Title	Emotion Analysis Model Using Dialect Corpus and Proposal of Flaming and Cyberbullying Detection Method
Author(s)	加藤, 大造
Citation	
Issue Date	2023-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/18753
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 修士(情報科学)

Master's Thesis

Emotion Analysis Model Using Dialect Corpus and Proposal of Flaming and Cyberbullying Detection Method

Taizo Kato

Supervisor NGUYEN, Minh Le

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September 2023

Abstract

It has been a long time since users were able to post their own feelings and thoughts freely and easily on the Internet through features such as posting to social networking services (SNS) and bulletin boards or commenting on video sharing services. With 74.2% of individuals using SNS in modern society, online conversations have become a part of our daily life. These conversations are typically spoken language and are often spoken with regional dialects that reflecting the user's place of residence or born and raised. The volume of such dialect-infused text data is on the rise, a natural language processing (NLP) models that can understand these dialects is required. In this study, we hypothesize that text containing dialects more strongly reflects the writer's emotions. We built a dialect corpus of approximately 320,000 instances gathered from dialect dictionaries and Twitter to train a variation of the BERT language model, which is named "DialectBERT". We fine-tuned this model for analyzing the intensity of eight emotions (Joy, Sadness, Anticipation, Surprise, Anger, Fear, Disgust and Trust) and their polarities. As a result, we confirmed that DialectBERT could correctly analyze six out of the eight emotions (Joy, Sadness, Anticipation, Surprise, Anger, Disgust) more accurately than existing models. DialectBERT also outperformed in terms of sentiment polarity analysis. Further, we demonstrated that comparable accuracy can be achieved with between 100,000 and 150,000 training instances.

The use of SNS becomes commonplace, problems such as online flaming and cyberbullying have become social issues. To address this, we collected conversational data from Twitter containing words related to flaming and cyberbullying, and then labelled data where these issues occurred. Using these data and the eight emotion analysis models, we analyzed the emotional intensity of each conversation and created emotional vectors. These vectors were then used to detect incidents of flaming and cyberbullying through vector similarity and machine learning algorithms. In all cases, models using DialectBERT yielded better detection accuracy. This study demonstrated that a BERT model trained with a dialect corpus can more accurately analyze emotional intensity, and that this model can effectively detect online flaming and cyberbullying incidents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

It has been a long time since users were able to freely and easily post their own feelings and thoughts on the Internet through features such as posting to social networking services (SNS) and bulletin boards or commenting on video sharing services. In Japan, the use of social networking services (SNS) increased around 2004 with the entrance of platforms like GREE[1] and mixi[2] [1]. Nearly 20 years have passed since then, and according to a survey by the Ministry of Internal Affairs and Communications in 2023, 74.2% of individuals use SNS [2]. As the use of SNS becomes a part of everyday life, conversations among friends and users often take place on these platforms. In the field of natural language processing(NLP) research, studies such as sentiment analysis [3, 4, 5] and flaming detection [6] have been conducted using conversational and emotional data posted on various SNS services.

Posts on the Internet, such as SNS, are usually made in spoken language. These posts are likely to contain many dialects from the regions where users live or born and raised. We are often felt that my posts or those of friends and acquaintances contain their dialects. Hirota et al. [7] stated, "ブログ等の CGM の普及により Web 上で方言が使用される機会が増えている.また、それに伴い,方言に対しても頑健な言語処理技術の必要性が高まっている(With the spread of CGM such as blogs, the use of dialects on the web is increasing, and the need for robust language processing technologies for dialects is growing.)". Given these characteristics of posted data, it is expected that using NLP models that understand dialects could lead to improved analysis accuracy in the analysis of Internet posted data.

As the use of SNS becomes more commonplace, problems like flaming and cyberbullying on SNS are becoming social issues. Yamaguchi [8] defined flaming as a "ある人物や企業が発信した内容や行った行為について、ソーシャルメディアに批判的なコメントが殺到する現象(Phenomenon where critical comments flood in on social media about the content disseminated or actions taken by a certain person or company.)" The Ministry of Education, Culture, Sports, Science, and Technology (MEXT) describes cyberbullying in its manual and casebook on 「ネット上のいじめ」に関する対応マニュアル・事例集（学校・教員向け） [9] as "bullying conducted through methods such as writing slander or defamation about a specific child on websites like bulletin boards on the Internet via mobile phones or computers, or sending emails." In 2020, there was an incident where a female professional wrestler who appeared on a

---

[1] https://gree.jp/

[2] https://mixi.jp/

TV program was excessively slandered on SNS due to her actions on the TV program, leading her to commit suicide. According to MEXT's 2023 survey on “令和３年度児童生徒の問題行動・不登校等生徒指導上の諸課題に関する調査結果の概要” [10], there were 21,900 cases of bullying using computers and mobile phones, and the trend is still increasing. Research on determining whether posted data on SNS is positive or negative [11] and on identifying posts leading to bullying [12] is also being conducted.

