBE with expanding is 0.61.**

Keywords — corpus of emotion; WNA; ANEW; Emotion Detection; Categorical model, Dimensional model, LDA

I. INTRODUCTION

Emotion detection of text is a research that is considered important in analyzing personal emotion. There are two ways to analyze the emotion in text using computing system.. There are Categorical Model and Dimensional Model [1]. Each model has different of advantages and disadvantages.

In Categorical Model, the value of discrete elements depends on the observed frequencies. The disadvantage of this model is the document probably classify in the wrong category. It happened because sometimes scientists use different label emotions in emotions categorization. Six basic Emotion of Ekman[1] became the label base of emotion for others detection emotion research. Wordnet-Affect Emotion[2] are affect lexical dataset which developed from Wordnet[3]. It used in text processing based on the categorical model. The WNA attributes are code term in wordnet, synset character (noun, adverb, adjective or verb) and also synonym synset.

Dimensional Model of Emotion presented in a dimensional space. Russell in 1979, use Valence and Arousal in his research[4]. It is illustrated with Circumplex model. In 1980, Plutchick developed the concept of space dimensions into a corner in a vector space. The named is Plutchick model Emotion[4]. In the middle of 1996, Mehrabian introduces the emotion concept with a variety of 3-dimensional. There is Pleasure-Arousal-Dominance[4]. ANEW[3] is Dataset that most frequently used for dimensional emotion model. It has a term attribute and value of Valance, Arousal, and Dominance. Using this model, the dimension value of each term has no effect to emotion categorization labels. But the disadvantages is human cannot indicate clearly the label emotions for the text document.

Francisco and Gervás [5] doing research for automated marking up of text with the emotional label. They create List of Emotional Words (LEW) which was indicating the combination of emotional categories and emotional dimension model. They used SAM for Corpus Annotation Method. 15 experts invited for contributing and evaluation of valence, arousal, and dominance scores. Valence, arousal, and dominance score is taken from the result of SAM and the category of emotion label is annotated from the expert. The automated tag in emotional categorization is done by checking the word in LEW. Category emotion of the sentence can be seen from the frequency of the number emotional label of words[5]. Francisco combines LEW, ANEW [3] and WordNet, as a knowledge base. But there were no automatic tagging steps on words that have no value VAD and emotion labels.

This research trying to merge the Dataset ANEW, WNA, WordNet, Similarity measure of words and the concept of distance between one node to another node on the space dimensions. This research needs the nearest neighbor nodes for getting the relation between terms. The result of this research is formed new dataset that has the valence, arousal, dominance scores and emotion label category. Euclidean Distance method is used to analyze the distance between two nodes in dimensional space. And for emotion label, CBE using six basic emotions of Ekman (Anger, Sadness, Joy, Disgust, Fear, and Surprise). In this research, there is a handling method for the case of a word that has no value VAD and or have no emotion label. It also has a step for expanding the corpus with the help of methods of LDA (Latent Dirichlet Allocation) [6].

978-1-5090-0890-2/16/$31.00 ©2016 IEEE

Wordnet is a lexical dataset that has a relation between the synset. The synset is a synonym set. Many types of research in Natural Language Processing, Data Mining, and other topic used this dataset. The example, a business process application is supported by wordnet[6].

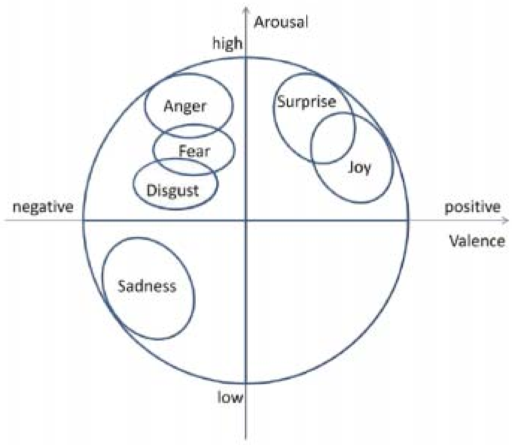

Fig 1. Mapping six basic emotion in dimensional space[7]

The purpose of this research is to establish a corpus-based of emotion with automatically that. That have a valence, arousal and dominance score (dimension scores) and emotion labels. So it can be used for emotion detection using categorical and dimensional models. If the label emotion does not match, the researchers can use dimension value as a reference. The problem is how to establish a model of automatic tagging for the case of an incomplete value of the term. The values are dimension score and categorization label. Thus the CBE becomes complex.

