Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7022–7028
Marseille, 20-25 June 2022

A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain

Haruya Suzuki†, Yuto Miyauchi†, Kazuki Akiyama†, Tomoyuki Kajiwara†, Takashi Ninomiya†, Noriko Takemura‡, Yuta Nakashima‡, Hajime Nagahara‡
†Ehime University, ‡Osaka University
{suzuki, miyauchi, k_akiyama}@ai.cs.ehime-u.ac.jp
{kajiwara, ninomiya}@cs.ehime-u.ac.jp
{takemura, n-yuta, nagahara}@ids.osaka-u.ac.jp

Abstract

We annotate 35,000 SNS posts with both the writer's subjective sentiment polarity labels and the reader's objective ones to construct a Japanese sentiment analysis dataset. Our dataset includes intensity labels (*none*, *weak*, *medium*, and *strong*) for each of the eight basic emotions by Plutchik (*joy*, *sadness*, *anticipation*, *surprise*, *anger*, *fear*, *disgust*, and *trust*) as well as sentiment polarity labels (*strong positive*, *positive*, *neutral*, *negative*, and *strong negative*). Previous studies on emotion analysis have studied the analysis of basic emotions and sentiment polarity independently. In other words, there are few corpora that are annotated with both basic emotions and sentiment polarity. Our dataset is the first large-scale corpus to annotate both of these emotion labels, and from both the writer's and reader's perspectives. In this paper, we analyze the relationship between basic emotion intensity and sentiment polarity on our dataset and report the results of benchmarking sentiment polarity classification.

Keywords: sentiment analysis, emotion detection, Japanese

1. Introduction

Emotion analysis is one of the major natural language processing tasks with many applications such as dialogue systems (Tokuhisa et al., 2008) and social media mining (Stieglitz and Dang-Xuan, 2013). Previous studies in emotion analysis have addressed two subtasks: classification of sentiment polarity (Sentiment Analysis) and classification of basic emotions (Emotion Detection). The former targets positive or negative sentiment polarity, while the latter targets about four to eight basic emotions, such as joy and sadness (Plutchik, 1980; Ekman, 1992).

As shown in Table 1, previous studies on emotion analysis have been diverse in terms of whether it deals with sentiment polarity (Sentiment) or basic emotions (Emotion), and whether it deals with the emotions of the writer of the text (Subjective Emotion) or the emotions of the reader of the text (Objective Emotion). However, the relationship between sentiment polarity and basic emotions, as well as the relationship between the writer's emotions and the reader's ones, are not clear, as there is no previous study that has comprehensively addressed these issues.

In this study, to extend the WRIME dataset (Kajiwara et al., 2021) with basic emotion intensity from both subjective and objective standpoints for Japanese SNS text, we annotated sentiment polarity labels from both subjective and objective standpoints for these texts. Specifically, one subjective annotator and three objective annotators annotated both labels of the sentiment polarity and the intensity of each of Plutchik's eight emotions (*joy*, *sadness*, *anticipation*, *surprise*, *anger*, *fear*, *disgust*, and *trust*) to 35,000 SNS posts collected from 60 subjective annotators. We annotated emotional intensity with a four-point scale (*none*, *weak*, *medium*, and *strong*) and sentiment polarity with a five-point scale (*strong positive*, *positive*, *neutral*, *negative*, and *strong negative*). Our comprehensively annotated corpus allows for analysis of relationships between labels and differences by annotator. Our dataset shown in Table 2 for Japanese emotion analysis is available on GitHub.[1]

In this paper, we analyze the relationship between basic emotion intensity and sentiment polarity on our dataset and report the results of benchmarking sentiment polarity classification. The correlation between emotional intensity and sentiment polarity shows that of Plutchik's eight emotions, *joy*, *anticipation*, and *trust* are positive emotions, *sadness*, *anger*, *fear*, and *disgust* are negative emotions, and *surprise* is neither. We also found that text in which multiple emotions co-occur is easier for readers to perceive the sentiment polarity of the writer than that of containing only a single emotion. Similar to the previous study on emotional intensity (Kajiwara et al., 2021), we also found that text readers weakly estimated the sentiment polarity of the writers. Experimental results on sentiment polarity classification showed that it is more difficult to estimate the subjective labels by the writer than the objective ones by the reader.