In NLP research, Devlin et al. [13] proposed the BERT model based on the Transformer by Vaswani et al. [14]. This BERT model has updated the state-of-the-art (SoTA) in many NLP tasks. BERT model conducts pre-training tasks called Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM is a task that hides multiple words in a sentence with [MASK] and predicts the hidden words. NSP is a task that is given two sentences and determines whether they are consecutive sentences. The model trained in pre-training is fine-tuned to solve individual tasks. Since the proposal of BERT, pre-trained BERT models have been made publicly available in various languages. In Japanese, a model trained using Wikipedia data by Tohoku University [15].

1.2 Objectives

This study assumes that texts that more strongly reflect the emotions of the writer include dialects. We collect text data containing dialects from SNS, develops an emotion analysis model using this data, and evaluate its accuracy. This study aims to demonstrate that more accurate sentiment analysis is possible by using such dialect data. We further pre-train the NLP model BERT additionally with dialect data, and then fine-tune using the sentiment analysis dataset (WRIME) that is released by Kajiwara et al. [3] and evaluate. Additionally, the method of detecting the occurrence of flaming and cyberbullying is evaluated using the highly accurate sentiment analysis model developed here. It demonstrates that this sentiment analysis model pre-trained with dialect data is effective in determining the occurrence of flaming and cyberbullying in a series of conversations exchanged on SNS. Furthermore, this study ensures versatility by excluding features dependent on specific platforms and verifies the effectiveness of the analysis method using only the posted texts. Figure 1 shows an overview of this study.

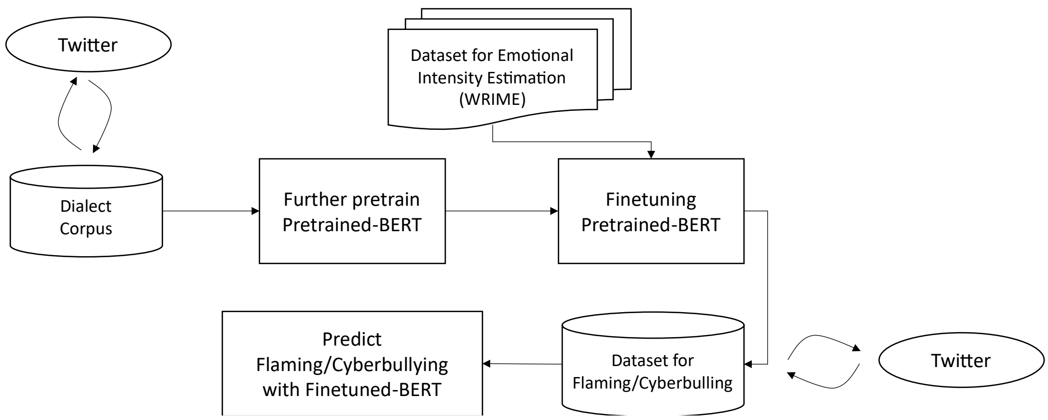

Figure 1 An overview of this study

1.3 Thesis outline

In Chapter 1, the background and objectives of this study were discussed. Chapter 2 covers related studies. Chapter 3 presents the acquisition of a sentiment analysis model using a dialect corpus and a method, as well as the experimental results. In Chapter 4, we discuss the detection of flaming and cyberbullying, detailing the data and method and the experimental results. Chapter 5 concludes this entire study.

Chapter 2
Related Works

In this chapter, related works are discussed. Section 2.1 covers studies related to dialects. Section 2.2 discusses studies on sentiment analysis, while Section 2.3 addresses studies on flaming and cyberbullying. Section 2.4 deals with studies on further pre-training, and finally, Section 2.5 discuss the challenges of the related studies.

2.1 Studies on Dialects

2.1.1. Studies on Dialects in Other Countries

Studies on dialects are actively conducted in Arabic. The reason for this is believed to be the broad usage of the language, ranging from countries on the Arabian Peninsula to Iraq, Syria, and countries on the North African continent, where different spoken languages are used in each country and region in daily conversation. Moreover, it is said that the number of speakers, including second language speakers, exceeds 400 million[3]. Mdhaffar et al. [16] studied sentiment analysis in Tunisian dialect. They collected 17,000 user comments that is used the local dialect on Facebook[4] and annotated them with positive or negative polarity. They trained the machine learning model with this dataset and then it achieved the highest accuracy. They also made this dataset publicly available as the Tunisian Sentiment Analysis Corpus (TSAC)[5]. Abdaoui et al. [17]collected about 1.2 million instances of Algerian dialect data from Twitter, trained a BERT model, and performed sentiment polarity and sentiment analysis. As a result, their BERT model (DziriBERT) achieved higher accuracy compared to other models (AraBERT [18] trained in standard Arabic, MARBERT [19] , QARiB [20] and CamelBERT [21] trained with Arabic dialect and classical Arabic data).

