Text Normalization for Sentiment Analysis in Japanese Social Media

Risa Kondo[†] **Ayu Teramen**[†] **Reon Kajikawa**[†] **Koki Horiguchi**[†] **Tomoyuki Kajiwara**[†‡]
Takashi Ninomiya[†] **Hideaki Hayashi**[‡] **Yuta Nakashima**[‡] **Hajime Nagahara**[‡]
[†]Ehime University [‡]Osaka University
{kondo@ai., teramen@ai., reon@ai., horiguchi@ai., kajiwara@}cs.ehime-u.ac.jp
ninomiya.takashi.mk@ehime-u.ac.jp
{hayashi, n-yuta, nagahara}@ids.osaka-u.ac.jp

Abstract

We manually normalize noisy Japanese expressions on social networking services (SNS) to improve the performance of sentiment polarity classification. Despite advances in pre-trained language models, informal expressions found in social media still plague natural language processing. In this study, we analyzed 6,000 posts from a sentiment analysis corpus for Japanese SNS text, and constructed a text normalization taxonomy consisting of 33 types of editing operations. Text normalization according to our taxonomy significantly improved the performance of BERT-based sentiment analysis in Japanese. Detailed analysis reveals that most types of editing operations each contribute to improve the performance of sentiment analysis.

1 Introduction

For research and development of sentiment analysis models, datasets with sentiment labels for text on social networking services (SNS) are available (Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018; Plaza del Arco et al., 2020; Bostan et al., 2020). In Japanese, sentiment analysis datasets for SNS posts such as WRIME[1] (Kajiwara et al., 2021; Suzuki et al., 2022) are available. Text from social media often contains informal Japanese expressions such as misspellings and Internet slang. These noisy texts may degrade the performance of natural language processing, including sentiment analysis.

In this study, to improve the performance of sentiment analysis in Japanese, various noisy expressions in SNS texts were manually normalized. We performed text normalization on 6,000 posts from the WRIME dataset, and organized the editing operations contained therein into 6 major categories and 33 subcategories. Then, our detailed analysis based on this Japanese text normalization taxonomy revealed which type of normalization contributes to improved performance of sentiment analysis in Japanese.

Experimental results showed that our text normalization improved the performance of sentiment analysis in Japanese. Furthermore, our detailed analysis reveals that most types of normalization contribute to improved performance in sentiment analysis. Among them, there were notable improvements due to the normalization of *casual/formal sentence endings*, *missing symbols*, *abbreviations*, and inconsistencies in *hiragana, katakana, and kanji*.[2] In contrast, since the normalization of *pronunciation variations* worsened the performance of sentiment analysis, the *pronunciation variations* may express the writer's emotions. We plan to release[1] our 6,000 normalized post pairs with our Japanese text normalization taxonomy.

2 Related Work

Noisy expressions found in social media deteriorate the performance of various natural language processing such as word segmentation and sentiment analysis. To address this issue, text normalization has been studied. Text normalization corpora have been developed for various languages, including English (Liu et al., 2011; Han and Baldwin, 2011; Yang and Eisenstein, 2013; Baldwin et al., 2015), German (Sidarenka et al., 2013), Spanish (Alegria et al., 2013, 2015), Turkish (Çolakoğlu et al., 2019), Danish (Plank et al., 2020), Italian (van der Goot et al., 2020), Thai (Limkonchotiwat et al., 2021), and Vietnamese (Nguyen et al., 2024), to facilitate the development of data-driven approaches for text normalization. For text normalization in Japanese, approaches to sequence labeling (Sasaki et al., 2013; Osaki et al., 2017) and sequence-to-sequence generation (Ikeda et al., 2016; Saito et al.,

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

[2] Japanese text can be written in three types of letters: hiragana, katakana, and kanji.

Proceedings of the Tenth Workshop on Noisy and User-generated Text, pages 149–157

2017) have been proposed. However, these previous studies are based on small parallel corpora of about 1,000 sentence pairs (Sasaki et al., 2013; Kaji and Kitsuregawa, 2014; Osaki et al., 2017; Higashiyama et al., 2021), automatically generated corpus (Ikeda et al., 2016), and non-public corpora (Saito et al., 2013, 2017). Therefore, a larger-scale parallel corpus that is freely available for Japanese text normalization is desired.