2. Related Work

Table 1 lists the popular datasets for emotion analysis in English and Japanese. These are summarized in terms of whether the target is sentiment polarity (Sentiment)

[1] https://github.com/ids-cv/wrime

	Sentiment	Emotion	Subj.	Obj.	Language	Size
IMDB (Maas et al., 2011)	✓	×	✓	×	English	50,000
SST (Socher et al., 2013)	✓	×	×	✓	English	11,855
ISEAR (Scherer and Wallbott, 1994)	×	✓	✓	×	English	7,666
SemEval-2018 (Mohammad et al., 2018)	×	✓	×	✓	English	12,634
SemEval-2007 (Strapparava and Mihalcea, 2007)	✓	✓	×	✓	English	1,250
Tsukuba sentiment-tagged corpus	✓	×	×	✓	Japanese	4,309
(Suzuki, 2019)	✓	×	×	✓	Japanese	534,962
WRIME (Kajiwara et al., 2021)	×	✓	✓	✓	Japanese	17,000
Ours	✓	✓	✓	✓	Japanese	35,000

Table 1: List of sentiment analysis datasets.

Text	I'm taking the summer off next month to go out! I'm looking forward to it!								
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Sentiment polarity
Writer	3	0	3	0	0	0	0	0	2
Reader 1	2	0	2	0	0	0	0	0	2
Reader 2	3	0	3	0	0	0	0	0	2
Reader 3	3	0	3	0	0	0	0	0	2

Text	My umbrella was stolen!!								
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Sentiment polarity
Writer	0	2	0	0	2	0	3	0	-2
Reader 1	0	0	0	0	2	0	0	0	-1
Reader 2	0	3	0	0	3	0	3	0	-2
Reader 3	0	0	0	0	3	0	0	0	-2

Text	Snowy morning with a light dusting of snow on the roof...								
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Sentiment polarity
Writer	0	0	0	2	0	0	0	0	0
Reader 1	0	0	0	0	0	0	0	0	0
Reader 2	1	0	0	1	0	0	0	0	0
Reader 3	0	0	0	1	0	0	0	0	0

Table 2: Examples of our dataset.

or basic emotions (Emotion), and whether the annotator is the text writer (Subj.) or the reader (Obj.). While previous studies have constructed datasets labeled for some of these perspectives, we annotate all of them simultaneously to construct a new dataset for a comprehensive analysis of text and emotion.

2.1. Corpus with Sentiment Polarity

Popular datasets for sentiment polarity classification include the Internet Movie Database (IMDB) (Maas et al., 2011) and Stanford Sentiment Treebank (SST) (Socher et al., 2013) in English, and the Tsukuba sentiment-tagged corpus[2] and Suzuki (2019)[3] in Japanese.

IMDB is a corpus that collects pairs of review texts and subjective rating values on a 10-point scale for movies and deals with the subjective sentiment polarity of the review texts by their writers. SST is also a corpus of sentiment polarity classification for movie review texts, but it deals with objective sentiment polarity, as a five-point scale sentiment polarity is estimated by annotators different from the writers of the review texts.

Tsukuba sentiment-tagged corpus is a corpus of review sentences from a hotel booking website with 6 types of labels: praise, complaint, request, neutral, no evaluation, and other. Suzuki (2019) annotates positive or negative sentiment polarity labels to Twitter posts that mention products. The ACP corpus[4] (Kaji and Kitsuregawa, 2006), which estimates positive or negative sentiment polarity labels to roughly one million sentences collected from the web, is the largest corpus for sentiment polarity classification in Japanese, but it is auto-

[2] http://www.nlp.mibel.cs.tsukuba.ac.jp/~inui/SA/corpus/

[3] http://www.db.info.gifu-u.ac.jp/sentiment_analysis/

[4] http://www.tkl.iis.u-tokyo.ac.jp/~kaji/acp/

	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Emotions	Sentiment
R1 vs. R2	0.580	0.474	0.528	0.468	0.610	0.430	0.399	0.208	0.508	0.608
R1 vs. R3	0.664	0.507	0.588	0.417	0.625	0.412	0.490	0.209	0.540	0.688
R2 vs. R3	0.657	0.589	0.596	0.471	0.638	0.468	0.378	0.234	0.562	0.530
W vs. R1	0.481	0.309	0.378	0.340	0.274	0.343	0.320	0.137	0.367	0.564
W vs. R2	0.587	0.434	0.441	0.402	0.297	0.357	0.452	0.114	0.450	0.493
W vs. R3	0.544	0.483	0.429	0.352	0.313	0.357	0.286	0.133	0.427	0.605
W vs. Avg. R	0.579	0.453	0.463	0.417	0.303	0.415	0.425	0.121	0.458	0.621

Table 3: Inter-annotator agreement by quadratic weighted kappa.

matically annotated.