---

[3] http://www.flang.keio.ac.jp/plurilingualism/column010.html

[4] https://www.facebook.com

[5] https://github.com/fbougares/TSAC

2.1.2. Studies on Japanese Dialects

Kudaka [22] pointed out in the context of machine translation that "近年では、大量の対訳コーパスから翻訳モデルを学習する統計的機械翻訳（SMT: Statistical Machine Translation）やニューラル機械翻訳が主流になっている.しかし、これらの翻訳方式の翻訳精度は対訳コーパスの量に大きく依存する.したがって、低言語資源の言語をこれらの方式で機械翻訳すると、翻訳の性能が低くなることが知られている.(In recent years, statistical machine translation and neural machine translation, which learn translation models from large-scale parallel corpora, have become mainstream. However, the translation accuracy of these methods depends greatly on the volume of the parallel corpora. Therefore, it is known that when languages with limited linguistic resources are translated using these methods, the performance of the translation decreases.)" He studied a method to address this issue by automatically expanding the corpus from existing small amounts of data. Shibata et al. [23] conducted research on a bidirectional machine translation system between dialect that are spoken in Yamagata prefecture and standard Japanese. They showed that even with a parallel corpus that tolerates a certain degree of sentence error, it is possible to achieve the same level of translation accuracy as previous studies for the dialect, which has almost no language resources.

2.2 Studies on Sentiment Analysis

Kajiwara et al. [3] created and made public a dataset (WRIME[6]) for conducting subjective (one author) and objective (three readers) sentiment analysis. They employed 50 people through crowdsourcing, each of whom labeled the intensity of their past SNS posts with Plutchik's [24] eight basic emotions (Joy, Sadness, Anticipation, Surprise, Anger, Fear, Disgust and Trust) on a four-level scale (none, weak, medium, strong) subjectively. Additionally, another three people (the readers) labeled similar data objectively. In this way, they conducted a validation of the prediction accuracy of emotional intensity from both subjective and objective perspectives. The results showed that the mean absolute error was larger for the subjective data evaluation than for the objective data evaluation, indicating that predicting the emotional intensity of the writer (subjective) is

---

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

challenging. Miyauchi et al. [25] employed a similar method to Kajiwara et al. and labeled 35,000 pieces of data with emotional intensity. In addition to this, they also subjectively and objectively labeled emotional polarity on a five-level scale (strongly negative, negative, neutral, positive, strongly positive) and made it public. Suzuki et al. [4] proposed a method for better subjective emotional intensity estimation by adding personality information to Kajiwara et al.'s WRIME and demonstrated its effectiveness. Bataa et al. [26] made predictions for a five-point rating and positive-negative sentiment using Rakuten product review and Yahoo! movie review data. They showed the effectiveness of text classification in the transfer learning of a BERT model pre-trained on the Japanese Wikipedia corpus.

2.3 Studies on Flaming and Cyberbullying Detection

2.3.1. Studies on Cyberbullying Detection

Zhang et al. [12] collected approximately 2.3 million data instances from Twitter that contained 36 Japanese words related to bullying. They then selected the top 3,450 instances based on the number of bullying-related words included and manually labeled each for the presence or absence of bullying. They eventually obtained 2,790 data (1,395 each of classified as bullying and classified as not bullying). Using this data, they predicted the presence or absence of bullying using various machine learning algorithms. For feature extraction, they used an approach involving n-grams, Word2Vec, Doc2Vec, the values that calculated from an emotion dictionary known as 'emotion values of tweets', and Twitter-specific characteristics such as the number of retweets, likes, hashtags, and URL. The results showed that using n-grams achieved an accuracy, precision, recall, and F-value of over 90%. However, they pointed out that because the number of bullying-related words is limited, a method to obtain new bullying-related words is necessary.

