A Classification Method for Japanese Sentences based on the Difficulty Level of Emotion Estimation

Sanae Yamashita[1], Yasushi Kami[1] and Noriyuki Okumura[2]
[1]*National Institute of Technology, Akashi College, 679–3 Nishioka, Uozumi-cho, Akashi-shi, Hyogo, 674–8501, Japan*
[2]*Otemae University, 6–42 Ochayasho-cho, Nishinomiya-shi, Hyogo, 662–8552, Japan*

Keywords: Emotion Extraction, Emotion Estimation, Response Time, Annotation.

Abstract: The existing systems to estimate emotions extract some emotions from the given sentences in any and all circumstances. However, there are many sentences whoever cannot estimate emotional features. It follows that the systems should not extract some emotions all the time. Systems should return "It is difficult to estimate" as we humans do so. This paper proposes a method to classify Japanese sentences based on the difficulty level of emotion estimation. Proposed system judges the difficulty level to estimate emotions using three conditions (negative expressions, emotive expression, and machine-learned classifications). As a result, proposed system achieved 0.8 of F_1 score based on mechanical evaluation.

1 INTRODUCTION

With the spread of SNS, there have been increasing in the opportunities of being read the messages we wrote. However, unintentional spreading and slandering of messages are increasing concurrently. Because of the messages on SNS frequently contain writer's emotions, it may be possible to prevent these problems if we can analyze the emotions included in the messages (Matsubayashi et al., 2016).

Many of the existing systems of estimating emotions have been presuming that they can always estimate some kind of emotion from given sentences. However, we found there are many difficult sentences for humans to estimate emotions actually. The interactive system should provide results not any emotions compulsorily but such as "difficult to estimate" in the same way as a human.

In this paper, we constructed a classification method for Japanese sentences to estimate the difficulty level to extract writer's emotion. The system consists of a combination of several decision conditions. For example, a sentence including any negative expressions is marked as "high difficulty" to estimate emotions. Also, if a sentence includes any emotive expressions, we regard it as "low difficulty." Accordingly, in this system, it decisions the difficulty of emotion estimation from Japanese sentences by the following combination: the existence of negative expressions, existence of emotive expressions, and prediction by machine-learned classifiers.

2 RELATED WORKS

Section 2.1 introduces some emotion classify methods being used by existing research. Section 2.2 presents the methods for determining whether or not sentences have any negative expressions.

2.1 Classify Emotions

Emotion classifications vary. One case, in Emotive Expression Dictionary (感情表現辞典), emotions are classified into 10 classes: 喜 (*joy*), 怒 (*anger*), 哀 (*sorrow*), 怖 (*fear*), 恥 (*shame*), 好 (*liking*), 厭 (*dislike*), 昂 (*excitement*), 安 (*relief*), 驚 (*surprise*). Ptaszynski uses this classification to his emotion analysis system ML-Ask (Ptaszynski et al., 2017).

Other cases, emotion models in psychology are also often used. Hasegawa and Saravia use Plutchik's wheel of emotions shown in Figure 1 (Hasegawa et al., 2014; Saravia et al., 2018). This model has 8 basic emotions, and each emotion has 3 levels. In the case of *joy*, *serenity*, *joy*, and *ecstasy* are subdivided emotions. The basic 8 emotions are *joy*, *sadness*, *trust*, *disgust*, *anger*, *fear*, *anticipation*, *sur-*

Yamashita, S., Kami, Y. and Okumura, N.
A Classification Method for Japanese Sentences based on the Difficulty Level of Emotion Estimation.
DOI: 10.5220/0008366303830390
In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, pages 383-390
ISBN: 978-989-758-382-7

Figure 1: Plutchik's wheel of emotions (Excerpt).

prise. These make 4 pairs as *joy-sadness*, *trust-disgust*, *anger-fear*, *anticipation-surprise*, and emotions in the pair are regarded to have the opposite meaning, respectively. This feature is useful in emotion systems, thus often be used as an emotion classify method. Also in this paper, Plutchik's wheel of emotions is used.

