Text Emotion Recognition using Affective Dictionary/Lexicon

Bagus Tris Atmaja

August 22, 2019

1 Background

In this third report, I present my works on text emotion recognition using affective dictionary and lexicon. Three dictionaries/lexicons are evaluated on this work,

1. New ANEW (Affective norms for English words)
2. VADER (Valence Aware Dictionary and sEntiment Reasoner)
3. SentiWordnet 3.0

For the dataset, two previous used datasets are evaluated namely

1. IEMOCAP
2. EmoBank

The goal of this part is to evaluate affective dictionary/lexicon-based text emotion recognition, without the use of deep learning, neural network or other machine learning technique. From affective dictionaries, the score of valence, which equal to sentiment intensity, can be obtained, and the total score for a sentence or utterance can be determined. For each affective dictionary, some strategies to obtain the total score of valence and the effect of data normalization can be performed. While the first dictionary (ANEW) provide all valence, arousal, and dominance score to determine dimensional emotion, the last two dictionaries only provided valence score. In this case, polarity of sentiment (or valence score only) from text (including speech transcription) could be predicted instead of all three degree of emotion dimensions (VAD).

2 Methods

2.1 ANEW-based Text Emotion Recognition

ANEW, by Bradley and Lang [1], is a dictionary contains of affective meaning of words. Originally, it only consists of 1,034 words. Warriner et al. [2] extends this dictionary to 13, 915 words. Three components of emotions are measured for each words: valence (the pleasantness of a stimulus), arousal (the intensity of emotion provoked by a stimulus), and dominance (the degree of control exerted by a stimulus). Table score of some words in ANEW affective dictionary.

The algorithm to obtain VAD score for each utterance can be described as follows,

Table 1: Excerpt of ANEW dictionary

Word	valence	arousal	dominance
able	6.64	3.38	6.17
abuse	1.53	6.21	2.9
cancer	1.9	5.14	2.9
crime	1.95	5.68	3.31
delight	8.21	5.02	7.29
happy	8.47	6.05	7.21
harass	2.95	6.1	3.58

1. Tokenize sentence into words
2. For each word, find it on ANEW dictionary
3. Obtain VAD score of corresponding word
4. Calculate the total score for each VAD using either mean, median, or Mika Mantyla [3] (hereafter referred as Mika) method.

For the mean method, the total score for each dimension can be calculated as follows,

$$\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i \tag{1}$$

For the median, the score for each VAD dimension of an utterance can be calculated,

$$w \rightarrow w_{med} \tag{2}$$

where,

$$med = \frac{n+1}{2}$$

If n is odd then the median is the middle value, if n is even, then median is the average of two middle values.

For the Mika method, the total score to compute each VAD score, w, can be formulated,

$$w_{mika} = \begin{cases} max(\bar{w}) - \bar{W}, & \text{if } min(w) > \bar{W} \\ \bar{W} - min(w), & \text{if } max(w) > \bar{W} \\ max(\bar{w}) - min(\bar{w}), & \text{if } min(w) \leq \bar{W} \leq max(w) \end{cases} \tag{3}$$

where $\bar{W}$ is average score of all words in lexicon (N is 13,915).

2.2 SENTIWORDNET 3.0

SENTIWORDNET [4] is a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. The version 3.0 is an improvement of previous version 1.0. SENTIWORDNET consist of three scores: positive, negative and objective (neutral), while the dictionary itself only lists positive and negative score. The objective score then is calculated as,

Table 2: Fragment of SENTIWORDNET score for some words.

Word	Pos	Neg
able	0.125	0
unable	0	0.75
abuse	0	0.625
cancer	0.125	0.625
worrying	0	0.875
light	0.25	0.25
thoughtful	0.5	0

$$Obj = 1 - (Pos + Neg) \tag{4}$$

In this research, valence score for each word is subtraction of positive (Pos) by negative (Neg) score,

$$V = Pos - Neg \tag{5}$$

Valence score for each utterance then is calculated using either mean, median or Mika method. Table 2 shows fragments of SENTIWORDNET lexicon for some words. As it used WordNet as dictionary, a set of synonyms is used to search the similar word meaning. I used degree of synset "0" to take the most common of a word given its part-of-speech (POS) in WordNet.

2.3 VADER

VADER [5] is a simple rule-based model for general sentiment analysis. It superseded the previous sentiment tools like ANEW and SENTIWORDNET. Although it provides positive, neutral and negative score in the similar ways on how SENTIWORDNET calculate valence, the compound score is the accurate single score valence as it is normalized and weighted composite score from those three scores. The compound score (Comp) is for each utterance not for each word, hence no further calculation is required.