2.3.2. Studies on Flaming Detection

Takahashi et al. [6] assigned polarity and influence values to symbols, emoticons, and degree adverbs (e.g., 'very,' 'slightly') that appear in Twitter post data (Tweets) and determined emotion labels (positive, slightly positive, neutral, slightly negative, negative) according to these polarity values. They detected as flaming those tweets for which the number of negative replies (responses to conversations) exceeded the positive ones. The detection accuracy by this method could not obtain enough results, so they then implemented detection using a decision tree with attributes assigned by Twitter's specific features, such as the number of followers of each tweet's poster and the number of favorites for that tweet. As a result, they showed that the emotions in replies to tweets and whether the replier is a follower or not are effective attributes for flaming detection.

2.4 Studies on Further Pre-training

Gururangan et al. [27] investigated whether it is useful to adjust a pre-trained model to the domain of the task to be solved. They performed a second phase of pre-training (Domain-Adaptive Pre-training) with data from the target domain and further adjusted the model using task-specific data (Task-Adaptive Pre-training). They conducted experiments on eight tasks (biomedical, computer science publications, news, reviews) across four fields. The results showed that for all tasks, the most favorable outcomes were achieved and demonstrated the effectiveness of additional pre-training.

2.5 Challenges in Related Studies

While studies related to Arabic dialects were discussed in 2.1.1, there are no known research studies in which sentiment analysis was conducted using Japanese dialect data. Similarly, there are no known studies in which a pre-trained BERT model was fine-tuned with dialect data.

In the research on flaming and cyberbullying mentioned in section 2.3, each piece of posted data is individually judged as to whether it is an aggressive post or whether it has positive or negative opinions. However, in determining flaming or cyberbullying, rather than judging each piece of posted data, it is necessary to consider the content of the entire conversation posted in a series of conversation groups and determine whether flaming or cyberbullying are occurring. Moreover, in the research by Takahashi et al. [6], a detection algorithm was developed using attributes specific to Twitter's features. However, flaming and cyberbullying are not limited to occurrences on only Twitter, but can potentially happen on any platforms where an unspecified number of people post their opinions online. Therefore, we believe that using attributes assigned by such Twitter features is insufficient. A method is sought that can make high-accuracy determinations from the posted text itself as a more universally applicable technique.

Chapter 3
The Sentiment Analysis Model Using a Dialect Corpus

3.1 Objectives

We assume that text containing dialects more strongly reflects writer's emotions. The aim is to learn from a conversation corpus containing dialects and to develop a model with high sentiment analysis accuracy. To this end, a dialect corpus is created from post data on SNS that contains dialects, and this corpus is used to further pre-train the deep learning NLP model, BERT. We evaluate the results and demonstrate that using a dialect corpus can lead to more accurate sentiment analysis.

3.2 The Dialect Corpus

This section describes the method of creating the dialect corpus. Figure 2 presents an overview of the process.

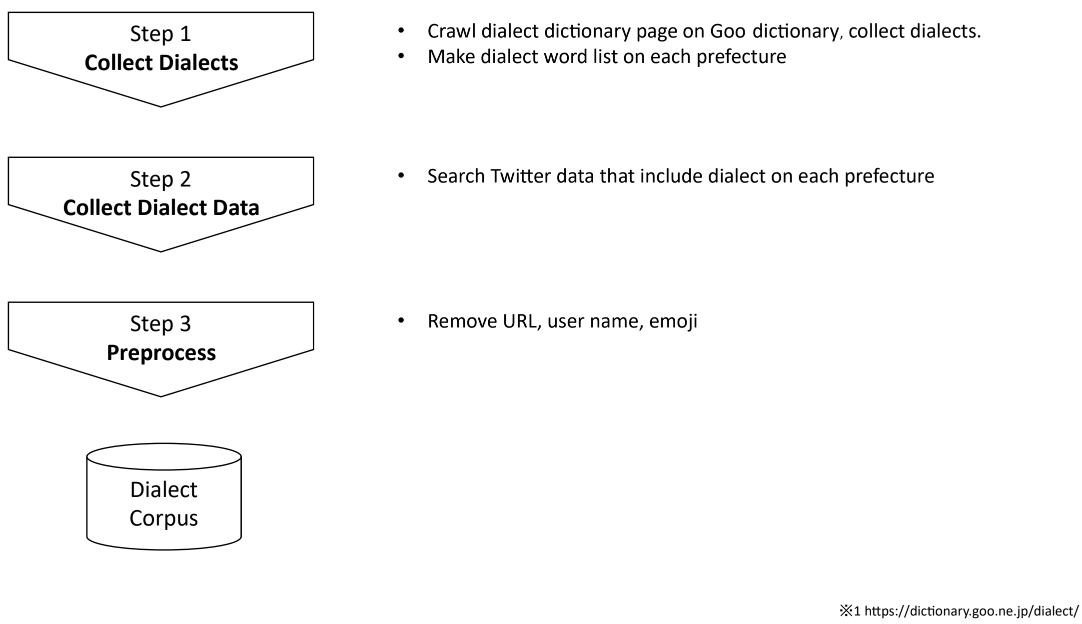

Figure 2 An overview of the process

3.2.1. Dialect Collection

Before acquiring data containing dialects, we first create a list of dialects spoken across Japan. For this, we use the National Dialect Dictionary published in a website, Goo Dictionary[7] operated by NTT Resonance Inc. This website provides dialects spoken by prefecture and region, along with their meanings and examples. We crawled all target pages of this website and obtained the dialects, the region where the dialect is spoken, its meaning, corresponding word in standard Japanese, and usage examples. As a result, 3,610 dialect words were collected.