Jung sugggests the word-association method. With this method, he tests what word is associated and how long is the response time to associate for a word. White shows the response time of pleasant words is shorter than the time of unpleasant words (White and Powell, 1936).

2.2 Detect the Negative Expressions

About Japanese sentences, most of the research on detecting the existence of negative expressions need the information of a part of speech (POS) of words. We use morphological analyzers [1] to extract the POS data, however, it is impossible to detect negative expressions, not in a dictionary of the analyzer, and also they cannot necessarily estimate POS of words or divide to words correctly. Therefore we research the method to decide that Japanese sentences have some negative expressions or not without depending POS data only (Yamashita et al., 2018). Consequently, Naive Bayes trained by following features is helpful: matching by POS and basic form, the position of representative negative expressions and negative polarity item (NPI) in the sentence.

[1]MeCab (http://taku910.github.io/mecab) and its dictionary NEologd (https://github.com/neologd/mecab-ipadic-neologd)

3 FEATURES WHEN HUMAN ESTIMATES EMOTIONS

In this section, to discover the feature when people estimate some emotions from sentences, we estimated the writer's emotions from Japanese sentences and annotated them manually. From the required time to annotate and the accuracy rate, the difficulties of estimation on each emotion and the features of the high difficulty sentences. In this section, 3.1 describes the method of experiment, 3.2 describes the result, 3.3 describes the features got from this experiment.

3.1 How to Examine the Features

This section shows the method of this experiment. 3.1.1 describes the classify methods of emotions, 3.1.2 describes the measure of annotation time, 3.1.3 describes the evaluated data and the annotating method.

3.1.1 Emotion Classify

For the annotation, we adopt Plutchik's Wheel of Emotions introduced in section 2.1. This model divides the basic 8 emotions into 3 levels for each emotion, but in this experiment the basic 8 emotions intact because subdivided emotions are not appropriate to manual annotation.

When annotating, 8 emotions are divided into 4 groups from the feature that each emotion of Plutchik's model has an opposite emotion, then select an emotion from each group including no emotions (*none*). The details of the groups are below.

- *joy*, *sadness*, and *none*
- *trust*, *disgust*, and n*one*
- *anger*, *fear*, and *none*
- *anticipation*, *surprise*, and *none*

3.1.2 Annotation Time Measurement

It is conceivable that the shorter the period people spend on annotating (annotation time), the easier the level of difficulty. For this reason, all logs of annotation time when chose emotions are recorded. In particular, the passage of time from displaying evaluated data to choosing each emotion in the selecting emotion area by clicking or tapping is recorded. For example, choosing again as the following case is also included.

Table 1: The examples of evaluated data.

Japanese (original)	English (translated)
袴たーのしー!	*hakama is enjoyable!*
気がついたら帰るには寒い時間になっちゃってつらい	*I'm very sorry because it's too cold to come back home when I realized*
今世紀最高の寝覚めを更新	*renewed the best awakening in this century*
お花見したいな [emoji:cherry_blossom]	*I'd like to go cherry-viewing*
@[USER] ありがとう、がんばる	*thanks, I'll do my best*

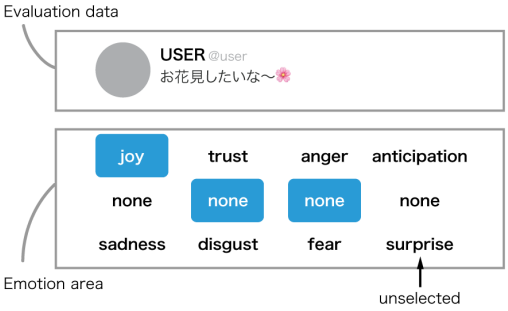

Figure 2: The view of the annotating application.