Table 3: Example of valence score of some sentences using VADER

Sentence	Comp	Neg	Neu	Pos
I am very happy today, I got a gift from my grandma.	0.7841	0.0	0.504	0.496
Are you angry to me...?	-0.5106	0.452	0.548	0.0
Oh my dear, you looks so beautiful.	0.822	0.0	0.397	0.603
Oh my god, how pity you are.	-0.0258	0.237	0.538	0.226
Let's take a rest, I am so tired.	-0.5777	0.428	0.572	0.0

2.4 Evaluation Metric

Due to different scale of dimensional emotion (VAD) and sentiment/valence score (V), the use of error metrics such as mean absolute error (MAE) and mean squared error (MSE) which are previously used is less important except the prediction scores are normalized

in the same scale, e.g. in range (-1, 1). Instead, we added the commonly used metric in dimensional emotion recognition i.e. by measuring concordance coefficient correlation (CCC) between true value (x) and predicted value (y). CCC is formulated as,

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2\sigma_y^2 + (\mu_x - \mu_y)^2} \tag{6}$$

Where ρ is Pearson correlation between x and y, μ and σ is mean and variance between those two variables. The CCC score is the main interest of the metric to evaluate the method.

MAPE (mean absolute percentage value) is not used in this part because some utterances have zeros score after normalization. Therefore, that metric cannot be used to evaluate the performance.

3 Results

3.1 Text emotion recognition using ANEW Dictionary

Three methods to calculate valence, arousal, and dominance by using ANEW dictionary has been performed, namely: mean, median and Mika method. As the focus is the degree of emotion, the goal is to obtain the closest score/value between true emotion dimension (from label in database) and predicted score using ANEW dictionary. The evaluation is measured by CCC score as main metrics with additional information of MAE and MSE. Two databases are used in this research, IEMOCAP [6] and EMOBANK [7].

Table 4 shows result of ANEW-dictionary based text emotion recognition by three different methods to calculate score within utterance from given word found in dictionary.

Table 4: Result of ANEW-based text emotion recognition using IEMOCAP database.

		MAE	MSE	CCC
mean	V	0.4659	0.3131	**0.1632**
	A	0.3640	0.2048	0.0049
	D	0.3606	0.2033	-0.0117
median	V	0.4821	0.3359	0.1518
	A	0.3821	0.2246	-0.0187
	D	0.3615	0.2048	-0.0011
Mika	V	0.4390	0.2995	0.0087
	A	0.3557	0.1975	**0.1953**
	D	0.3775	0.2207	**0.1537**

The bold score shows the highest score for each emotion dimension. This result on IEMOCAP dataset shows that the mean method obtained highest score for valence, while Mika method got both arousal and dominance highest results. For EMOBANK dataset, the result shows in Table 5. In that Table, mean method outperforms other methods for calculating utterance emotion score from given words found in that utterance. For both databases, the CCC score calculation was performed after the label is normalized to (-1, 1) so does the prediction score. The inconsistency result from this two databases where IEMOCAP achieves highest result on arousal and dominance using Mika method while EMOBANK achieves the highest VAD CCC score using mean method should be investigated in the future or should be compared to other databases.

Table 5: Result of ANEW-based text emotion recognition using EMOBANK database.

		MAE	MSE	CCC
mean	V	0.2210	0.0750	**0.3804**
	A	0.2601	0.1019	**0.1372**
	D	0.1930	0.0642	**0.1555**
median	V	0.2390	0.0882	0.3191
	A	0.2827	0.1176	0.1169
	D	0.2062	0.0718	0.1308
Mika	V	0.2798	0.1328	0.0324
	A	0.2964	0.1340	0.1097
	D	0.2631	0.1115	0.0286

3.2 Sentiment Analysis using SENTIWORDNET and VADER

For the second experiment, a sentiment analysis using SENTIWORDNET and VADER have been conducted. For the SENTIWORDNET, three methods to calculate valence score same as in 3.1 was used i.e. 'mean', 'median' and 'Mika' methods while VADER use its own heuristic calculation method to predict valence score within utterance. Table 6 shows Result of valence prediction on IEMOCAP data using SENTIWORDNET and VADER. It is shown that the more recent sentiment analysis tool i.e. VADER outperforms significantly SWN results from three different methods. The use of five heuristic approaches based on punctuation, capitalization, degree of modifiers, contrastive conjunction and tri-gram lexical features in VADER effectively catch sentiment intensity within each sentence in IEMOCAP dataset. The result obtained by VADER is also higher than the highest one obtained by ANEW approach in IEMOCAP dataset.

Table 6: Result of sentiment analysis (valence prediction) using SENTIWORDNET (SWN) and VADER on IEMOCAP dataset.

		MAE	MSE	CCC
SWN	mean	2.7812	8.6106	0.0078
	median	0.4511	0.3113	0.0882
	Mika	2.7229	8.2918	0.0053
VADER		0.4470	0.3032	**0.2112**

For EMOBANK dataset, the sentiment intensity result obtained by SWN and VADER is shown in Table 7. It is shown in that table that VADER also outperforms SWN method as in IEMOCAP result. Compared to ANEW result on valence prediction, the result obtained by VADER is slightly higher in term of CCC.

To summarize valence prediction by three methods: ANEW, SWN, and VADER, figure 3.2 shows highest CCC score from each method on IEMOCAP and EMOBANK datasets. For both datasets, VADER gain the highest score which mean that the predicted degree of valence or sentiment intensity by VADER method has correlation with true valence score obtained by human annotation within the dataset.