3.2.2. Dialect Data Collection

Next, for each of these dialect words, we used the API provided by Twitter Inc.[8] to obtain post data containing the dialect. 1,460 dialect words out of 3,610 were in use. Table 1 shows the details of the numbers obtained and used by region. Some of the post data (Tweets) on Twitter retain the user's location information at the time of posting. This time, we added a search condition to match the region where the dialect defined in the Goo Dictionary is spoken and the location information attached when the user posted the tweet. For example, for the Osaka dialect "akan(あかん)", the conditions for searching for post data containing "akan" would be:

- The text part of the post data contains "あかん".
- The data was posted in Osaka Prefecture.

This time, we made the search target area at the prefectural level, but we excluded Tokyo and Kanagawa prefecture from this search target areas. These two prefectures are thought to have a relatively high use of Standard Japanese within the region, as they have a large number of migrants from other regions compared to other areas. Table 2 shows some examples of the data obtained in this way. The data obtained this time was 324,899 items, with one tweet considered as one data. The number(#) of data column in Table 1 shows the number of data obtained for

[7] https://dictionary.goo.ne.jp/

[8] https://developer.twitter.com/en (API has been deprecated as of July 2023)

each region.

Table 1 Number of dialects by region, number used, and data collected.

	# of dialects	# of dialects actually used	# of data
Hokkaido(北海道)	78	49	14,157
Tohoku(東北)	474	245	32,117
Kanto(関東)	406	136	27,582
Chubu(中部)	798	135	41,310
Kinki(近畿)	498	283	101,994
Chugoku(中国)	388	183	37,598
Shikoku(四国)	327	157	18,378
Kyushu(九州)	641	272	51,763
Total(合計)	3,610	1,460	324,899

Table 2 Dialect data samples

Area	Dialect	Meaning of Standard Japanese	Usage Samples
Hokkaido(北海道)	ちょす	触る	アイリさんのお子さんもうスマホをちょすくらい大きくなったのか…
	けっぱる	頑張る	ちょいと風がある朝ですね。今年は帽子が欲しい…ソダシ帽が…しかし…いいのかそれで私…。ま、いっか。今日もけっぱるよー！
Tohoku(東北)	もぞこい	かわいそうだ	骨折中の息子。1日12時間以上寝てる。襧豆子のよう。寝て寝て治せ。もぞこいなぁ。
	きどごろね	うたたね	きどごろねすっと、風邪引くべぇ〜w
Kanto(関東)	おしゃらぐ	おしゃれ	ワークマン女子行ってみました❣️広い店内でゆっくり見られておしゃらぐ✨
	めためた	めったやたらに	めためた眠いし、今日初外出🙄がんばるんば
Chubu(中部)	やっとかめ	久しぶり	明日は学校〜〜みんなにやっとかめに会えるし〜〜2日いきゃ土日だし〜〜さいこうじゃん！
	おぞい	わるい	家の中、PCで聴いてもいいんだけど、PC用スピーカーがおぞいもんで、購入を検討中
Kinki(近畿)	がめつい	けちな	給料だけはしっかりしてくれ・・ お金にがめつい事は言いたく無いがこちらも生活掛かってるからな
	いらう	触る	まだ。起きてる。スマホ、いらうと、寝れなくなるから。落ち着かないと
Chugoku(中国)	はぶてる	すねる	わたくしはそんなことではぶてるようなケチな人間じゃあない。
	ちばける	ふざける	わしも、一つの仕事だけしてないけんな。あんまりちばけるなよ。 かぐらお交番の警察も勘違いするなよ。
Shikoku(四国)	やねこい	めんどう	ありがとうございますm(_ _)m次は香川といきたいところですが、琴電がやねこい💦
	じょんならん	どうにもならない	風呂はいりたすぎてじょんならん😰
Kyushu(九州)	しぇからしか	うるさい	えーい、しぇからしか。朝イチ取りに行こかねw…いかん、いかん。。。
	めんそーれー	いらっしゃい	風が強くなってきた。台風さんめんそーれー

3.2.3. Preprocess

Given the features of tweets, there are items directly unrelated to the content of the post, such as mentions (usernames written after the @ symbol), genre or topic tags added after the # (hashtag), and links to webpages (URL) or images. Also, two type of emojis, one is composed of symbols and alphanumeric characters (examples: v(^^)v, m(_ _)m) and the other is defined in Unicode U+0023 to U+1FA95 (examples: 😆👍) are often included. In this study, we aim to perform analysis without depending on a specific platform and solely based on text. Considering these objectives and the characteristics of tweets, we conducted the preprocessing as follows. The text before and after preprocessing is shown in Table 3.