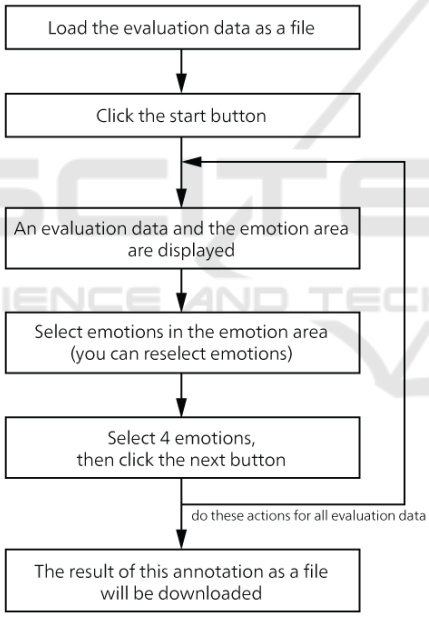

Figure 3: A flow of the annotation.

1. Select *joy*,
2. Cancel *joy*,
3. And then select *sadness*.

3.1.3 Annotation Method

For the evaluated data, the author's tweets (1,000 tweets, posted from 2018/01 to 2018/03) [2] are used. Table 1 shows these data. They are annotated emotions.

[2]https://twitter.com/yamasy1549

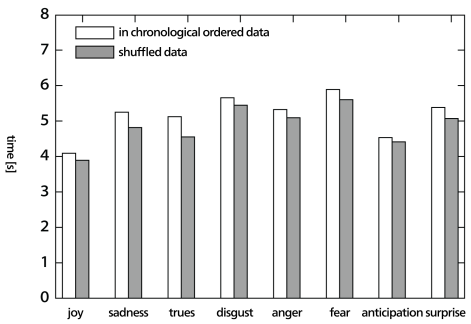

Figure 4: The median of annotation time for each emotion.

A web application is constructed for recording selected emotions with annotation time. Figure 2 shows the view of the application and Figure 3 shows the order to annotation. Selecting emotion area has 4 subareas, *joy-sadness*, *trust-disgust*, *anger-fear*, and *anticipation-surprise*, are reordered randomly on each evaluated data. When the position of the subareas are fixed, if an annotator tends to annotate from left to right, the right subarea is the possibility of being had some bias. By default, all emotions are not selected and then annotators select 4 emotions totally from each column one by one.

In this experiment author's tweets are used for evaluated data. Annotators were 26 persons, including 4 acquaintances of the author, by an above application for annotation.

3.2 Annotation Time and Accuracy of each Emotion

The medians of annotation time for each emotion are shown in Figure 4. For the view of annotation time, *joy* is an easy emotion to estimate relatively.

The accuracy of each emotion is shown in Figure 5. *joy*, *sadness*, and *anticipation* are high accuracies. Especially *joy* is regarded to be an easy emotion to estimate.

3.3 Feature of Difficult Sentence to Estimate Emotion

This section describes the features when people estimate emotions from sentences got from this experi-

Table 2: The examples of evaluated data.

Feature	Japanese (original)	English (translated)
Only proper noun	東加古川高専です 応用起床技術者試験	*Higashi-Kakogawa KOSEN is* *Applied Waking Up Technology Engineer Examination*
Only onomatopoeia	おっ ふぇぇ	*oh* *ew*
Including an intent of question	黒軸ってやっぱ長く使うと疲れるんですかね	*will I get tired when using CHERRYMX Black for a long time?*
Suggesting fact	部屋寒いので重たいのブン回して暖を取っている	*because the room is cold, executing heavy processing and keeping warm*

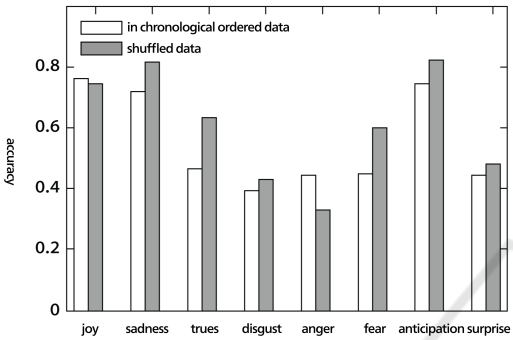

Figure 5: The accuracy of each emotion.

ment. As the example described in section 3.1.2, we focus on the case of reselecting emotions. When an annotator selects some emotions and reselects *none*, the evaluated data is regarded to be a high difficulty for the annotator. We examined these data and recovered the following features. Table 2 shows these features with some example sentences.