3 Japanese Text Normalization for Sentiment Analysis in Social Media

This section describes what types of text normalization are covered in this study and how we perform text normalization.

3.1 Japanese Text Normalization Taxonomy

Combining the 14 types of Japanese text normalization employed in previous studies (Saito et al., 2013; Sasano et al., 2013; Osaki et al., 2017; Higashiyama et al., 2021) and the 19 new types of normalization that we found by analyzing Japanese SNS texts in WRIME, we define a Japanese text normalization taxonomy consisting of 6 major categories and 33 subcategories.[3] Table 1 lists the taxonomy and its examples.

Typos and Misspellings As in the previous study (Saito et al., 2013), we define misspellings as separate subcategories of *misuse* of kanji and *typos*. We also employ the *missing characters* that have been employed in the previous study (Osaki et al., 2017). In addition, since *conjugation errors* were frequently observed, this is newly added as an independent category.

Even minor changes such as the presence or absence of punctuation can affect the performance of sentiment analysis. We therefore introduce a new subcategory, *missing symbols*. This type of normalization not only completes punctuation but also encloses proper nouns in parentheses.

Dialect In addition to characteristic expressions such as *Internet slang* and *censored words*, SNS texts frequently contain expressions that reflect the writer's personality, such as *regional dialect* and *unique sentence endings* that are rarely seen in normal written language. As in previous studies (Saito et al., 2013; Osaki et al., 2017), we employ these types of normalization.

Also, casual and formal forms are made consistent. Because of the frequency of each, this study divides the subcategories according to whether the editing point is sentence-ending or not, thus providing subcategories for *casual/formal sentence endings* and *casual/formal functional expressions*.

Alternative Spellings Alternative spellings, which have been employed in previous studies (Saito et al., 2013; Osaki et al., 2017; Higashiyama et al., 2021), are often found on SNS text. Since abbreviations are often used due to character count constraints, we use the category of *abbreviations* independently of changes in character types: *hiragana, katakana, and kanji*.

Along with *pronunciation variations*, *homophones*, and *small/large characters* employed in many previous studies (Saito et al., 2013; Sasano et al., 2013; Osaki et al., 2017; Higashiyama et al., 2021), we also employ *synonyms* (Sasano et al., 2013) and *loanwords* (Higashiyama et al., 2021). Considering compatibility with pre-trained language models, *synonyms* are paraphrased into the most frequent expressions, and *loanwords* are translated or transliterated into hiragana or kanji.

There was also variation in the use of parentheses and other symbols. Therefore, we also add the category of *symbol conversion*.

Emphasis Expressions *Inserted sounds*, *inserted symbols*, and *repetition* of characters and symbols, which have been employed in the previous study (Osaki et al., 2017), are also frequently used in social media for the purpose of emphasis. To eliminate redundancy and to make these expressions consistent across the corpus, they are also normalized in this study.

Some posts list parallel items with *bullet points* or *word order changes* to uncommon or unreadable sentences. We newly normalize and edit them into fluent and complete sentences.

Simplification As a new major category, we introduce a new category of "simplification" to paraphrase complex expressions or to complement missing information. We employ five types of subcategories: *lexical/phrasal simplification* to paraphrase complex expressions and SNS-specific expressions such as neologisms and coined words, *completion*

[3] The "similar forms" employed by previous studies (Saito et al., 2013; Sasano et al., 2013) were not employed in this study because they did not appear in our analysis. For example, this category includes ネ申 → 神, うれ∪い → うれしい, etc. Our analysis covers 6,000 posts from the WRIME dataset, which consists of SNS texts posted from 2010 to 2020.