Preprocessing Procedure

1. Extract only the text attributes from the tweet objects responded from the Twitter API.
2. Remove @username, #hashtags, and webpage and image URLs.
3. Consolidate consecutive symbols (example: !!!, ???), spaces, and line break characters into one.
4. Remove both type of emojis.

Table 3 Examples of text before and after preprocessing

Before Preprocessing	After Preprocessing
今週末はカターレの試合無いから、あいそんないわ😫#kataller #カターレ富山	今週末はカターレの試合無いから、あいそんないわ。
うぁ😩かんにんしてや~~🐤💦#阪神タイガース	うぁ。かんにんしてや~~
＼('ω')／ｳｵｵｵｵｵｱｱｱｰｰｰｯ！ #京都飛ばし ！かんにんやで～！#ビッカメ娘	かんにんやで～
@BGM_10ve10ve やっぱ鈍感だ😂😂😂それかいけず～やな😂😂新潟はいつでもいけるがな笑	やっぱ鈍感だ。それかいけず～やな。新潟はいつでもいけるがな笑
えーっ！もち米って何？えーい！とにかくこれで炊くのだ！断定！あんじょーだざ！ http://t.co/8Y4JhkPDx	えーっ！もち米って何？えーい！とにかくこれで炊くのだ！断定！あんじょーだざ！
うわ～、和枝ちゃん以上のいけず(￣Д￣)ﾉ #とと姉ちゃん #ごちそうさん	うわ～、和枝ちゃん以上のいけず

3.3 Proposal Method

As mentioned in the previous section, the data acquired and preprocessed from Twitter is used as the dialect corpus and use for further pre-training. As the base pre-training model for further pre-training, we use the BERT model made publicly available by Inui et al. [15]. Using the further pre-trained model (Dialect BERT) as a base model, we fine-tune it for intensity of eight emotions and their polarities, thus creating BERT models specialized for each emotion analysis.

For fine-tuning, we use the dataset WRIME [3] made publicly available by Kajiwara et al., which is used for emotion analysis. The whole process of our proposed method is shown in Figure 3. Our experiment comprises three major steps. We conduct the evaluation of emotion analysis accuracy using the final emotion BERTs.

1. A dialect dictionary is created from the dialect list made during the dialect corpus creation. This dictionary is be converted into a format installable into MeCab[9], thus creating a MeCab capable of understanding dialects, which we call DialectMeCab.

2. Using the dialect corpus and DialectMeCab, further pre-training is conducted on the pre-training model to create DialectBERT. The parameters used in this training are the same as those applied by Inui et al.

3. The DialectBERT is fine-tuned to predict the values of intensity of eight emotions and emotional polarity labeled in the WRIME dataset, thus creating individual emotion BERTs for each emotion (JoyBERT, SadnessBERT, AnticipationBERT, SurpriseBERT, AngerBERT, FearBERT, DisgustBERT, TrustBERT).

