WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations

Tomoyuki Kajiwara
Graduate School of Science and Engineering
Ehime University, Japan
kajiwara@cs.ehime-u.ac.jp

Chenhui Chu
Graduate School of Informatics
Kyoto University, Japan
chu@i.kyoto-u.ac.jp

Noriko Takemura **Yuta Nakashima** **Hajime Nagahara**
Institute for Datability Science, Osaka University, Japan
{takemura, n-yuta, nagahara}@ids.osaka-u.ac.jp

Abstract

We annotate 17,000 SNS posts with both the writer's subjective emotional intensity and the reader's objective one to construct a Japanese emotion analysis dataset. In this study, we explore the difference between the emotional intensity of the writer and that of the readers with this dataset. We found that the reader cannot fully detect the emotions of the writer, especially *anger* and *trust*. In addition, experimental results in estimating the emotional intensity show that it is more difficult to estimate the writer's subjective labels than the readers'. The large gap between the subjective and objective emotions implies the complexity of the mapping from a post to the subjective emotional intensities, which also leads to a lower performance with machine learning models.

1 Introduction

Emotion analysis is one of the major NLP tasks with a wide range of applications, such as a dialogue system (Tokuhisa et al., 2008) and social media mining (Stieglitz and Dang-Xuan, 2013). Since emotion analysis has been actively studied, not only the classification of the sentiment polarity (positive or negative) of the text (Socher et al., 2013), but also more detailed emotion detection and emotional intensity estimation (Bostan and Klinger, 2018) have been attempted in recent years. Previous studies on emotion analysis use six emotions (*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*) by Ekman (1992), eight emotions (*anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust*, and *anticipation*) by Plutchik (1980), and VAD model (*Valence*, *Arousal*, and *Dominance*) by Russell (1980).

Table 1 lists datasets with emotional intensity.[1] These existing emotion analysis datasets include subjective emotional intensity labels by the writers (Scherer and Wallbott, 1994) and objective ones by the readers (Aman and Szpakowicz, 2007; Strapparava and Mihalcea, 2007; Buechel and Hahn, 2017; Mohammad and Bravo-Marquez, 2017a; Mohammad and Kiritchenko, 2018; Bostan et al., 2020), whereas the latter is mainly done by, e.g., expert or crowdsourcing annotators.

It depends on the applications whether the writer's emotions or the reader's ones to be estimated in NLP-based emotion analysis. For example, in a dialogue system, it is important to estimate the reader's emotion because we want to know how the user feels in response to the system's utterance. On the other hand, in applications such as social media mining, we want to estimate the writer's emotion. In other applications such as story generation, it is worth considering the difference between the emotions the writer wants to express and the emotions the reader receives. As shown in Table 1, most existing datasets have collected only objective emotions.[2] Therefore, previous studies on emotion analysis have focused on estimating objective emotional intensity.

In this study, we introduce a new dataset, WRIME,[3] for emotional intensity estimation. We collect both the subjective emotional intensity of the writers themselves and the objective one annotated by the readers, and explore the differences between them. In our data collection, we hired 50

[1] In this paper, the emotions of the text writers themselves are called *subjective* emotions, and the emotions that the readers receive from the text are called *objective* emotions.

[2] EmoBank (Buechel and Hahn, 2017) is a dataset that aims to collect the emotional intensity of both writers and readers. However, crowdsourcing annotators, who are different from the text writer, infer the writer's emotions, so they are not able to collect the writer's subjective emotions.

[3] Dataset of **w**riters' and **r**eaders' **i**ntensities of e**m**otion for their **e**stimation. An expanded version of 40,000 posts is available. https://github.com/ids-cv/wrime

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2095–2104
June 6–11, 2021.