1. Typos and Misspellings	Example
Missing Symbols	暑い→ 暑い。, 天地明察を見たい→ 『天地明察』を見たい
Missing Characters‡	みんな起きている→ みんなが起きている, とこ→ ところ
Conjugation Errors	見てたら→ 見ていたら, 起きれて→ 起きられて
Typos*‡§	腸がが→ 腸が, きます！れ→ きます!!
Misuse*	以外に少ない→ 意外に少ない
2. Dialect	Example
Casual/Formal Sentence Endings	～だ。→ ～です。, 食いたい。→ 食べたいです。
Casual/Formal Functional Expressions	っていう話→ という話, 奪われるから→ 奪われるので
Internet Slang*‡	ワロタでした→ 笑いました , ググったら→ 検索すると
Regional Dialect*‡	やん→ でしょうね, おめんど→ あなたたち
Unique Sentence Endings	ますわよ→ ますよ, っす→ です
Censored Words*	N__K → NHK
3. Alternative Spellings	Example
Hiragana/Katakana/Kanji*‡§	欲しい→ ほしい, スカート＋ヒール→ スカートとヒール
Abbreviations*‡	ネット→ インターネット, コロナ→ 新型コロナウイルス感染症
Pronunciation Variations*†‡	いくん→ いくの, こりゃ→ これは
Synonyms	本日→ 今日, お菓子→ 菓子
Symbol Conversion	「悪の教典」→ 『悪の教典』, 。。。→ …。
Loanwords§	good night → おやすみなさい, オーダー→ 注文
Homophones*†‡	行けそーな→ 行けそうな, ℃ → 度
Small/Large Characters*†‡§	まぁまぁ→ まあまあ, ヮィャレス→ ワイヤレス
4. Emphasis Expressions	Example
Inserted Sounds*†‡§	よーし→ よし, 雨かあ→ 雨か
Inserted Symbols‡	"一般的な人" → 一般的な人
Word Order Changes	そのまま私が食べるパンを→ 私が食べるパンをそのまま
Repetition‡	え ??????? → え ??, いやいやいやいやいや→ いやいや
Bullet Points	結論: → 結論として言えるのは、
5. Simplification	Example
Completion	撮ればよかったな→ 撮ればよかったなと後悔しています
Lexical/Phrasal Simplification	カットに行く→ 美容院に行く, ノミの心臓→ 臆病
Deletion	男（ひと）→ 男, せいで(おかげで) → おかげで
Fusion	今朝方のツイート。酔っていた→ 今朝方のツイートは酔っていた
Splitting	買い物に行き、買った服を→ 買い物に行きました。買った服を
6. Emotional Expressions	Example
Numerical Expressions	21時→ <num>時, ひとつ→ <num>つ, 数回→ <num>回
Emotional Symbols	(笑) → <joy>, (怒) → <anger>
Emoticons	(●´３｀●) → <joy>, orz → <sadness>
Emojis	☆→ <joy>, 🎶 → <joy><joy>

Table 1: Japanese text normalization taxonomy as defined in this study and examples for each subcategory. The symbols in the subcategory represent the type of normalization employed in previous studies, where ∗ is (Saito et al., 2013), † is (Sasano et al., 2013), ‡ is (Osaki et al., 2017), and § is (Higashiyama et al., 2021), respectively.

of missing information, *deletion* of redundant information, *splitting* and *fusion* of sentences to improve readability across sentences.

Emotional Expressions In SNS text, emoticons and emojis are frequently used to express the writer's emotions. While these can be valuable cues for sentiment analysis, there are diverse expressions, for example, "（笑）" and "www" to express feelings of joy. Therefore, to effectively utilize these for sentiment analysis, a new major category of "emotional expressions" is defined. This type of normalization groups *emoticons*, *emojis*, and *emotional symbols* such as "（笑）" into Plutchik's basic eight emotions (Plutchik, 1980)

and replaces them with special tokens such as <joy> and <sudness> that are assigned to each emotion. In addition, *numerical expressions* are also replaced with the special token <num>.

3.2 Details of Our Text Normalization

This section provides details on text normalization methods for each major category. Note that, as shown in Figure 1, multiple parts of a post may be normalized at the same time, and that multiple types of normalization may be applied to one expression.

Typos and Misspellings All errors are revised to the correct wording. In addition, missing punctuation should be completed, and proper nouns, including the titles of books and movies, should be consistently enclosed in parentheses with 『』.