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

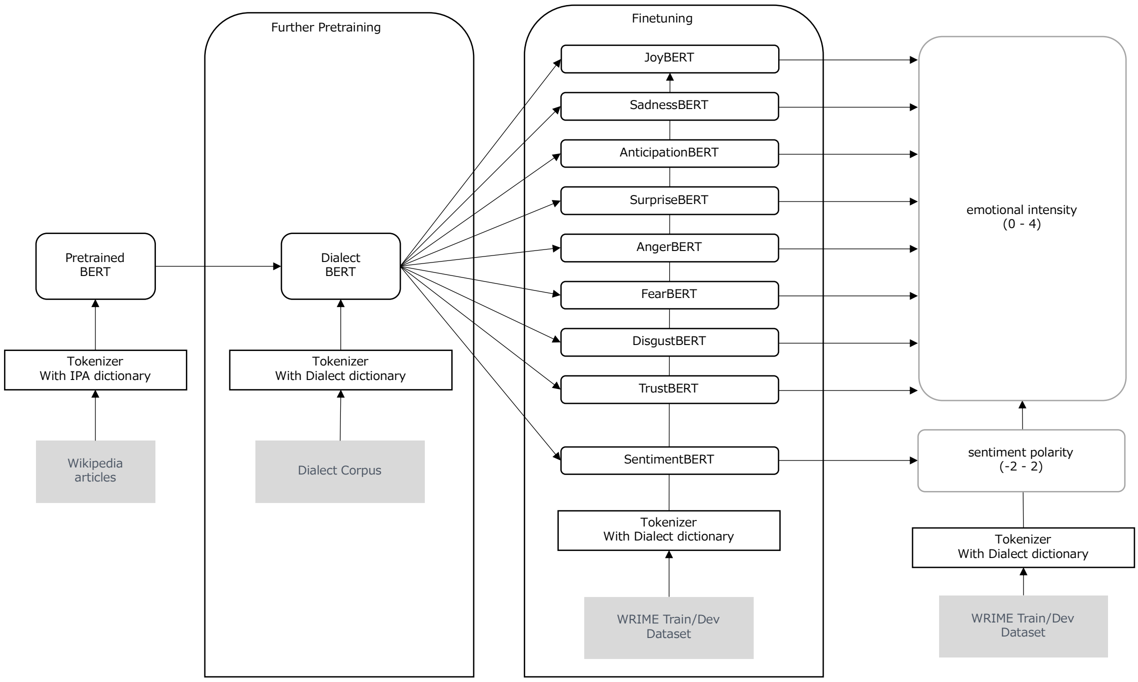

Figure 3 Overall process of the emotion analysis model

3.3.1. WRIME

As mentioned earlier, WRIME[10] is a subjective and objective emotion analysis dataset publicly released by Kajiwara et al. [3]. In this study, we use the Ver.2 dataset of 35,000 Twitter posts collected from 60 authors. From both the subjective (one writer of the text) and objective (three employed from crowd workers) perspectives, each post data is labeled with Plutchik's [24] basic eight emotions (Joy, Sadness, Anticipation, Surprise, Anger, Fear, Disgust and Trust) intensity in four stages (none, weak, medium, strong), and emotional polarity in five stages (strong negative, negative, neutral, positive, strong positive). Plutchik proposed in his emotion theory that humans have eight basic and primitive emotions, and all other emotions are either a mix or derivative of these eight emotions. Also, these emotions can pair with their opposite emotions (Joy and Sadness, Trust and Disgust, Fear and Anger, Surprise and Anticipation). Table 4 shows samples of the WRIME dataset labeled by emotion intensity and emotional polarity. In this study, we use the data labeled from the subjective perspective for

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

the analysis of the writer's emotions, considering the detection of flaming and cyberbullying.

Table 4 WRIME Data Sample, Source [4]

Text	I'm taking the summer off next month to go out! I'm looking forward to it!								
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Sentiment polarity
Writer	3	0	3	0	0	0	0	0	2
Reader 1	2	0	2	0	0	0	0	0	2
Reader 2	3	0	3	0	0	0	0	0	2
Reader 3	3	0	3	0	0	0	0	0	2
Text	My umbrella was stolen!!								
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Sentiment polarity
Writer	0	2	0	0	2	0	3	0	-2
Reader 1	0	0	0	0	2	0	0	0	-1
Reader 2	0	3	0	0	3	0	3	0	-2
Reader 3	0	0	0	0	3	0	0	0	-2
Text	Snowy morning with a light dusting of snow on the roof...								
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Sentiment polarity
Writer	0	0	0	2	0	0	0	0	0
Reader 1	0	0	0	0	0	0	0	0	0
Reader 2	1	0	0	1	0	0	0	0	0
Reader 3	0	0	0	1	0	0	0	0	0

3.3.2. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model for NLP proposed by Devlin et al. [13], based on the Transformer model proposed by Vaswani et al. [14]. BERT is structured using the Encoder part of the Transformer. The Transformer is a deep learning model based on the encoder-decoder architecture with an attention mechanism (Figure 4). The Encoder has six layers, each consisting of a multi-head attention layer and a feedforward layer. The part of Decoder is similar to the Encoder but has an additional multi-head attention layer that processes the output from the Encoder. Thanks to this attention mechanism, the Transformer has solved the issue of long-term dependencies that was a problem in traditional RNN models. Devlin et al. implemented and tested a model with 12 layers of the Transformer, $\text{BERT}_{\text{BASE}}$ and a model with 24 layers, $\text{BERT}_{\text{LARGE}}$. The BERT provided by Inui et al. [15], which

we use as a pre-training model, is the same size as $BERT_{BASE}$.
BERT is pre-trained on tasks known as the Masked Language Model (MLM) and Next Sentence Prediction (NSP), using unlabeled data. MLM is a task that hides multiple words in a single sentence with [MASK] and predicts the hidden words. The pre-training data was created by replacing 80% of the data set with [MASK], replacing 10% with randomly chosen words instead of [MASK], and leaving 10% unchanged. NSP is a task in which two sentences are given connected by a [SEP] token, and it is determined whether these are continuous sentences.