2.2. Corpus with Basic Emotions

Popular datasets for emotional intensity estimation of basic emotions include ISEAR (Scherer and Wallbott, 1994) and SemEval-2018 (Mohammad et al., 2018) in English and WRIME (Kajiwara et al., 2021) in Japanese.

ISEAR is a corpus for subjective emotional classification that collects pairs of text and emotion labels written about one's events in the past. In SemEval-2018 Task1[5], a corpus for objective emotional intensity estimation was constructed using crowdsourcing to estimate emotional intensity to Twitter posts.

WRIME is a corpus of Japanese SNS posts with Plutchik's eight emotional intensities (Plutchik, 1980) from the perspective of both the writer and the reader. These corpora for basic emotions are expensive because they label multiple types of emotions, and are smaller in scale than the corpora for sentiment polarity classification introduced in Section 2.1.

A dataset that deals with both basic emotions and sentiment polarity is the SemEval-2007 Task14[6] (Strapparava and Mihalcea, 2007) corpus. This is a corpus from English news headlines objectively annotated with sentiment polarity and Ekman's six emotion intensities (Ekman, 1992), respectively. There is no corpus in Japanese that deals with both basic emotions and sentiment polarity.

3. Sentiment Polarity Annotation

3.1. Annotating Subjective Labels

We employed[7] 60 subjective annotators using the crowdsourcing service Lancers[8]. Our annotators consist of the gender breakdown of the annotators was 21 males and 39 females, and the age breakdown was 28 people in their 20s, 22 people in their 30s, and 10 people over 40 years old. The subjective annotators labeled the intensity of each of Plutchik's eight emotions (Plutchik, 1980) to their past posts on the SNS at a four-point scale (none, weak, medium, and strong), and also annotated sentiment polarity at a five-point scale (strong positive, positive, neutral, negative, and strong negative). Here, for the purpose of emotion analysis from text, posts with images or URLs were excluded. Each annotator labeled 100 to 1,000 posts, and a total of 35,000 posts were collected. All subjective annotators labeled both emotional intensity and sentiment polarity labels to all posts that they provided. We did not set any restrictions on the time of post, as a result, posts were collected for a range of 10 years, from August 2010 to November 2020.

To evaluate the quality of the collected sentiment polarity labels, 30 posts were randomly selected for each annotator. One of our undergraduate students evaluated the posts and the sentiment polarity labels on a four-point scale based on the following criteria.

- 3: I fully agree with the given label.
- 2: I can find the relevance between the post and label.
- 1: I hardly find the relevance between the post and label.
- 0: I do not think the annotator seriously engaged for this post.

The average of the evaluation results for each annotator was 2.4 with a minimum score of 1.9 and a maximum score of 2.8. There were no posts with a score of zero. Four annotators scored below the average of 2 points, but there were no annotators with significantly low quality.

3.2. Annotating Objective Labels

Three objective annotators were hired,[9] also using Lancers. The annotators consist of two women in their 30s and one in her 40s. The objective annotators labeled all 35,000 posts collected in Section 3.1 with both emotional intensity and sentiment polarity in the same way as the subjective annotations. However, while the subjective annotators labeled the emotions of themselves, the writers of the text, the objective annotators labeled the emotions of the writers that the readers estimated from the text.

[5] https://competitions.codalab.org/competitions/17751

[6] http://web.eecs.umich.edu/~mihalcea/affectivetext/

[7] We paid 10 JPY per post for subjective annotators.

[8] https://www.lancers.jp/

[9] We paid 3 JPY per post for objective annotators.

	-2	-1	0	+1	+2
Writer	4,105	6,465	10,380	9,415	4,635
Avg. Readers	1,687	10,468	11,462	9,138	2,245
Reader 1	2,254	10,316	8,741	11,216	2,473
Reader 2	1,056	4,029	20,147	8,510	1,258
Reader 3	9,581	4,256	10,687	2,841	7,635

Table 4: Distribution of sentiment polarity labels by the annotator.