Dialect Styles of sentence endings and functional expressions consistently transfer from casual to formal. Other types of dialects are normalized while using web searches as much as the annotator can detect.

Alternative Spellings Pronunciation variations, homophones, and small/large characters are revised to the correct wording. For symbol conversion, a sequence of punctuations is replaced by an ellipsis, and a comma at the end of a sentence is replaced by a period. Here, parentheses are consistently used with a single 「」 for utterances and a double 『』 for proper nouns.

Loanwords written in alphabetic or katakana characters are replaced with their Japanese counterparts when fluency can be improved by translation or transliteration. In addition, hiragana/katakana/kanji, abbreviations, and synonyms are replaced with high-frequency words. Here, word frequencies are counted from the Japanese edition of the CC-100[4] (Wenzek et al., 2020), a large-scale Web corpus, by word segmentation[5] (Kudo et al., 2004) of the text. Note that we therefore do not replace high-frequency abbreviations. For example, common abbreviations, such as "TV", are left as abbreviations because they are more frequent than the formal name of "television". However, proper nouns are not abbreviated regardless of their frequency.

[4] https://data.statmt.org/cc-100/

[5] https://github.com/neologd/mecab-ipadic-neologd

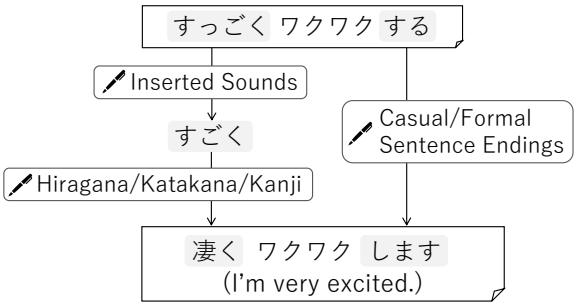

Figure 1: Example of our text normalization.

Emphasis Expressions Repetition of symbols, characters, words, or phrases should be limited to two times following the previous study (Osaki et al., 2017). Redundant sounds and symbols are also removed. Bullet points are expanded into sentences and word order is reformatted to improve fluency.

Simplification To improve readability, long compound sentences that should be expressed in multiple sentences are split, while short multiple sentences that should be expressed in one sentence are fused. Missing information should be completed if it can be inferred by the annotator, while redundant information should be deleted for simplicity. We paraphrase technical terms, low-frequency words, onomatopoeia, and other difficult-to-understand expressions into general and objective expressions.

Emotional Expressions Emojis, emoticons, and other emotional symbols are replaced with following special tokens according to Plutchik's basic eight emotions (Plutchik, 1980): <anger>, <disgust>, <fear>, <joy>, <sadness>, <surprise>, <trust>, and <anticipation>. Annotators choose which of the special tokens to replace the emotional symbols with, based on the context. We replace all numbers with the special token <num>, without regard to how large or small the numerical expressions are. However, we do not edit numerical expressions that are part of idioms, because replacing them would change their meaning.

4 Experiment

Our experiments evaluate the performance of sentiment polarity classification on sentences with individual or all normalizations, and assess the effectiveness of preprocessing with text normalization.

4.1 Settings

Task We evaluate the performance of Japanese sentiment polarity classification on the WRIME dataset (Kajiwara et al., 2021; Suzuki et al., 2022). This is a dataset of Japanese SNS posts labeled with five levels of sentiment polarity (-2, -1, 0, 1, 2) by the text writer. We used quadratic weighted kappa (QWK) (Cohen, 1968) as our evaluation metric.

Annotation For this experiment, we manually performed the text normalization described in the previous section on a total of 6,000 posts from WRIME, consisting of 5,000 posts from the training set and 500 posts each from the validation and evaluation sets. Annotations of text normalization were performed by three of the authors. First, one of the authors performed text normalization on the original posts. Then, another one of the authors evaluated the acceptability of their normalization and modified them as necessary. Finally, the remaining one author categorized each text normalization example based on our taxonomy.