We decide the sentences consist only proper nouns as high difficulty without having any words associated specific emotions strongly: a name of a theme park is associated with *joy*. On the sentences consist only onomatopoeia, for example, an interjection おっ (*oh*) is able to estimate both of *surprise* and *anticipation*. Therefore without considering around context, the difficulty of emotion estimation of the only sentences.

4 DETECT THE EMOTIVE EXPRESSIONS

The objective of this experiment is to get a standard to decide whether or not a given sentence includes emotive expressions. Section 4.1 describes the method of this experiment, and section 4.2 describes the results and considerations.

4.1 Definition of the Emotive Expressions

We define emotive expressions as words or phrases suggesting some emotions. In this experiment, 2,100 emotive expressions used in existing emotion analysis system, ML-Ask [3], are based. These expressions are not necessarily covering all of existing emotive expressions, and hence we calculate a words' similarity (Cosine similarity) by Word2Vec, then if a word is similar to an emotive expression in ML-Ask, we regard the word as an emotive expression too. The process is as follow. In this experiment, the objective is to decide this similarity score of θ.

1. Consider emotive expressions in ML-Ask as emotive expression list.
2. Split given a sentence to words by morphological analyzer MeCab [4] and get a set of appeared words.
3. Calculate the similarities of between the words in emotive expression list and appeared words, then a max similarity will be the score of the given sentence.
4. If the score is greater than θ, the sentence is regarded as it contains emotive expressions.

We made 26 annotators to annotate emotions of tweets (Yamashita et al., 2019), if over 25% of annotators get lost to annotate, then the sentence is regarded as a difficult data to estimate emotions, else is regarded as an easy data.

4.2 Examine the Similarity

About each of difficult data and easy data, a rate of data including emotive expressions is shown in Figure 16. Emotive expressions are the high rate in difficult data, not in easy data. An example of difficult data having a max score (1.0) is　えーこれは自画

[3]http://arakilab.media.eng.hokudai.ac.jp/~ptaszynski/repository/mlask.htm

[4]http://taku910.github.io/mecab/

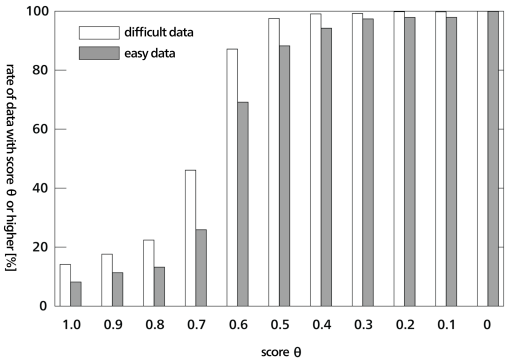

Figure 6: The rate of data including emotive expressions.

自賛しても怒られない、公開待ちなのが惜しい (*well, I cannot get angry if I praise myself, it's regrettable to be open to the public*). This example has a fit of emotive expression *anger*, thus this sentence can be regarded to be easy data. But this includes a negative expression, so this sentence can be regarded to be difficult data, too.

An example of Over 0.7 score sentence is るっびーかいぎ、試験と丸被りの予感がするし今年は諦めかなぁ (*RubyKaigi, I have a bad feeling that it may conflict with the exam and I'll give up this year*), we regard these sentences as including emotive expressions. In practical use, deleting most of the difficult data by some way is the best to detecting emotive expressions.

5 DECIDE THE DIFFICULTY BY CLASSIFIERS

In this experiment, we aim to decide the difficulty of emotion estimation. Section 5.1 describes the method of this experiment, and Section 5.2 describes the results of this experiment. In section 5.3, we compare the results with the baseline.