	Emotion	Intensity	Subj.	Obj.	Language	Size
ISEAR (Scherer and Wallbott, 1994)	E6	n/a	✓	×	English	7,666
Blogs (Aman and Szpakowicz, 2007)	E6	{Low, Med., High}	×	✓	English	5,025
SemEval-2007 (Strapparava and Mihalcea, 2007)	E6	$[0, 100]$	×	✓	English	1,250
WASSA-2017 (Mohammad and Bravo-Marquez, 2017b)	M4	$[0, 1]$	×	✓	English	7,097
SemEval-2018 (Mohammad et al., 2018)	M4	$[0, 1]$	×	✓	English	12,634
EmoBank (Buechel and Hahn, 2017)	VAD	$\{1, 2, 3, 4, 5\}$	×	✓	English	10,062
GoodNewsEveryone (Bostan et al., 2020)	P8	{Low, Med., High}	×	✓	English	5,000
WRIME (Ours)	P8	$\{0, 1, 2, 3\}$	✓	✓	Japanese	17,000

Table 1: List of datasets with emotional intensity. In the "Emotion" column, datasets with E6 adopt the six emotions by Ekman (1992): *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, ones with P8 adopts the eight emotions by Plutchik (1980): *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust*, *anticipation*, and ones with M4 adopts the four emotions by Mohammad et al.: *joy*, *sadness*, *anger*, *fear*.

participants via crowdsourcing service. They annotated their own past posts on a social networking service (SNS) with the subjective emotional intensity. We also hired 3 annotators, who annotated all posts with the objective emotional intensity. Consequently, our Japanese emotion analysis dataset consists of 17,000 posts with both subjective and objective emotional intensities for Plutchik's eight emotions (Plutchik, 1980), which are given in a four-point scale (no, weak, medium, and strong).

Our comparative study over subjective and objective labels demonstrates that readers may not well infer the emotions of the writers, especially of *anger* and *trust*. For example, even for posts written by the writer with a strong *anger* emotion, our readers (i.e., the annotators) did not assign the *anger* label at all to more than half of the posts with the subjective *anger* label. Overall, readers may tend to underestimate the writers' emotional intensities. In addition, experimental results on emotional intensity estimation with BERT (Devlin et al., 2019) show that predicting the subjective labels is a more difficult task than predicting the objective ones. This large gap between the subjective and objective annotations implies the challenge in predicting the subjective emotional intensity for a machine learning model, which can be viewed as a "reader" of the posts.

2 Related Work

To estimate the emotional intensity of the text, datasets labeled with Ekman's six emotions (Ekman, 1992) and Plutchik's eight emotions (Plutchik, 1980) has been constructed for languages such as English, as shown in Table 1. EmoBank[4] (Buechel and Hahn, 2017), which is most relevant to ours, labels the emotional intensity of both the writers and readers of the text. However, the annotators for EmoBank are not writers, and readers are required to guess the writer's emotion; therefore, to be strict, this dataset only contains the objective labels. Our dataset is the first to collect the subjective emotional intensity of the writers themselves.

ISEAR (Scherer and Wallbott, 1994) is a dataset with subjective emotional labels. This is a dataset in which annotators describe their own past events in each emotion. They use a label set that adds *shame* and *guilt* to Ekman's six emotions. Although ISEAR is the only dataset with subjective emotional labels, their intensity is not considered.

Early studies in collecting objective emotional labels were annotated by experts. Aman and Szpakowicz (2007) labeled each sentence of English blog posts with Ekman's six emotions and their intensity on a three-point scale. Strapparava and Mihalcea (2007) labeled Ekman's six emotional intensities to English news headlines and held a competition of SemEval-2007 Task 14.[5]

In recent years, there have been many studies on collecting objective emotional labels using crowdsourcing. Mohammad and Bravo-Marquez (2017a); Mohammad and Kiritchenko (2018) labeled tweets in English, Arabic, and Spanish with the intensity of four emotions (*joy*, *sadness*, *anger*, and *fear*). Using these datasets, they held a series of competitions to estimate the emotional intensity in WASSA-2017 Shared Task on Emotion Intensity[6] (Mohammad and Bravo-Marquez, 2017b) and SemEval-2018 Task 1[7] (Mohammad et al., 2018).

[4] https://github.com/JULIELab/EmoBank

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

[6] https://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html

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

	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Overall
Reader 1 vs. Reader 2	0.697	0.607	0.594	0.342	0.627	0.359	0.527	0.203	0.547
Reader 1 vs. Reader 3	0.662	0.545	0.567	0.443	0.581	0.429	0.455	0.196	0.549
Reader 2 vs. Reader 3	0.700	0.597	0.632	0.415	0.630	0.476	0.512	0.295	0.585
Writer vs. Reader 1	0.622	0.461	0.423	0.348	0.363	0.333	0.394	0.089	0.439
Writer vs. Reader 2	0.633	0.526	0.432	0.339	0.386	0.361	0.442	0.153	0.465
Writer vs. Reader 3	0.624	0.450	0.459	0.396	0.374	0.380	0.467	0.134	0.463
Writer vs. Avg. Readers	0.683	0.536	0.498	0.441	0.401	0.433	0.514	0.132	0.515

Table 2: Inter-annotator agreement by quadratic weighted kappa.