To evaluate the quality of annotations, we calculated the inter-annotator agreement using quadratic weighted kappa[10] (Cohen, 1968). The upper part of Table 3 shows the agreement between the objective annotators. Although *anger* has a substantial agreement of $\kappa > 0.6$ and *trust* has a fair agreement of $\kappa < 0.4$, the overall moderate agreement of the eight *emotions* is $0.5 < \kappa < 0.6$. In terms of sentiment polarity, moderate to substantial agreement was also confirmed.

4. Analysis

The lower part of Table 3 shows the inter-annotater agreement between subjective and objective annotators. Similar to the previous study (Kajiwara et al., 2021), for basic eight emotions, the overall agreement between subjective and objective annotators is lower than the agreement between objective annotators. A similar trend was observed for the newly labeled sentiment polarities in the present study. Note that when the labels of the three objective annotators were averaged, the overall agreement with the subjective annotators slightly improved.

Other characteristics related to the basic eight emotions showed the same trend as in the previous study (Kajiwara et al., 2021). Although the number of subjective annotators and the number of labeled posts increased in this study, this is not surprising because the annotation method for emotional intensity are follows the previous study (Kajiwara et al., 2021). In the following sections, we will investigate the newly labeled sentiment polarities in this study. Here, the labels for strong positive, positive, neutral, negative, and strong negative will be denoted as +2, +1, 0, -1, and -2, respectively.

4.1. Distribution of Sentiment Polarity

Table 4 shows the distribution of the sentiment polarity labels for each annotator. The sentiment polarity labels of the subjective annotators (writers) are most often neutral, with relatively few extreme labels of strong positive and strong negative. The three objective annotators (readers) showed different characteristics. Reader 1 labeled more positives and negatives than neutrals. Reader 2 has a similar tendency to the writers, but estimated more neutral labels and less extreme labels (*i.e.*, strong positive and strong negative). Reader 3 labels more strong positives and strong negatives. The average over the three annotators shows similar trends to the subjective annotators, but with fewer extreme strong positive and strong negative and more negative labels.

	Avg. Readers				
Writer	-2	-1	0	+1	+2
+2	0.2	5.0	24.3	47.5	23.0
+1	0.6	8.2	32.7	48.1	10.5
0	1.8	27.8	49.3	19.4	1.6
-1	9.3	63.7	23.2	3.7	0.2
-2	20.3	59.9	15.7	3.8	0.2

Table 5: Confusion matrix of subjective and objective sentiment polarity labels. (%)

Table 5 shows the confusion matrix between the sentiment polarity labels of the subjective annotators and the averaged labels of the three objective annotators. In the posts where the subjective annotators labeled strong positive, the objective annotators estimated positives (47.5%) and strong positives (23.0%) to 70.5% of them. In addition, the subjective annotators labeled strong negative, the objective annotators estimated negatives (59.9%) and strong negatives (20.3%) to 80.2% of them. Note that the percentage of positive and negative reversals between subjective and objective annotators is sufficiently small, ranging from 3.9% to 8.8%. These observations indicate that although objective annotators succeed in roughly estimating the sentiment polarity of subjective annotators, they tend to estimate the intensity of sentiment polarity more weakly.

In the previous study (Kajiwara et al., 2021), it was pointed out that objective annotators tend to estimate emotional intensity weaker than subjective annotators for basic emotions. Similarly, we found that objective annotators tended to be less likely to give extreme sentiment polarity labels on average, and to estimate sentiment polarity more weakly than subjective annotators.

4.2. Relationship between Emotional Intensity and Sentiment Polarity

Table 6 shows the Pearson correlation between emotional intensity and sentiment polarity. Here, the correlation coefficients were calculated by expressing the four-levels of emotional intensity as 0, 1, 2, and 3 (no, weak, medium, and strong), and the five-levels of sen-

[10] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust
Writer	0.585	-0.526	0.381	0.052	-0.353	-0.298	-0.467	0.296
Avg. Readers	0.665	-0.539	0.400	0.037	-0.229	-0.410	-0.470	0.252

Table 6: Pearson correlation coefficient between emotional intensity and sentiment polarity.