5.1 Classifiers

Difficulty deciding by similarity without some classifiers to be a baseline. As the classifiers, we make SVM, CNN, and LSTM that these features are vectors of Word2Vec. Preparing features excluding some POS from 11 list of POS: 名詞 (noun), 助詞 (postpositional particle = PP), 動詞 , (verb), 助動詞 (auxiliary verb = AV), 記号 (symbol), 形容詞 (adjective), 副詞 (adverb), 感動詞 (interjection), 連体詞 (pre-noun adjectival = PA), フィラー (filler), 接続詞 (conjunction), we find useful POS to classify experimentally.

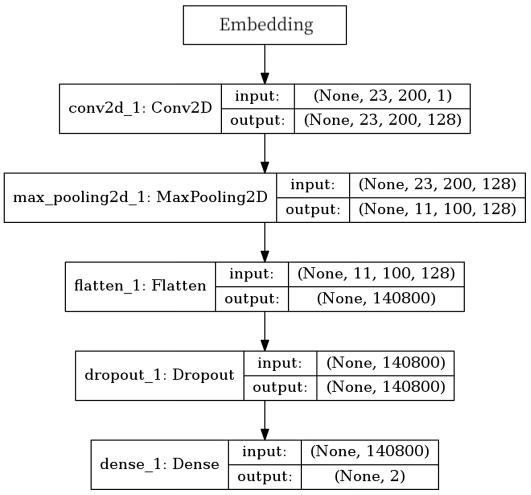

Figure 7: The CNN model.

For training and evaluating, we use the same data as section 3. In the whole 998 data, difficult data is 635, and easy data is 373.

5.1.1 The Baseline

Calculating a similarity between given data and difficult data, if the similarity is over the specific value, this given data is regarded to be difficult data. Word Mover's Distance is used to calculating a similarity.

5.1.2 SVM with Word2Vec Features

Split the train data to word lists with MeCab, and get 200 dimensions Word2Vec vectors. The sum of word vectors composing train data is regarded as a sentence vector, and this vector to be the feature of SVM. Then 10-fold cross-validate and evaluate the model.

5.1.3 CNN with Word2Vec Features

An example of an implementation to classify texts by CNN is Kim's model (Kim, 2014). In our method, we refer the model Kim proposed and make a model up in Figure 7. At the embedding layer, same as SVM, make the sentence vectors having 200 dimensions each word. At the convolution layer, make 128 filters size of 3x200, and convolute each other. And then pooling each filter at pooling layer, output 128 neurons. At last, fully connect and output class probabilities by Softmax. Dropout is selected as the best score from 0.1 steps in range 0.0∼1.0.

5.1.4 LSTM with Word2Vec Features

The network is built like Figure 8. At the embedding layer, create sentence vectors the same as SVM and

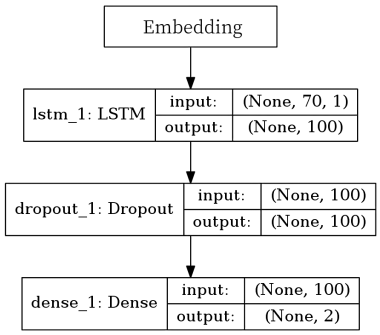

Figure 8: The LSTM model.

CNN. Next, provide an LSTM layer with 100 hidden layers. Fully connect onward are the same as CNN.

5.2 The Results of this Experiment

Compare the performance of the baseline and three classifiers. As a performance indicator F_β value ($\beta = 0.4$) was is used, because we look upon a proportion of actual easy data in data decided to be easy by the system as important. Confusion matrixes are displayed as order [TP FN FP TN]. TP shows difficult data predicted as difficult, FN shows difficult data predicted as easy, FP shows easy data predicted as difficult, TN shows easy data predicted as easy.

5.2.1 The Baseline

The similarities in max F_β scores and other score are shown in Table 3. Most of the sentences are predicted as difficult data.