Some datasets (Kaji and Kitsuregawa, 2006; Suzuki, 2019) are available in Japanese. However, these are sentences with sentiment polarity, and do not cover the various emotions dealt with in this study. Our study is the first to label Japanese texts with various emotional intensity.

3 Emotional Intensity Annotation

3.1 Annotating Subjective Labels

We hired 50 participants via crowdsourcing service *Lancers*.[8] Those participants include 22 men and 28 women, where 2 are teens, 26 are in their 20s, 18 are in their 30s, and 4 are above 40 years old. They copy and paste their own past SNS posts and then labeled the posts with the subjective emotional intensity according to Plutchik's eight emotional intensities (Plutchik, 1980) with a four-point scale (0: no, 1: weak, 2: medium, and 3: strong). They did not provide us with all the posts, but chose only those posts that they could agree to publish. Here, for the purpose of emotion analysis from the text, posts with images or URLs were excluded. Each participant labeled 100 to 500 posts, resulting in 17,000 posts in total. We did not limit the posts to be annotated based on when they are posted. As a result, our dataset contains posts in the 9-year range from June 2011 to May 2020. We assumed that each post would require 50 seconds for annotation and paid 21.5 JPY per post. This roughly corresponds to 15 USD per hour, which is a good reward for crowdsourcing.[9]

To assess the quality of annotations, we randomly sampled 30 posts for each participant. One of our graduate students evaluated the posts and the corresponding eight emotional intensity labels on a four-point scale based on the following criteria.

- 3: I fully agree with the label given.
- 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 score for each participant was 2.1, where 1.8 at minimum, and 2.5 at maximum. There were no posts rated as 0. We had five annotators whose average score was below 2, but reviewing their posts and labels does not necessarily show obvious clues of improper annotation.

3.2 Annotating Objective Labels

We hired three objective annotators via the same crowdsourcing service as in Section 3.1. Annotators include two women in their 30s and one woman in their 40s. They labeled all the 17,000 posts with Plutchik's eight emotional intensities (Plutchik, 1980) in the same way as subjective annotation. Note that while the subjective annotators labeled their own emotions as the writer of each post, the objective annotators labeled each post based on the emotions they received from the post. Objective annotators do not have to fill in the text, so their task is simply to label emotional intensity. We assumed that each post takes 10 seconds and paid 3.8 JPY per post, which results in the reward of roughly 13 USD per hour.

To assess the quality of annotations, we calculated the quadratic weighted kappa[10] (Cohen, 1968) as a metric of the inter-annotator agreement. The upper part of Table 2 shows the agreement between the objective annotators. The best case, *joy*, shows

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

[9] One of the popular crowdsourcing services, Prolific, has a minimum payment of 6.5 USD per hour. https://www.prolific.co/pricing