Model Our sentiment analysis models were built by fine-tuning pre-trained Japanese BERT (Devlin et al., 2019) on the training set described above. For fine-tuning, AdamW (Loshchilov and Hutter, 2019) was used for optimization, the batch size was set to 64, and training was terminated when the QWK in the validation set stopped improving by 3 epochs. The learning rate was chosen from $\{1, 2, 3, 4, 5\} \times 10^{-5}$ to achieve the highest QWK in the validation set. We used two types of BERT, a base[6] model and a large[7] model, and added nine types of special tokens to the vocabulary for *emotional* and *numerical expressions*. In the following sections, we report the average score of 5 experiments conducted while changing the random seed.

4.2 Result

Table 2 shows the experimental results. The "Manual" columns that we trained and evaluated using our normalized dataset perform better in sentiment analysis than the "Baseline" columns that we trained and evaluated using the dataset without normalization. The performance improvement in sentiment analysis by text normalization is consistent for the two types of BERT models. These experimental results show that the text normalization based on our taxonomy is effective for sentiment analysis in Japanese.

	Baseline	Manual	Automatic
BERT-base	0.506	0.582	0.517
BERT-large	0.511	0.589	0.522

Table 2: Evaluation of sentiment polarity classification by quadratic weighted kappa. "Baseline" is the performance for text without normalization, "Manual" is for manually normalized text, and "Automatic" is for automatically normalized text, respectively.

4.3 Analysis: Evaluation by Subcategory

To clarify which type of text normalization contributes to improved performance in sentiment analysis, Table 3 shows the results of training and evaluating the BERT-large model with datasets normalized to each subcategory exclusively. The other experimental settings are the same as in Section 4.1.

For most subcategories, our text normalization improved the performance of sentiment analysis. The worse performance of sentiment analysis when only *pronunciation variations* were normalized suggests that changes in pronunciation are more likely to express the writer's emotions.

Text normalization for the four categories of *casual/formal sentence endings*, *missing symbols*, *hiragana/katakana/kanji* variations, and *abbreviations* achieved significant performance improvements of more than 3 points each. The diversity of texts, including these spelling inconsistencies, is a factor that hinders the training of sentiment analysis models.

4.4 Analysis: Automatic Text Normalization

We tried automatic text normalization by fine-tuning BART[8] (Lewis et al., 2020), a pre-trained sequence-to-sequence model, using our text normalization dataset. In fine-tuning, we applied vocabulary expansion as in BERT in Section 4.1, used AdamW (Loshchilov and Hutter, 2019) for optimization, set the batch size to 8, and terminated training when the cross-entropy loss on the validation set stopped improving by 3 epochs.

The performance of text normalization was evaluated by BLEU (Papineni et al., 2002) on the evaluation set, and the results showed a significant improvement from BLEU=47.4 without normaliza-

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

[7] https://huggingface.co/tohoku-nlp/bert-large-japanese

[8] https://huggingface.co/ku-nlp/bart-large-japanese

Category	Subcategory	#	QWK
	Baseline (w/o normalization)		0.511
	Apply all types of normalization		**0.589**
Typos and Misspellings	Missing Symbols	4,453	**0.555**
	Missing Characters	3,604	**0.529**
	Conjugation Errors	1,328	**0.526**
	Typos	55	**0.538**
	Misuse	45	**0.513**
Dialect	Casual/Formal Sentence Endings	5,321	**0.559**
	Casual/Formal Functional Expressions	1,923	0.511
	Internet Slang	539	**0.515**
	Regional Dialect	319	**0.522**
	Unique Sentence Endings	128	**0.523**
	Censored Words	16	**0.527**
Alternative Spellings	Hiragana/Katakana/Kanji	2,480	**0.550**
	Abbreviations	1,262	**0.541**
	Pronunciation Variations	1,031	0.509
	Synonyms	886	**0.525**
	Symbol Conversion	461	**0.537**
	Loanwords	273	**0.518**
	Homophones	132	**0.532**
	Small/Large Characters	63	**0.514**
Emphasis Expressions	Inserted Sounds	963	**0.518**
	Inserted Symbols	331	**0.538**
	Word Order Changes	293	**0.538**
	Repetition	288	**0.525**
	Bullet Points	43	**0.525**
Simplification	Completion	918	**0.520**
	Lexical/Phrasal Simplification	771	**0.530**
	Deletion	220	**0.518**
	Fusion	105	**0.517**
	Splitting	38	**0.531**
Emotional Expressions	Numerical Expressions	968	**0.533**
	Emotional Symbols	259	**0.535**
	Emoticons	180	**0.515**
	Emojis	47	**0.521**