5.2.2 SVM with Word2Vec Features

The scores of SVM trained by features excluded specify POS is shown in Table 4. Comparing this model with the model trained by all POS features, in particular, the feature excluding postpositional particle, adjective, and adverb gives F_β value upper. Though, these 3 POSes are not looked upon useful for features to decision difficulty of emotion estimation. The case of training by excluding postpositional particle, the average number of characters per a given data is 39.7 characters, and the median is 33.0 characters. On the difficult data predicted as easy, the average is 44.2, the median is 21.0, and long sentences intended to advertise often exist. On the easy data predicted as easy, over 40% includes URLs of images or websites, but the rate that other data includes URLs is limited to be less than 10%.

Table 3: The scores of the baseline.

Recall	Prec.	F_β	Matrix
0.901	0.519	0.552	(336 37 311 62)

Table 4: The scores of SVM.

Excluded	Recall	Prec.	F_β	Matrix
–	0.545	0.552	0.551	(48 40 39 55)
Noun	0.482	0.562	0.549	(41 44 32 51)
PP	0.636	0.644	**0.643**	(56 32 31 63)
Verb	0.568	0.593	**0.590**	(54 41 37 50)
AV	0.620	0.516	0.528	(49 30 46 57)
Sym.	0.516	0.605	**0.591**	(49 46 32 54)
Adj.	0.593	0.622	**0.618**	(51 35 31 65)
Adv.	0.470	0.644	**0.613**	(47 53 26 56)
Int.	0.535	0.590	**0.582**	(46 40 32 63)
PA	0.609	0.582	**0.586**	(53 34 38 57)
Filler	0.525	0.602	0.**590**	(53 48 35 46)
Conj.	0.556	0.610	**0.602**	(50 40 32 60)
PP, Adj.	0.633	0.538	0.549	(57 33 49 43)
PP, Adv.	0.550	0.647	**0.632**	(55 45 30 52)
Adj., Adv.	0.711	0.602	**0.615**	(59 24 39 60)

Table 5: The scoers of CNN.

Excluded	Recall	Prec.	F_β	Matrix
–	0.500	0.676	0.645	(46 46 22 68)
Noun	0.735	0.581	0.598	(61 22 44 41)
PP	0.667	0.667	**0.667**	(66 33 33 50)
Verb	0.535	0.662	0.641	(53 46 27 56)
AV	0.740	0.640	**0.652**	(71 25 40 46)
Sym.	0.60	0.698	**0.682**	(60 40 26 55)
Adj.	0.772	0.617	0.635	(71 21 44 46)
Adv.	0.710	0.660	**0.666**	(66 27 34 55)
Int.	0.787	0.914	**0.894**	(74 20 7 80)
PA	0.729	0.642	**0.653**	(70 26 39 47)
Filler	0.605	0.571	0.576	(52 34 39 57)
Conj.	0.823	0.612	0.635	(79 17 50 36)

5.2.3 CNN with Word2Vec Features

On the case of CNN, the scores of a model trained by features excluding specify POS is shown in Table 5. As a result of excluding interjection, the easy data predicted as difficult are not including any parentheses, brackets, or braces. And same as SVM, difficult data predicted as easy often include URL comparatively.

5.2.4 LSTM with Word2Vec Features

On the case of LSTM, the scores of a model trained by features excluding specify POS is shown in Table 6. As a result of excluding verb and auxiliary verb, the average number of characters per a given data is 40.4 characters, the median is 32.5 characters. On the difficult data predicted as easy, the average is 32.1, the

Table 6: The scores of LSTM.