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

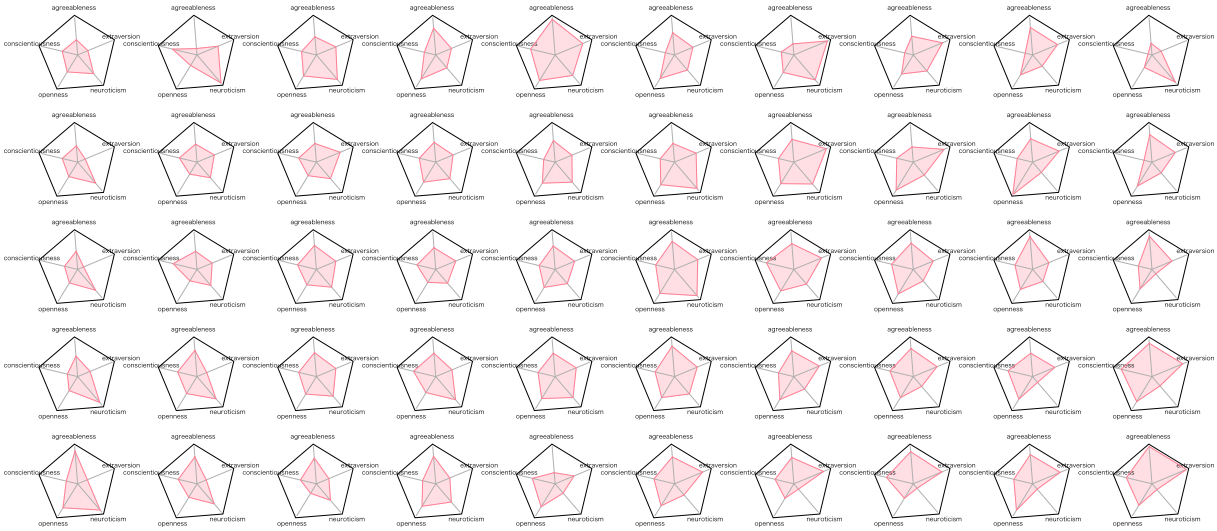

Figure 1: Results of personality diagnosis. Clockwise from top: agreeableness, extraversion, neuroticism, openness, conscientiousness.

Text	Today is the perfect weather. I'll clean and play.							
Writer	joy: 3	sadness: 0	anticipation: 3	surprise: 0	anger: 0	fear: 0	disgust: 0	trust: 1
Reader 1	joy: 3	sadness: 0	anticipation: 3	surprise: 0	anger: 0	fear: 0	disgust: 0	trust: 0
Reader 2	joy: 2	sadness: 0	anticipation: 2	surprise: 0	anger: 0	fear: 0	disgust: 0	trust: 0
Reader 3	joy: 3	sadness: 0	anticipation: 3	surprise: 0	anger: 0	fear: 0	disgust: 0	trust: 0
Text	The tire of my car was flat. I heard that it might be mischief.							
Writer	joy: 0	sadness: 3	anticipation: 0	surprise: 1	anger: 3	fear: 0	disgust: 0	trust: 0
Reader 1	joy: 0	sadness: 3	anticipation: 0	surprise: 3	anger: 1	fear: 2	disgust: 1	trust: 0
Reader 2	joy: 0	sadness: 2	anticipation: 0	surprise: 2	anger: 0	fear: 0	disgust: 0	trust: 0
Reader 3	joy: 0	sadness: 2	anticipation: 0	surprise: 2	anger: 0	fear: 1	disgust: 1	trust: 0

Table 3: Examples of our dataset.

a substantial agreement ($\kappa > 0.6$), but *trust*, is with a fair agreement ($\kappa < 0.4$). Overall, we confirmed a moderate agreement ($0.5 < \kappa < 0.6$) among the objective annotators.

The lower part of Table 2 shows the agreement between the subjective and the objective annotators. These are discussed in Section 4.2.

3.3 Writers' Personality Assessment

We also performed personality assessments of our writers (i.e., subjective annotators) in order to explore the relationship between personality and emotion. Through 60 questions (Saito et al., 2001) based on the Big Five personality traits (Goldberg, 1992), the following five factors were assessed: agreeableness, extraversion, neuroticism, openness, and conscientiousness. In this personality assessment, the writer's own applicability to each of 60 adjectives, such as "cheerful" and "honest" is reported on a 7-point scale, and the five factors of personality indicators are derived.

Figure 1 shows the results of the personality assessment over all 50 writers, where we can see various personalities. For example, well-balanced writers can be seen near the center of the figure, and writers with low neuroticism appear in the lower right. In Section 5, we shall show how the personality helps to improve emotional intensity estimation.

4 Analysis

Table 3 shows some examples of labeled posts in our dataset. The first post was written with a strong emotions of both *joy* and *anticipation*. Readers can have similar emotions as the writer for this post. The second post was written with a strong emotions of both *sadness* and *anger*. Readers can share emotions of *sadness*, but they are more *surprised* than *angry*.