Table 3: Performance of sentiment polarity classification by BERT-large evaluated with quadratic weighted kappa (QWK) when only the subcategories in each row are normalized. If that normalization improves performance over the baseline, the values in the QWK column are highlighted in bold. The # column shows the number of normalizations that fall into each subcategory out of the 6,000 posts we analyzed.

tion to BLEU=62.0, indicating the effectiveness of automatic text normalization. The "Automatic" column in Table 2 shows the performance of sentiment analysis trained and evaluated using an automatically normalized dataset. Not surprisingly, automatic text normalization did not contribute to the improved performance of sentiment analysis as much as its manual counterpart. Nevertheless, consistent performance improvements were achieved for both types of BERT models. More training data would improve the performance of automatic text normalization, but that is left as our future work.

	Text	Label
Original post	しもんぬきゃわ	Negative
Automatic normalization	仕事に行きません。	Very Negative
Manual normalization	下野紘が可愛いです。	Very Positive
Reference	Hiro Shimono is cute.	Very Positive
Original post	ふふってなった	Negative
Automatic normalization	ふふっていました。	Negative
Manual normalization	ふふっとなりました。	Neutral
Reference	It made me smile.	Positive
Original post	あたまもおなかもいたい。どっちかにしてほしい	Neutral
Automatic normalization	あたまもお腹も痛いです。どっちかにしてほしいです。	Negative
Manual normalization	頭もお腹も痛いです。 どちらかにしてほしいです。	Negative
Reference	I have a headache and a stomachache. Pick a side!	Negative
Original post	私３Ｆ3列26	Positive
Automatic normalization	私は列です	Positive
Manual normalization	私は**<num>**階の**<num>**列**<num>**番の席です。	Neutral
Reference	I am on the third floor, row 3, seat 26.	Very Negative

Table 4: Examples of text normalization and its sentiment analysis. Reference rows are the English translation of the normalized text and the correct emotional polarity label annotated by the writer who posted the original text.

4.5 Qualitative Evaluation

Table 4 shows examples of text normalization and the results of its sentiment analysis. As in these examples, sentences consisting only of hiragana characters deteriorate the performance of sentiment analysis. Conversely, sentences that do not contain hiragana characters, as in the bottom example, are also difficult. If these can be properly normalized, expressions such as "可愛い (cute)" and "痛い (ache)" appear as cues to positive or negative emotions, contributing to improved performance of sentiment analysis. In the bottom example, the numerical expression represents the negative emotion of distant, but normalization of the numerical expression has made it difficult to read that emotion. Although the normalization of numerical expressions contributes to sentiment analysis on average, it can also have a negative impact, as in this example. In some cases, automatic text normalization almost works, as in the second and third examples, but in others, as in the first example, it generates text that is off the mark.

5 Conclusion

In this study, we worked on text normalization as a preprocessing to improve the performance of sentiment analysis for Japanese SNS texts. We defined a Japanese text normalization taxonomy consisting of 33 types of editing operations and manually normalized 6,000 posts. Experimental results showed that both automatic and manual text normalization consistently improved the performance of sentiment analysis. In manual text normalization, most types of normalization improved the performance of sentiment analysis, respectively. Our detailed analysis reveals that *pronunciation variations* should not be edited, and are a useful linguistic phenomenon for sentiment analysis.

Limitations

We released a dataset of manually normalized Japanese text from 6,000 posts (about 11,000 sentences) on social media. Our corpus is larger, considering that the Japanese text normalization corpora available in previous studies are about 1,000 sentence pairs. However, it is an insufficient size compared to corpora available for other text-to-text generation tasks such as machine translation, grammatical error correction, and text simplification.

Acknowledgments

This work was supported by Innovation Platform for Society 5.0 from Japan Ministry of Education, Culture, Sports, Science and Technology (JPMXP0518071489).