Excluded	Recall	Prec.	F_β	Matrix
–	0.788	0.670	0.684	(67 18 33 64)
Noun	0.642	0.612	0.616	(52 29 33 54)
PP	0.753	0.615	0.631	(64 21 40 57)
Verb	0.609	0.709	**0.693**	(56 36 23 67)
AV	0.545	0.714	**0.685**	(55 46 22 59)
Sym.	0.681	0.627	0.634	(64 30 38 49)
Adj.	0.698	0.615	0.625	(67 29 42 44)
Adv.	0.529	0.672	0.648	(45 40 22 75)
Int.	0.591	0.658	0.648	(52 36 27 66)
PA	0.653	0.711	**0.702**	(64 34 26 58)
Filler	0.625	0.640	0.637	(55 33 31 63)
Conj.	0.624	0.624	0.624	(53 32 32 65)
Verb, AV	0.717	0.724	0.723	(71 28 17 56)

median is 17.5, and the median tends to be lower similar to SVM. Also same as SVM or CNN, the difficult data predicted as easy are often including URLs.

Auxiliary verbs are often used as previous, completion, or affirmation meanings: 月がわりときれいですね (*the moon is beautiful so-so*). This POS is not useful to estimate emotions because it is hard to associate some emotions with these auxiliary verbs. However, sentences including like ない (*no*), one of the negative expressions, some auxiliary verb, are regarded to be high difficulty to decision them emotion estimation so it is not always correct to exclude all auxiliary verb.

5.3 Compare Methods with the Baseline

Extract the results shown in section 5.2 and compare in Table 7. From F_β values, for deciding the difficulty of emotion estimation by classifiers, CNN trained by features excluding interjection is adopted. The baseline by words similarity decisions almost data to be difficult. It is expected to be easy to decide as difficult when using word similarities. On the other hand, the case based on word distributed representation, easy and difficult data are correctly decided over 60%, the decision has not a bias. Especially CNN, the proportion that easy data are correctly predicted to be easy is over 90%.

Table 7: Compare the methods with the baseline.

Method (Excluded POS)	Recall	Prec.	F_β
Baseline	0.901	0.519	0.552
SVM(PP)	0.636	0.644	0.643
CNN(Int.)	0.787	0.914	**0.894**
LSTM(Verb, AV)	0.717	0.724	0.723

6 BUILDING A DIFFICULTY DECISION SYSTEM

In section 4 and 5, we are can decide the difficulty of emotion estimation from the existence of emotive expressions and classifiers. In this section, build a deciding the difficulty of emotion estimation system to combine these deciding methods. In section 6.1 describes the construction of this system and section 6.2 describes the evaluation of the system.

6.1 The Construction of the System

This system receives Japanese sentences and then returns difficulties of emotion estimation "high difficulty" or "low difficulty" for each sentence. Inside the system, which decisions by a combination of 3 conditions: (1) existence of negative expressions, (2) existence of emotive expressions, (3) prediction by classifiers. The decision of the existence of negative expressions, Naive Bayes is used (Yamashita et al., 2019). The sentence including some negative expressions is considered to be "high difficulty", so if the sentence is decided that it includes some negative expressions by Naive Bayes, the decision of the sentence becomes high. Including emotive expressions or not, is suggested in chapter 4, is decided by the words similarity score (over 0.7 or not). The sentence including emotive expressions are regarded to be easy to decide the difficulty, so the sentence predicted including emotive expressions becomes easy. In the case of prediction by classifiers, decide the difficulty by classifiers suggested in section 4.

6.2 Evaluation of the System

To evaluate the system, use the annotation data that 8 people who know the writer (author) annotated 254 author's tweets. This data is not included in the data used on each above experiments. Same as section 4, separate this data into the difficult data and the easy data.

The evaluations are shown in Table 8. Deciding by 2 steps, the existence of negative expressions and classifiers is the best score. 70% of the difficult data are correctly predicted, but the easy data could not be predicted correctly 20%. In the case of the decision including emotive expressions, one of the features of the FN (= False Negative) data which is actually difficult but predicted easy is including したい (*want to do*). This expression shows the writer's hope or request, but usually, it is not written that what kind of emotion the writer can give to do it really, so the emotive expressions are hard to be detected. In FP (=

Table 8: The evaluation of the system.