	Joy				Sadness				Anticipation			
Intensity	W	R1	R2	R3	W	R1	R2	R3	W	R1	R2	R3
0	9,942	12,043	12,379	11,213	10,472	13,205	11,961	12,559	9,991	11,714	12,509	10,796
1	2,454	291	1,074	1,397	2,837	389	1,881	2,123	2,996	610	1,245	2,683
2	2,283	2,285	2,055	3,475	2,140	2,127	2,168	1,846	2,172	2,507	1,825	2,119
3	2,321	2,381	1,492	915	1,551	1,279	990	472	1,841	2,169	1,421	1,402

	Surprise				Anger				Fear			
Intensity	W	R1	R2	R3	W	R1	R2	R3	W	R1	R2	R3
0	11,148	10,534	14,143	10,974	14,408	16,278	16,180	16,223	13,355	13,285	15,163	13,877
1	2,605	997	1,429	2,840	1,284	156	304	311	1,815	384	626	1,478
2	1,778	3,234	971	2,027	661	285	315	266	1,082	2,032	838	1,070
3	1,469	2,235	457	1,159	647	281	201	200	748	1,299	373	575

	Disgust				Trust				Overall			
Intensity	W	R1	R2	R3	W	R1	R2	R3	W	R1	R2	R3
0	13,258	14,538	14,333	11,959	12,682	16,165	15,920	15,979	95,256	107,762	112,588	103,580
1	1,882	449	1,248	2,436	2,162	470	466	609	18,035	3,746	8,273	13,877
2	959	1,190	934	1,242	1,239	239	395	300	12,314	13,899	9,501	12,345
3	901	823	485	1,363	917	126	219	112	10,395	10,593	5,638	6,198

Table 4: Distribution of emotional intensity labels. W, R1, R2, and R3 represent writer, reader 1, reader 2, and reader 3, respectively.

4.1 Distribution of Emotional Intensity

Table 4 shows the distribution of emotional intensity labels. For all emotions, intensity 0 is most frequently assigned. This is not surprising, as it is rare for a single post to come with many emotions, which may be contradictory to each other, at the same time.[11] However, for emotions of *anger* and *trust*, about 95% of labels by the objective annotators have an intensity 0, which is particularly high. In other words, with regard to emotions of *anger* and *trust*, readers may tend to underestimate the emotions of the writers. In addition, we can see some characteristics of each objective annotator, e.g., the number of times that reader 1 gives intensity 1 is small.

4.2 Difference between Writers and Readers

The lower part of Table 2 shows the agreement between the subjective and the objective annotators. As with the agreement between the objective annotators in Section 3.2, we calculated the quadratic weighted kappa (Cohen, 1968). Agreement between subjective and objective annotators are lower than agreement between objective annotators (the upper part of Table 2). Especially for the emotion of *anger*, there is a large gap between the reader–reader agreements and writer–reader agreements. In addition, for the emotion of *trust*, the writer–reader agreement is even lower, although the reader–reader agreements are also low. These results imply that there is a large difference between the subjective and objective emotion.

Table 5 shows the confusion matrix between the subjective emotional intensity labels and the objective ones for respective emotions. For example, in posts where the writer labeled intensity 0 for *joy*, the percentages where the reader labeled intensities 0, 1, 2, and 3 were 91.7%, 3.1%, 4.0%, and 1.2%, respectively. This confusion matrix shows the fine-grained differences in emotional intensity between writers and readers, which reinforces our discussion in Section 3 that readers hardly detect the emotions associated with the post. Focusing on the emotion of *anger* in the confusion matrix, in 58.6% of the posts where the writer labeled intensity 3 (strong *anger*), the reader labeled intensity 0 (no emotions of *anger*). This is more prominent in the emotion of *trust*: for 81.5% of posts that the writer labeled intensity 3, the reader labeled intensity 0. This clearly demonstrates that the readers cannot infer the emotion *trust* of the writer. As for other emotions, readers are most likely to label an intensity 0 in posts labeled with an intensity 2 or less by the writer. Overall, the readers tend to underestimate the writer's emotions, and they rarely label intensity 1 or more when the writer label intensity 0.

[11] 90% of posts have less than 4 emotions at the same time.