References

Iñaki Alegria, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2015. TweetNorm: A Benchmark for Lexical Normalization of Spanish Tweets. *Language Resources and Evaluation*, 49(4):883–905.

Iñaki Alegria, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluís Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2013. Introducción a la Tarea Compartida Tweet-Norm 2013: Normalización Léxica de Tuits en Español. In *Proceedings of the Tweet Normalization Workshop co-located with 29th Conference of the Spanish Society for Natural Language Processing*, pages 1–9.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.

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 Twelfth Language Resources and Evaluation Conference*, pages 1554–1566.

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

Talha Çolakoğlu, Umut Sulubacak, and Ahmet Cüneyd Tantuğ. 2019. Normalizing Non-canonical Turkish Texts Using Machine Translation Approaches. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 267–272.

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.

Bo Han and Timothy Baldwin. 2011. Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378.

Shohei Higashiyama, Masao Utiyama, Taro Watanabe, and Eiichiro Sumita. 2021. User-Generated Text Corpus for Evaluating Japanese Morphological Analysis and Lexical Normalization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5532–5541.

Taishi Ikeda, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Japanese Text Normalization with Encoder-Decoder Model. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pages 129–137.

Nobuhiro Kaji and Masaru Kitsuregawa. 2014. Accurate Word Segmentation and POS Tagging for Japanese Microblogs: Corpus Annotation and Joint Modeling with Lexical Normalization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 99–109.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 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.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, and Sarana Nutanong. 2021. Handling Cross- and Out-of-Domain Samples in Thai Word Segmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1003–1016.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, Deletion, or Substitution? Normalizing Text Messages without Pre-categorization nor Supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the Seventh International Conference on Learning Representations*.

Saif Mohammad and Felipe Bravo-Marquez. 2017. 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.

Thanh-Nhi Nguyen, Thanh-Phong Le, and Kiet Nguyen. 2024. ViLexNorm: A Lexical Normalization Corpus for Vietnamese Social Media Text. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1421–1437.

Ayaha Osaki, Yoshiaki Kitagawa, and Mamoru Komachi. 2017. Nihongo Twitter bunsho wo taishou to shita keiretsu labeling ni yoru hyouki seikika. *IPSJ SIG Technical Report*, 2017-NL-231(12):1–6. (In Japanese).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish Nested Named Entities and Lexical Normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662.

Flor Miriam Plaza del Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin. 2020. EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1492–1498.

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

Itsumi Saito, Kugatsu Sadamitsu, Hisako Asano, and Yoshihiro Matsuo. 2013. Extracting Derivational Patterns based on the Alignment of a Standard Form and its Variant towards the Japanese Morphological Analysis for Noisy Text. *IPSJ SIG Technical Report*, 2013-NL-214(5):1–9. (In Japanese).

Itsumi Saito, Jun Suzuki, Kyosuke Nishida, Kugatsu Sadamitsu, Satoshi Kobashikawa, Ryo Masumura, Yuji Matsumoto, and Junji Tomita. 2017. Improving Neural Text Normalization with Data Augmentation at Character- and Morphological Levels. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 257–262.

Akira Sasaki, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui. 2013. Normalization of Text in Microblogging Based on Machine Learning. In *Proceedings of the 27th Annual Conference of the Japanese Society for Artificial Intelligence*. (In Japanese).

Ryohei Sasano, Sadao Kurohashi, and Manabu Okumura. 2013. A Simple Approach to Unknown Word Processing in Japanese Morphological Analysis. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 162–170.

Uladzimir Sidarenka, Tatjana Scheffler, and Manfred Stede. 2013. Rule-based Normalization of German Twitter Messages. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*.

Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2022. A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7022–7028.

Rob van der Goot, Alan Ramponi, Tommaso Caselli, Michele Cafagna, and Lorenzo De Mattei. 2020. Norm It! Lexical Normalization for Italian and Its Downstream Effects for Dependency Parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6272–6278.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012.

Yi Yang and Jacob Eisenstein. 2013. A Log-Linear Model for Unsupervised Text Normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 61–72.