The Python scripts to run this experiment on affective lexicon-based text emotion recognition (VAD prediction) is provided in open repository[1] for result reproducibility.

[1] `http://www.github.com/bagustris/text-vad`

Table 7: Result of sentiment analysis (valence prediction) using SENTIWORDNET (SWN) and VADER on EMOBANK dataset.

		MAE	MSE	CCC
SWN	mean	0.1582	0.0449	0.2120
	median	0.2651	0.1361	0.1852
	Mika	0.2782	0.1694	0.1333
VADER		0.2611	0.0449	**0.3877**

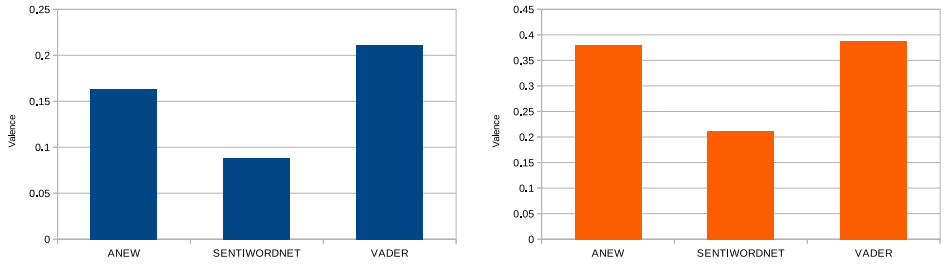

Figure 1: Comparison of valence prediction among three approaches on IEMOCAP (left/blue) and EMOBANK (right/orange) dataset.

3.3 Comparison with Acoustic-based emotion recognition

As the IEMOCAP dataset is a multimodal dataset consist of video, audio and motion movement recordings, it is also possible to perform emotion recognition task by using those modalities. To extend this work, I compared the result I obtained from this text emotion recognition task with acoustic feature-based emotion recognition in term of CCC score obtained from [8]. Table 8 shows comparison of those two results.

Table 8: Comparison of text-based and acoustic-based emotion recognition performance (CCC score) on the same IEMOCAP dataset.

Modality	Valence	Arousal	Dominance
Text	**0.2112**	0.1953	0.1537
Acoustic	0.11	**0.43**	**0.36**

For the text modality, the highest valence is obtained from VADER method, while arousal and dominance score are carried from ANEW with Mika calculation. It is interesting while acoustic-based emotion recognition gained low score on valence but higher score on arousal and dominance, the text-based emotion recognition shows the opposite. The use of both modalities may improve overall score as human emotion perception also derived from those two modalities.

4 Conclusion

A text emotion recognition study is performed by evaluating three affective lexicons: ANEW, SENTIWORDNET, and VADER, using two datasets: IEMOCAP and EMOBANK. From

the first affective lexicon i.e. ANEW, three emotion attributes/dimensions can be obtained i.e. valence (V), arousal (A), and dominance (D). To calculate score of those emotion dimensions, three methods have been used: 'mean', 'median', and Mika method [3]. Not only for ANEW, those three calculation methods are also applicable to calculate valence score using SENTIWORDNET lexicon. The last lexicon, VADER, has its heuristic model to calculate valence score form given utterance/sentence. The result shows that VADER lexicon performs best among three approaches for calculating valence score on two datasets. For arousal and dominance which only apply on ANEW, the Mika method shows better result on IEMOCAP dataset while 'mean' method attain higher score on EMOBANK dataset. The comparison of this result with acoustic-based emotion recognition on same IEMOCAP dataset shows that text-based emotion recognition performs better on valence prediction while acoustic feature predicts arousal and dominance more precisely than text feature.

References

[1] Bradley, M. M., and Lang, P. J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings (Technical Report No. C-1). Gainesville, FL: University of Florida, NIMH Center for Research in Psychophysiology

[2] Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

[3] Mäntylä, M., Adams, B., Destefanis, G., Graziotin, D., and Ortu, M. (2016). Mining Valence, Arousal, and Dominance - Possibilities for Detecting Burnout and Productivity? https://doi.org/10.1145/2901739.2901752

[4] Baccianella, S., Esuli, A., Fabrizio Sebastiani, and Sebastiani, F. (2009). SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining Stefano. Foundations and Trends® in Information Retrieval, 2(1–2), 2200–2204. Retrieved from http://en.scientificcommons.org/52462955

[5] Hutto, C. J., & Gilbert, E. E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).". Retrieved from http://sentic.net/

[6] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[7] Buechel, S., and Hahn, U. (2017). EMOBANK: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In Proc. European Chapter of the Assoc. for Computational Linguistics (Vol. 2). Retrieved from http://www.julielab.de

[8] Atmaja, B. T., Elbarougy, R., and Akagi, Masato (2019). RNN-based Dimensional Speech Emotion Recognition. Acoustical Society of Japan, Autumn Meeting 2019, Ritsumeikan University Biwako-Kusatsu Campus, Shiga, Japan.