	Joy				Sadness				Anticipation			
Writer \ Reader	0	1	2	3	0	1	2	3	0	1	2	3
0	91.7	3.1	4.0	1.2	90.2	4.7	4.0	1.1	84.1	6.7	6.0	3.1
1	60.7	9.7	22.5	7.0	57.9	15.7	19.9	6.5	55.2	13.5	19.7	11.6
2	37.2	9.4	34.1	19.3	45.1	15.4	26.8	12.7	46.8	11.5	23.3	18.3
3	18.2	6.6	37.7	37.4	33.6	12.4	31.7	22.3	32.4	10.1	24.6	32.8

	Surprise				Anger				Fear			
Writer \ Reader	0	1	2	3	0	1	2	3	0	1	2	3
0	80.9	7.8	7.8	3.5	98.6	0.7	0.5	0.2	89.0	3.9	5.1	2.1
1	56.5	16.0	17.7	9.7	87.5	5.2	4.8	2.5	70.3	8.7	14.0	7.0
2	48.8	14.8	22.0	14.3	77.4	7.2	9.6	5.8	57.5	9.3	19.7	13.5
3	35.8	14.0	23.8	26.3	58.6	6.7	15.2	19.5	44.2	6.9	22.8	26.1

	Disgust				Trust				Overall			
Writer \ Reader	0	1	2	3	0	1	2	3	0	1	2	3
0	87.9	6.3	3.8	2.0	96.2	2.2	1.1	0.5	90.2	4.3	3.9	1.6
1	62.2	16.1	13.1	8.6	92.6	4.2	2.2	1.1	65.6	11.8	15.4	7.2
2	49.1	14.5	18.9	17.5	86.3	6.5	5.1	2.0	51.8	11.6	22.5	14.1
3	34.7	11.4	21.3	32.6	81.5	7.2	6.4	4.8	36.9	9.7	25.9	27.6

Table 5: Confusion matrix of subjective and objective labels. (%) This is a total of the three sub-matrices. Each sub-matrix is a confusion matrix for each reader.

5 Emotional Intensity Estimation

We conduct experiments on the four-class classification as an ordinal classification to estimate emotional intensity {0, 1, 2, 3} using the dataset constructed in Section 3.

5.1 Experimental Settings

In this experiment, we divided the dataset[12] into training set of 15,000 posts from 30 writers, validation set of 1,000 posts from 10 writers, and evaluation set of 1,000 posts from 10 writers. That is, there is no duplication of writers between the splits. We used MeCab (IPADIC-2.7.0)[13] (Kudo et al., 2004) to tokenize Japanese text.

The performance of the emotional intensity estimation models is evaluated by the mean absolute error (MAE) and the quadratic weighted kappa (QWK). We evaluated the model using both the emotional intensity labels given by the subjective annotators (subjective labels) and the average of the emotional intensity labels given by the three objective annotators (objective labels).

Following the standard emotional intensity estimation models (Acheampong et al., 2020), we train the following three types of four-class classification models for each emotion.

- **BoW+LogReg** employs Bag-of-Words to extract features and Logistic Regression to the estimate emotional intensity.
- **fastText+SVM** vectorizes each word with fastText[14] (Bojanowski et al., 2017) and estimates the emotional intensity with a Support Vector Machine based on their average vector.
- **BERT** is a model that fine-tunes the pre-trained BERT[15] (Devlin et al., 2019) and estimates the emotional intensity as $\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{h}\boldsymbol{W})$, where $\boldsymbol{h}$ is a feature vector obtained for the [CLS] token of BERT. We investigate the performance of both BERT trained with subjective labels (Subj. BERT) and BERT trained with objective labels (Obj.

[12] Each writer provided 500 posts for the training set and 100 posts for the validation and test sets.

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

[14] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.ja.300.bin.gz

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

Subjective labels	MAE									QWK
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Overall	Overall
Random	1.390	1.383	1.419	1.313	1.492	1.420	1.411	1.407	1.404	0.001
Modal Class	0.896	0.713	0.907	0.684	**0.218**	0.344	0.435	0.429	0.578	0.000
BoW+LogReg	0.863	0.817	0.919	0.752	0.313	0.479	0.545	0.555	0.655	0.156
fastText+SVM	0.896	0.754	0.910	0.723	0.250	0.397	0.489	0.510	0.616	0.120
Subj. BERT	0.734	0.666	0.899	0.684	**0.218**	0.344	0.443	0.432	0.553	0.135
Subj. BERT w/ Pc	0.784	0.698	0.870	0.659	**0.218**	**0.343**	0.457	0.429	0.557	0.153
Subj. BERT w/ Pa	0.740	0.665	0.850	0.665	**0.218**	0.351	0.441	0.429	0.545	0.183
Obj. BERT	**0.674**	**0.623**	**0.789**	**0.634**	**0.218**	0.356	**0.432**	**0.427**	**0.519**	**0.242**
Reader 1	0.545	0.544	0.713	0.686	0.211	0.523	0.522	0.428	0.522	0.417
Reader 2	0.521	0.520	0.720	**0.571**	0.201	**0.347**	**0.375**	0.426	**0.460**	0.442
Reader 3	0.526	0.533	0.738	0.694	0.200	0.610	0.520	0.432	0.532	0.439
Avg. Readers	**0.491**	**0.466**	**0.658**	0.584	**0.198**	0.458	0.420	**0.425**	0.463	**0.486**