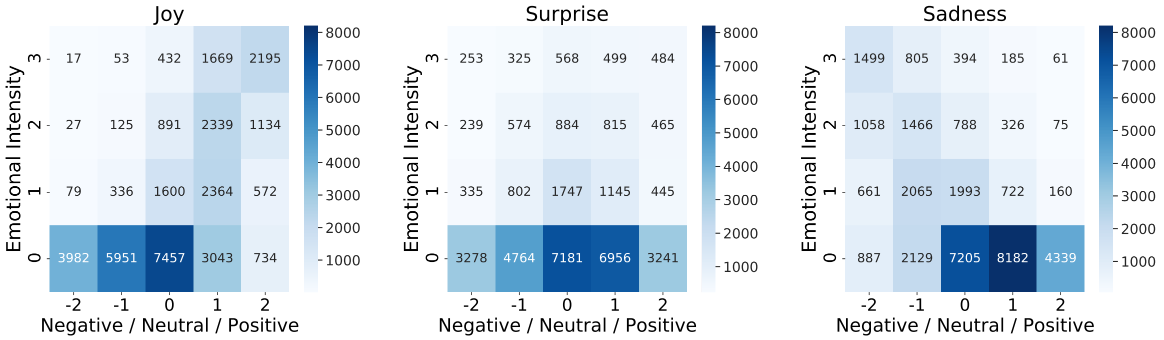

Figure 1: Distribution of emotional intensity and sentiment polarity of subjective annotators.

timent polarity from -2 to +2. From this table, we can see whether each emotion is positive or negative. For example, since there is a positive correlation between emotional intensity and sentiment polarity for *joy*, the more joyful a post is, the more positive it is.

From this analysis, we can say that three of Plutchik's eight emotions, *joy*, *anticipation*, and *trust*, are positive emotions, and four of them, *sadness*, *anger*, *fear*, and *disgust*, are negative emotions. The correlation between emotional intensity and sentiment polarity was particularly strong for *joy* among the positive emotions and for *sadness* among the negative emotions.

Figure 1 shows the distribution of emotional intensity and sentiment polarity. This figure also confirms the positive correlation between emotional intensity and sentiment polarity in the emotion of *joy* and the negative correlation in the emotion of *sadness*. It also shows that the emotion of *surprise* appears in both positive and negative posts.

In addition, we found that the agreement between labels by subjective annotators and those by objective annotators increased for posts in which multiple emotions co-occurred. In this analysis, the dataset was divided into three parts according to the subjective emotional intensity as follows.

- No emotion: All emotional intensities are none or weak.
- Single emotion: Only one emotional intensity is medium or strong.
- Multiple emotions: Two or more emotional intensities are medium or strong.

Table 7 shows the relationship between the co-occurrence of emotions and the agreement of subjective-objective sentiment polarity. The agreement between sentiment polarity labels by subjective annotators and those by objective annotators is higher for single-emotion posts than for no-emotion posts. Furthermore, there is more agreement on multiple-emotions posts than on single-emotion posts.

	# Posts	QWK
No emotion	11,395	0.496
Single emotion	12,281	0.590
Multiple emotions	11,324	0.697

Table 7: Agreement between sentiment polarity labels by subjective annotators and those by objective annotators.

5. Sentiment Polarity Classification

For future research and development of emotion analysis models, we evaluate the performance of baseline models based on machine learning and deep learning on the dataset in Section 3. Our experiment evaluates the performance of a five-class classification that estimates sentiment polarity $\{-2, -1, 0, 1, 2\}$ from a given text.

5.1. Setting

We divided the dataset into training set of 30k posts from 40 writers, validation set of 2.5k posts from 10 writers, and evaluation set of 2.5k posts from 10 writers. That is, there is no duplication of writers between the splits. The performance of the sentiment polarity classification models is automatically evaluated by three metrics: accuracy, mean absolute error (MAE), and quadratic weighted kappa (QWK). Two types of experiments are conducted: evaluation using labels by subjective annotators and evaluation using the average of labels by three objective annotators.

	Subjective			Objective		
	Accuracy	MAE	QWK	Accuracy	MAE	QWK
BoW+LogReg	0.344	0.924	0.359	0.443	0.695	0.444
BERT (Wikipedia)	0.386	0.824	0.512	0.573	0.483	0.695
BERT (SNS)	0.391	0.778	0.558	**0.615**	**0.426**	**0.743**
Subj. BERT (SNS)	0.391	0.778	0.558	0.443	0.646	0.627
Obj. BERT (SNS)	**0.436**	**0.694**	**0.595**	**0.615**	**0.426**	**0.743**

Table 8: Experimental results of sentiment polarity classification.

We compare the following three types of classifiers.