II. Methods

Basically, WNA Dataset and ANEW Dataset has the different concepts. WNA Dataset only has label emotion and ANEW Dataset only has scores of dimension. It cause-effect when the method merged. There would be data without label or dimension score. There are three stages in the establishment of CBE:

1. Merging WNA and ANEW
2. Automatic tagging of incomplete data
3. Expand CBE

Figure 2 show the step of the method proposed. By using WNA and ANEW Dataset we can process incomplete data become the complete data and complete dataset.

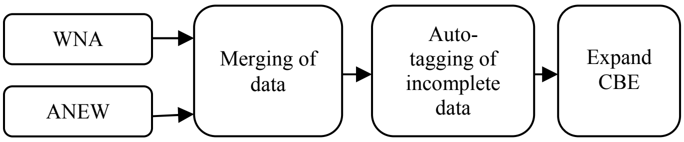

Fig 2. Step of the method proposed

The last step was Expand CBE. This step makes CBE can be dynamic corpus when there are new training data. CBE was able to adapt with mapping the existing term in the training data into CBE.

A. Merging WNA and ANEW

WNA and Anew have different data format. So, before merging, it needed to assign standardization of the attributes into CBE attribute. CBE have attributes: term, value of Valence-Arousal-Dominance and label emotion.

B. Automatic Tagging of incomplete data

Autotagging data process is done when there are incomplete data in the merged process. From merging process, would be found the data which has no label emotion and or no value dimensions. Figure 3 explain the process of auto-tagging.

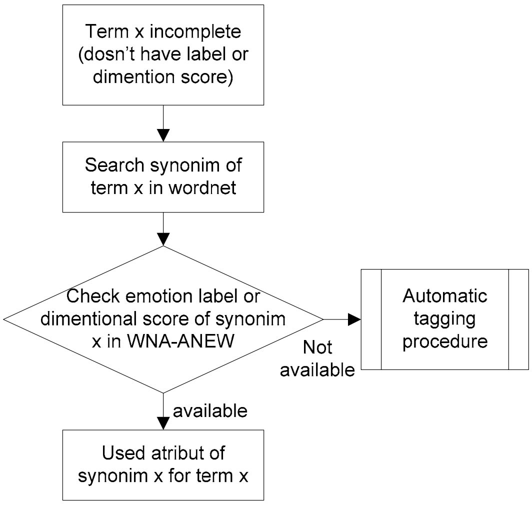
Term x incomplete (dosn't have label or dimention score)
Search synonim of term x in wordnet
Check emotion label or dimentional score of synonim x in WNA-ANEW
Not available
Automatic tagging procedure
available
Used atribut of synonim x for term x

Fig 3. Step of Autotagging Incomplete Data

It begins with searching synonym of the term in WordNet. And automatic tagging procedure executed when the system doesn't found a relation synonyms in a dataset. Autotagging Incomplete data used similarity measure method for finding the near neighbor of term x. There are several methods for similarity measure: Wu-Palmer, JCN, LCH, LIN, RES, PATH, LESK, Adapted LESK, and HSO [8]. Before using one of that methods, this research trial to found correlation between the distance of term (Euclidean Distance) and several methods of the similarity measure. This research used the Adapted LESK Method[9] because the score of Pearson Correlation[10] indicated by 0.97.

As well as the concept of K-Nearest neighbors (KNN), this study uses the close nodes to the model. The difference is K-nearest is used for classification [11], while this research is used to look for the score of VAD.

The main concept of this procedure is to look for the value of VAD term with the help of the 5 nodes nearest neighbor. So this procedure takes a few node that has a high Adapted-LESK measure. Assumed that high score of Adapted-LESK measure, it means high similarity. This research using 5 nodes of VAD:

VADpc(VAD central of the cluster), VADy and VADz(VAD of the nearest neighbor node), VADmin (VAD minimum), and VADx(VAD of term x).