Table 6: Evaluation of MAE and QWK in estimating subjective emotional intensity.

BERT), in both evaluations on subjective and objective labels.

We also evaluate the following two baselines.

- **Random** outputs one of the four emotional intensity labels {0, 1, 2, 3} randomly with the uniform distribution.
- **Modal Class** always outputs the most frequent intensity label for each emotion. As shown in Table 4, in this dataset, intensity 0 has the highest frequency for all emotions, so in practice, this baseline always gives label 0.

We used scikit-learn[16] (Pedregosa et al., 2011) implementation for both BoW+LogReg and fastText+SVM models. For the hyper-parameter of C, the optimum value over the validation set was selected from {0.01, 0.1, 1, 10, 100}.

As for BERT-based models, we used the implementation in Transformers[17] (Wolf et al., 2020). We used the whole-word-masking model with a batch size of 32, a dropout rate of 0.1, a learning rate of 2e-5, and Adam (Kingma and Ba, 2015) for optimization. The training stopped after 3 epochs without improvement in the validation loss.

In the evaluation of subjective labels, the personality of the writers is considered in the Subj. BERT in the following two ways.

- **w/ Pc**: Feature extraction is performed with $\boldsymbol{h}_c = [\boldsymbol{u}; \boldsymbol{v}]\boldsymbol{W}^c$ in consideration of personality. Here, $\boldsymbol{v}$ is a 768-dimensional text representation obtained from the [CLS] token of BERT, and $\boldsymbol{u}$ is a representation of the Big Five personality indicators given by linearly transforming the five indicator values into a 768-dimensional vector. When estimating the emotional intensity, $\boldsymbol{h}_c$ is used instead of $\boldsymbol{h}$.
- **w/ Pa**: Feature extraction is performed with $\boldsymbol{h}_a = \mathrm{attention}(\boldsymbol{u}\boldsymbol{W}^Q, \boldsymbol{v}\boldsymbol{W}^K, \boldsymbol{v}\boldsymbol{W}^V)$ in consideration of personality. That is, in the calculation of the attention mechanism, the personality representation $\boldsymbol{u}$ is used as the query, and the text representation $\boldsymbol{v}$ is used as both the key and the value. $\boldsymbol{h}_a$ is used instead of $\boldsymbol{h}$ for emotional intensity estimation.

5.2 Results

The performance of each model on subjective and objective labels is shown in Tables 6 and 7, respectively. Regardless of the method, the evaluation of subjective label estimation gets a larger mean absolute error than the evaluation of objective labels. In our previous discussion, we have stated that it is difficult for readers to estimate the emotions of writers; this also applies to machine learning models.

5.2.1 Evaluation with Subjective Labels

In the evaluation of subjective labels, the traditional models of BoW+LogReg and fastText+SVM achieved lower mean absolute errors than the Random baseline, but were inferior to the Modal Class baseline. The BERT methods achieved a mean absolute error lower than the Modal Class baseline. Surprisingly, Obj. BERT trained with objective labels, rather than Subj. BERT trained with subjective labels, achieved the highest performance.