- BoW+LogReg: Bag-of-Words is used for feature extraction, and logistic regression is used for sentiment polarity classification. MeCab (IPADIC-2.7.0)[11] (Kudo et al., 2004) is used for word segmentation.
- BERT (Wikipedia)[12]: Japanese BERT (Devlin et al., 2019) pre-trained using Wikipedia is fine-tuned for this task and the sentiment polarity is estimated as $\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{h}\boldsymbol{W})$. Where $\boldsymbol{h}$ is the feature vector obtained from the [CLS] token of BERT and $\boldsymbol{W}$ is the trainable parameter.
- BERT (SNS)[13]: Japanese BERT pre-trained using SNS text is fine-tuned for this task.

To implement the BoW+LogReg model, we use scikit-learn[14] (Pedregosa et al., 2011). For the hyper-parameter of C, we select the optimal value over the validation set from $\{0.01, 0.1, 1, 10, 100\}$. To implement the BERT models, we use HuggingFace Transformers[15] (Wolf et al., 2020). The batch size is 32, the dropout rate is 0.1, the learning rate is 2e-5, the optimization is Adam (Kingma and Ba, 2015), and early-stopping is applied at 3 epochs.

5.2. Results

The experimental results are shown in Table 8, where BERT models outperform the BoW+LogReg model, with the best performance achieved by BERT (SNS) pre-trained with SNS text whose domain matches our dataset. The overall performance of BERT on subjective data is lower than that on objective data, indicating that the estimation of subjective sentiment polarity is more difficult.

The bottom rows of Table 8 show the results of the evaluation in objective data by subjective BERT (SNS) trained with subjective labels and evaluation in subjective data by objective BERT (SNS) trained with objective labels. These results show that the performance of models trained using objective data is consistently high, regardless of whether the evaluation target is subjective or objective data. As mentioned earlier, since it is difficult to estimate the sentiment polarity of the writer (subjective data), simple training did not provide sufficient performance.

6. Conclusion

In this study, we extended the dataset for Japanese emotion analysis[1] (Kajiwara et al., 2021), approximately doubling its size (35,000 entries) and annotating it with sentiment polarity labels. This is the first corpus of emotion analysis in Japanese annotated with both the basic emotions and sentiment polarity, and the first effort to annotate these emotion labels from both the subjective standpoint of the writer and the objective standpoint of the reader, even in other languages including English. This corpus has enabled us to conduct a comprehensive analysis of text and emotion.

Our analysis revealed that three emotions, *joy*, *anticipation*, and *trust*, are positive, while four emotions, *sadness*, *anger*,*fear*, and *disgust*, are negative. We also found that text in which multiple emotions co-occur is easier for readers to perceive the sentiment polarity of the writer than that of containing only a single emotion.

The experimental results on the sentiment polarity classification show that the estimation of subjective sentiment polarity by the writer is more difficult than the estimation of objective sentiment polarity by the reader. We also showed that it is difficult to train a high-quality sentiment polarity classification model from a labeled corpus of subjective sentiment polarity using simple supervised learning framework.

In future work, we plan to improve the performance of subjective sentiment polarity classification by multi-task learning with emotional intensity estimation and personalize by considering posting history.

Acknowledgment

This work was supported by Innovation Platform for Society 5.0 from Japan Ministry of Education, Culture, Sports, Science and Technology.

[11] https://taku910.github.io/mecab/

[12] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

[13] https://github.com/hottolink/hottoSNS-bert

[14] https://scikit-learn.org/

[15] https://github.com/huggingface/transformers

Bibliographical References

Cohen, J. (1968). Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion*, 6(3–4):169–200.

Kaji, N. and Kitsuregawa, M. (2006). Automatic Construction of Polarity-Tagged Corpus from HTML Documents. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 452–459.

Kajiwara, T., Chu, C., Takemura, N., Nakashima, Y., and Nagahara, H. (2021). WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104.

Kingma, D. P. and Ba, J. L. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Plutchik, R. (1980). A General Psychoevolutionary Theory of Emotion. *Theories of Emotion*, 1:3–31.

Scherer, K. R. and Wallbott, H. G. (1994). Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Stieglitz, S. and Dang-Xuan, L. (2013). Emotions and Information Diffusion in Social Media — Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems*, 29(4):217–248.

Strapparava, C. and Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 70–74.

Suzuki, Y. (2019). Filtering Method for Twitter Streaming Data Using Human-in-the-Loop Machine Learning. *Journal of Information Processing*, 27:404–410.

Tokuhisa, R., Inui, K., and Matsumoto, Y. (2008). Emotion Classification Using Massive Examples Extracted from the Web. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 881–888.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.