VAD score of a testing term is obtained using Gauss-Jordan Elimination Method. Gauss-Jordan [12] have characteristics of matrix reduction and processing matrix. Gauss-Jordan can be used to solve linear equations with two or more variables. Equation result of step no.10 can not be categorized as linear equations with three variables. It takes one more help coordinate node (VADmin), in order to obtain the distance between the term x and the center coordinates (ed).

$$\left[\begin{array}{ccc|c} Vx_{1,1} & Ax_{1,2} & Dx_{1,3} & k1 \\ Vx_{2,1} & Ax_{2,2} & Dx_{2,3} & k2 \\ Vx_{3,1} & Ax_{3,2} & Dx_{3,3} & k3 \end{array}\right]$$

Fig 4. A typical 3x3 Augmented Matrix

The all process in Autotagging incomplete data is as follow:

1. The first step, we used to determine the core of the label emotion, its serve as a center cluster of the label emotions (VADpc). At this step, researchers determined the core of the label emotion by taking a term that was definitely the core of it. Selected cores are:

   a. 'Joy' : term 'joy' with VAD (8.21,6.49,6.63)
   b. 'Sad' : term 'sad' with VAD (1.61,4.13,3.45)
   c. 'Anger' : term 'angry' with VAD (2.85,7.17,5.55)
   d. 'Disgust': term 'disgusted', VAD(2.45,5.42,2.59)
   e. 'Fear': term 'fear', VAD(2.76,6.96,3.22)
   f. 'Surprise': term 'surprised', VAD(7.47,7.47,6.11)

2. Matching its incomplete term to six center cluster of VADpc by checking the similarity of Adapted LESK (LESK2).
3. Choose the label emotion of the center cluster with the highest LESK. Higher LESK close to similarity with "x" term.
4. Took the center cluster of VAD --- (VAD_{pc})
5. Through the CBE, examine similarities of LESK between "x" term with other terms that had the same label emotion and VAD's value.
6. Took the VAD's value with the highest LESK (LESK1 and LESK4). Example : VAD term y --- (VAD_y) and VAD term z (VAD_z)

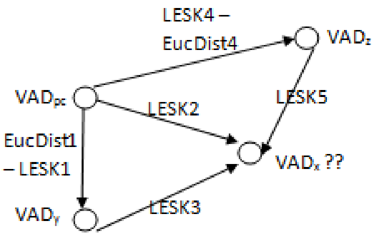

Fig 5. Mapping point of term

7. As formulas (1), find the EucDist1 between VADy and VADpc. Then find EucDist4 between VADpc and VADz.

$$EucDis(pc\,;y) = \sqrt{(Vpc - Vy)^2 + (Apc - Ay)^2 + (Dpc - Dy)^2} \quad (1)$$

8. Finding the value LESK1, LESK similarity between VADpc with VADy; LESK2 value, LESK similarity between VADpc with VADx; LESK3 value, LESK similarity between VADy with VADx; LESK4 value, LESK similarity between VADpc with VADz; LESK5 value, LESK similarity between VADz with VADx with Adapted LESK methods.
9. Calculate EucDist2, EucDist3, and EUC Dist5 with the proportionality concept. For EUC Dist2 can be searched by the formula (2), but the score of LESK need normalized.

$$\frac{LESK1}{EucDist1} = \frac{LESK2}{EucDist2} \quad (2)$$

10. Finding term minimal, the term that has VAD score minimum in emotional category label.
11. Finding Adapted LESK between term x and term min (LESKmin)
12. Finding EucDistmin:

$$\frac{LESK5}{EucDist5} = \frac{LESKmin}{EucDistmin} \quad (3)$$

13. Finding ed score:

$$ed = EucDistmin + (\sqrt{Vmin^2 + Amin^2 + Dmin^2}) \quad (4)$$

14. Forming of equation 1, 2 and 3 of the values above, the formula (5) is:

$$EucDist2^2 - Vpc^2 - Apc^2 - Dpc^2 - ed^2 = -2Vx.Vpc - 2Ax.Apc - 2Dx.Dpc$$

$$EucDist3^2 - Vy^2 - Ay^2 - Dy^2 - ed^2 = -2Vx.Vy - 2Ax.Ay - 2Dx.Dy$$

$$EucDist5^2 - Vz^2 - Az^2 - Dz^2 - ed^2 = -2Vx.Vz - 2Ax.Az - 2Dx.Dz$$

15. Find the value of V, A and D to x, using Gauss-Jordan Elimination method.

Within algorithms, incomplete data would be fixed automatically.