Combination	Acc.	Recall	Prec.	F_1	F_β	Matrix
Negative	0.268	0.241	0.951	0.384	0.676	(58 183 3 10)
Emotive	0.575	0.593	0.935	0.726	0.866	(143 98 10 3)
Classify	0.646	0.660	0.952	0.779	0.897	(159 82 8 5)
Negative + Emotive	0.689	0.718	0.940	0.814	0.902	(173 68 11 2)
Negative + Classify	**0.740**	**0.768**	0.949	**0.849**	**0.919**	(185 56 10 3)
Emotive + Negative	0.154	0.116	0.933	0.207	0.474	(28 213 2 11)
Emotive + Classify	0.413	0.411	0.934	0.571	0.794	(99 142 7 6)
Classify + Negative	0.173	0.133	0.970	0.234	0.519	(32 209 1 12)
Classify + Emotive	0.413	0.411	0.934	0.571	0.794	(99 142 7 6)
Negative + Emotive + Classify	0.579	0.598	0.935	0.729	0.867	(144 97 10 3)
Negative + Classify + Emotive	0.579	0.598	0.935	0.729	0.867	(144 97 10 3)
Emotive + Negative + Classify	0.465	0.473	0.927	0.626	0.819	(114 127 9 4)
Emotive + Classify + Negative	0.102	0.054	**1.000**	0.102	0.292	(13 228 0 13)
Classify + Negative + Emotive	0.484	0.490	0.937	0.643	0.832	(118 123 8 5)
Classify + Emotive + Negative	0.102	0.054	**1.000**	0.102	0.292	(13 228 0 13)

False Positive) data which is easy data predicted as hard, there are some data including words not being regarded as emotive expressions. 神 (*god*) and たのしい (*enjoyable*) are samples of these words. For example, 神 is used to express such as joy and trust. In the decide system constructed in this research, high similarity words to 神 are 恩寵 (*grace*) and 慈悲 (*mercy*), but these words are not regarded to be emotive expressions because similarities are less than 0.7. Although たのしい (written in hiragana – the Japanese cursive syllabary) is the same meaning with 楽しい (written in kanji – the Chinese ideographs), the similarity of these 2 words is 0.51 counterintuitively.

7 CONCLUSION

Even though it cannot compare the scores because this theme has no existing research, 70% of high difficulty data are decided correctly. On the other hand, 80% of easy data are decided incorrectly. Following two sentences are proved to be hard to decide the difficulty: (1) Including emotive expressions but failed to detect. (2) Not including any emotions.

In the future, we aim to improve the score of the system by considering the annotation times and the features of miss-decided sentences. Also, we try to improve versatility by using not only the author's tweets but also sentences written by other people.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 18K11455.

REFERENCES

Hasegawa, T., Kaji, N., Yoshinaga, N., and Toyoda, M. (2014). Predicting and evoking listener's emotion in online dialogue. *Transactions of the Japanese Society for Artificial Intelligence : AI*, 29(1):90–99.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746—-1751.

Matsubayashi, K. et al. (2016). An emotion estimation method from twitter tweets and the application. *Proceedings of the 78th National Convention of IPSJ*, 2016(1):79–80.

Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K., and Masui, F. (2017). Ml-ask: Open source affect analysis software for textual input in japanese. *Journal of Open Research Software*, 5(1).

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. (2018). CARER: Contextualized affect representations for emotion recognition. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697.

White, M. M. and Powell, M. (1936). The differential reaction-time for pleasant and unpleasant words. *The American Journal of Psychology*, 48(1):126–133.

Yamashita, S., Kami, Y., Kato, E., Sakai, T., and Okumura, N. (2018). An evaluation method for estimating the degree of difficulty to extract writer's emotion based on response time in annotating emotion. *Technical Report of IEICE Document Communication*, 1(1):1–6.

Yamashita, S., Kami, Y., and Okumura, N. (2019). 品詞情報とルールベースによる否定表現有無の判定 (in japanese). *Proceedings of the 25th annual meeting of the Association for Natural Language Processing*, 1(1):1447–1450.