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

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

Objective labels	MAE									QWK
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Overall	Overall
Random	1.353	1.333	1.333	1.291	1.516	1.354	1.391	1.450	1.378	0.001
Modal Class	0.595	0.459	0.713	0.518	**0.044**	0.420	0.383	0.026	0.395	0.000
BoW+LogReg	0.560	0.460	0.606	0.525	0.064	0.432	0.412	0.057	0.390	0.277
fastText+SVM	0.595	0.459	0.713	0.518	0.059	0.406	0.383	0.051	0.398	0.040
Subj. BERT	0.489	0.446	0.685	0.518	**0.044**	0.414	0.381	0.039	0.377	0.209
Obj. BERT	**0.403**	**0.411**	**0.475**	**0.442**	**0.044**	**0.386**	**0.348**	**0.024**	**0.317**	**0.442**
Reader 1	**0.222**	0.246	0.277	**0.276**	**0.035**	**0.233**	**0.226**	0.053	**0.196**	**0.830**
Reader 2	0.224	0.264	**0.268**	0.349	0.045	0.277	0.269	0.027	0.215	0.769
Reader 3	0.237	**0.241**	0.332	0.360	0.046	0.346	0.310	**0.021**	0.237	0.808

Table 7: Evaluation of MAE and QWK in estimating objective emotional intensity.

Since it is difficult to estimate subjective labels, which are the emotion of the writer, a simple model may not provide sufficient performance.

Therefore, we examined Subj. BERT w/ Pc and Subj. BERT w/ Pa to assist training using the personality information of the writer. As a result, Subj. BERT w/ Pc, which simply concatenates the personality representation and the text representation, was not effective, but Subj. BERT w/Pa, which considers personality representation with weighting, achieved higher performance than simple Subj. BERT. The evaluation by QWK also shows the usefulness of using the personality information of the writer. However, even with personality information, the performance is not comparable with that of Obj. BERT. Improving methods for accurate estimation of subjective emotions is our future work.

Below the dotted line in Table 6, the performance of the human readers is shown for comparison. Estimating the emotional intensity of writers is difficult for both human readers and machine learning models.

5.2.2 Evaluation with Objective Labels

In the evaluation of objective labels (Table 7), the traditional models of BoW+LogReg and fastText+SVM were comparable to the Modal Class baseline. Similar to the evaluation in the subjective labels, the BERT-based models achieved mean absolute errors lower than the Modal Class baseline, and Obj. BERT achieved the highest performance.

Below the dotted line in Table 7, the performance of the human readers is shown for comparison. Note that the objective labels are the average of each of these readers. Compared to each reader, Obj. BERT does not reach human performance.

6 Conclusion

We introduce a new dataset, WRIME, for Japanese emotional intensity estimation. Our dataset is based on Plutchik's eight emotions (Plutchik, 1980), labeling both the writer's subjective emotional intensity and the reader's objective one in SNS posts.

Overall, the readers tend to underestimate the writer's emotions. Even the strong emotions of the writer cannot be detected by the reader, especially in the emotions of *anger* and *trust*.

Experimental results on emotional intensity estimation show that it is more difficult to estimate the writer's subjective labels than the readers'. The large gap between the subjective and objective emotions imply the complexity of the mapping from a text to the subjective emotional intensities, which also leads to a lower performance with machine learning models.

Estimating the writer's subjective emotions with higher accuracy is future work. We have shown the possibility of improving the performance of subjective emotional intensity estimation by considering the personality of the writer. It may be worth considering the writer's meta information, including personality, and the writer's past posting history.

Ethical Considerations

We ensure that our work is conformant to the ACM Code of Ethics.

Acknowledgments

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

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based Emotion Detection: Advances, Challenges, and Opportunities. *Engineering Reports*, 2(7):1–24.

Saima Aman and Stan Szpakowicz. 2007. Identifying Expressions of Emotion in Text. In *International Conference on Text, Speech and Dialogue*, pages 196–205.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Laura-Ana-Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 578–585.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

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

Lewis R. Goldberg. 1992. The Development of Markers for the Big-Five Factor Structure. *Psychological Assessment*, 4(1):26–42.

Nobuhiro Kaji and Masaru Kitsuregawa. 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.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 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.

Saif Mohammad and Felipe Bravo-Marquez. 2017a. Emotion Intensities in Tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pages 65–77.

Saif Mohammad and Felipe Bravo-Marquez. 2017b. WASSA-2017 Shared Task on Emotion Intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49.

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

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 198–209.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

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

James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Takako Saito, Tomoyasu Nakamura, Toshihiko Endo, and Madoka Yokoyama. 2001. Standardization of Big Five Scales Using the Adjective Check List. *Kyushu University Psychological Research*, 2:135–144.

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

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 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.

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

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

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

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion Classification Using Massive Examples Extracted from the Web. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 881–888.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.