C. Expand CBE

Expand CBE was used to increase the number of terms in the CBE with dimension values and label emotions. In expanding CBE used ISEAR[1] (International Survey of Emotion Antecedents and Reactions) as training Dataset. There were 7666 sentences with label emotions in ISEAR Dataset and seven label emotions, which are : 'Anger', 'Shame', 'Sadness', 'Joy', 'Fear', 'Disgust', and 'Guilty'. CBE known as the system analyzes six basic emotions, then we would not implement all the labels of ISEAR emotions.

This research leaves the “Guilty” and “Shame” label documents.

Latent Dirichlet Allocation (LDA) is the topic of extraction method [6]. With the LDA method expected the term mapped out in categorical models of emotion. From the trials, there was no model optimization of the number of clusters. So by trying some number of clusters, it found the best 7 cluster.

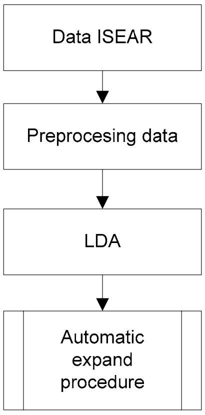

Fig 6. Step of Expand CBE

Before using LDA method, we used to preprocess the ISEAR Dataset by lowercase, punctuation removal, stopword removal and number removal analyze.

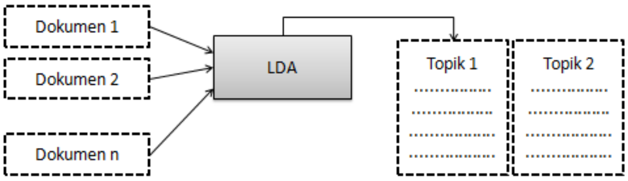

Fig 7. Schema of Input-Output LDA

Automatic Expand Procedure is the process to choose and complete the data to be added in the CBE. The steps of Automatic Expand Procedure are:

- From CBE, the center cluster of term emotional chosen with the highest probability. It must have label emotion in CBE, so we can find the center cluster from the result of LDA analyze.
- For all term emotional, find the Euclidean Distance with the center of the cluster by sorting the highest to lowest Euclidean Distance.
- By using 50 terms of threshold, so there was a possibility the term was removed from the cluster because it has too long distance to the center cluster.
- To find the complex Attribute, then we need to process the term emotional by using Automatic tagging procedure.

With all the steps required above, CBE process could be expected to expand for another training dataset, so that dataset obtained more complete.

III. Result and Discussion

In the process of accuracy testing, there were two things that to be a concern. Focus on the process of accuracy testing was a concern to a quality of corpus and the result of emotion detection. The emotional corpus used was divided into three cases: WNA+ANEW corpus, non-expand CBE corpus and CBE with Expand corpus.

For all Dataset from each corpus seen in table 1. On table 1 we can see the total number of a dataset, total incomplete and complete dataset each corpus.

TABLE I. Total data of Emotional Corpus

Detail	WNA-ANEW	CBE not expand	CBE with Expand
Total of data	1.384	1.384	2.056
Total of data complete	154	982	1.223
Total of data incomplete	1.230	402	833

Based on table 1, seen that CBE with expanding corpus has the complete and the highest total of data. But, it needs to analyze the best result of emotional detection to know corpus quality.

In this research, there was no interpretation for the position of the word in the sentences, inverting of the word and the other linguistic rule. Thus, the results of the detection label were only seen from the highest frequency scattered on the words in sentences. The final labeling emotion is seen on formulas (6).

$$\text{Emotion Label} = \max(\Sigma \text{ label emotion of term}) \qquad (6)$$

ISEAR Dataset used as a data testing. Without “Guilty” and “Shame” label, the amount of data was equal to 5.477. There are three cases of trials, the first case was used WNA + ANEW corpus, the second case using a non-expand CBE corpus, and the third case using the CBE with expanding corpus. And Evaluation measure used is Recall, Precision, and F-Measure.

Table 2. Result of experiment

Case	Recall	Precision	F-Measure
Case 1	0.63	0.42	0.50
Case 2	0.65	0.47	0.54
Case 3	0.73	0.53	0.61

From table 2, shown that emotional detection using CBE with expanding as corpus with the F-Measure 0.61 better than others.

[1] http://www.affective-sciences.org/system/files/page/2636/ ISEAR.zip, downloaded on December 14, 2014. Linked from http://www.affective-sciences.org/researchmaterial

The value of F-measure likely can be enhanced by developing the quality of corpus and use of linguistic rules in the detection of emotions. Improving the quality of the corpus can be done with the use of Part Of Speech Tagging (POS Tagging)[13], use the decision tree to choice the right rules [14] for ambiguity word cases, and use weight of word with calculate the similarity based on structural and behavioral similarity method [15] in term analysis. Also the implementation of other methods to expand the corpus, not only LDA. Linguistic rules can be developed from the analysis of the sentence structure with use of POS tagging.

IV. Conclusion

From the experiment, shown that CBE with expanding has the better dataset than others. CBE with expanding had the ability in automatically tagging incomplete data and automatic expand corpus. Only CBE with expands had the ability. But, seen from the lowest F-Measure 0.61, It needs the other text-merging concept to analyze the Natural Language Processing problem.

In the future work, this CBE would be used as a supporting text emotion detection process. Eventually, this research would be developed by adding the concept of POS Tagging, Word Sense Disambiguation, inverting of a word in the sentence and the other linguistic rule.

References

[1] P. Ekman, “An-Argument-For-Basic-Emotions.pdf,” *Cogn. Emot.*, vol. 6, no. 3, pp. 169–200, 1992.

[2] M. M. Bradley, P. J. Lang, M. M. Bradley, and P. J. Lang, “Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings,” 1999.

[3] S. Poria, A. Gelbukh, E. Cambria, P. Yang, A. Hussain, and T. Durrani, “Merging SenticNet and WordNet-Affect Emotion Lists for Sentiment Analysis,” 2012.

[4] E. Cambria, A. Livingstone, and A. Hussain, “The Hourglass of Emotions,” *Cogn. Behav. Syst.*, pp. 144–157, 2012.

[5] V. Francisco and P. Gerv, “Automated Mark Up of Affective Information in English Texts,” in *International Conference on Text, Speech and Dialogue*, 2006.

[6] R. Sarno., C. A. Djeni., I. Mukhlash., and D. Sunaryono., “Developing A Workflow Management System for Enterprise Resource Planning,” *J. Theor. Appl. Inf. Technol.*, vol. 72, no. 3, 2015.

[7] S. Mac Kim, *Recognising Emotions and Sentiments in Text*, no. April. Philosophy in the School of Electrical and Information Engineering, University of Sydney, 2011.

[8] B. S. Rintyarna and R. Sarno, “Adapted Weighted Graph for Word Sense Disambiguation,” in *IcoICT*, 2016.

[9] S. Banerjee and T. Pedersen, “An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet,” in *Third International Conference on Computer Linguistics and Intelligent Text Processing*, 2002, pp. 136–145.

[10] S. Management and B. W. Naukowe, “Comparison Of Vales of Pearson’s and Spearman's Correlation Coefficients On the Same Sets of Data,” *Quaest. Geogr.*, vol. 30, no. 2, pp. 87–94, 2011.

[11] B. Y. Pratama and R. Sarno, “Personality Classification Based on Twitter Text Using Naive Bayes , KNN and SVM,” in *2015 International Conference on Data and Software Engineering (ICoDSE)*, 2015, pp. 170–174, DOI: 10.1109/ICODSE.2015.7436992.

[12] L. Smith and J. Powell, “An Alternative Method to Gauss-Jordan Elimination : Minimizing Fraction Arithmetic,” *Math. Educ.*, vol. 20, no. 2, pp. 44–50, 2011.

[13] C. D. Manning, “Part-of-Speech Tagging from 97 % to 100 %: Is It Time for Some Linguistics ?,” *CICILing*, 2011.

[14] R. Sarno, H. Ginardi, E.W. Pamungkas, D. Sunaryono, "Clustering of ERP Business Process Fragments", *International Conference on Computer, Control, Informatics, and Its Applications (IC3INA)*, 2013, DOI: 10.1109/IC3INA.2013.6819194.

[15] R. Sarno, P. L. I. Sari, H. Ginardi, D. Sunaryono, I. Mukhlash, "Decision Mining for Multi Choice Workflow Patterns", *International Conference on Computer, Control, Informatics, and Its Applications (IC3INA)*, 2013, DOI: 10.1109/IC3INA.2013.6819197.