After pre-training, fine-tuning is performed according to the task to be solved (downstream tasks). During fine-tuning, a layer of the appropriate shape for the task to be solved is added after the last layer of BERT. Additionally, BERT has the characteristic that there is little difference between the pre-trained architecture and the architecture during fine-tuning. The overall image of pre-training and fine-tuning is shown in Figure 5.

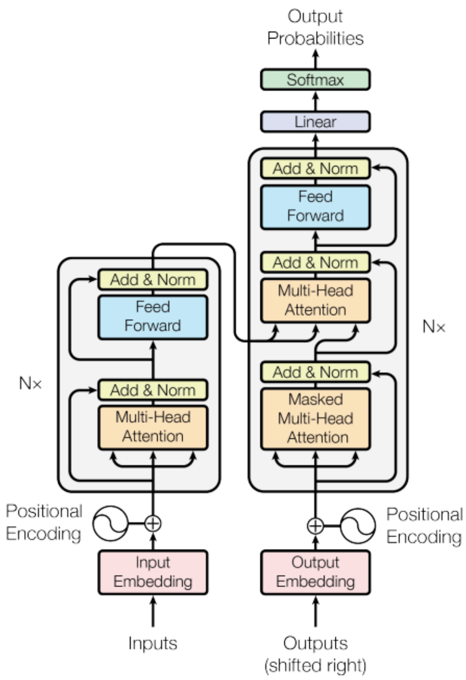

Figure 4 The Transformer – model architecture, Source [14]

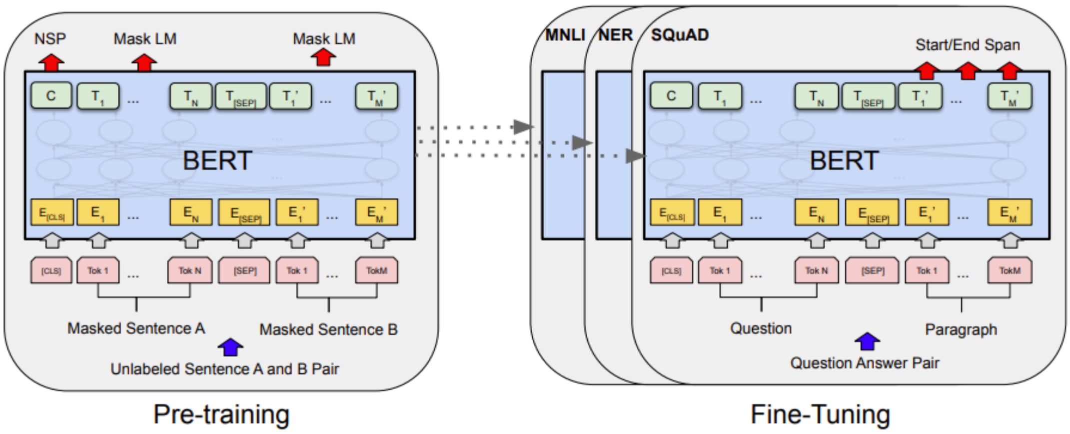

Figure 5 An Overall pre-training and fine-tuning procedures for BERT, Source [13]

3.3.3. Morphological Analysis

In this study, we use MeCab[11] for morphological analysis. MeCab comes with a standard system dictionary called ipadic. In addition, there is mecab-ipadic-NEologd, which has added new words derived from language resources on the Internet. In this study, we use this mecab-ipadic-NEologd[12] dictionary as a base dictionary. In addition to this dictionary, we create a new dialect dictionary and install it in MeCab. This MeCab, which has the dialect dictionary installed, is called DialectMeCab, and we customize it so that it can correctly perform morphological analysis of dialects. An excerpt from the dialect dictionary we created this time is shown in Table 5. Among the items needed in the dictionary, left context ID, right context ID, and cost columns are set to 0 because they are unused items according to the specifications[13]. For the other features, since there are no predetermined items in the MeCab specification in order to enhance the versatility of the system, we set the same part of speech information as the standard Japanese word with the same meaning (from column 5 to column 11). Columns 12 and 13 are reading and pronunciation, and since all dialects are in hiragana, we set the same as the dialect for each. In column 14, we set the string "Dialect(方言)" as a flag so that it is understood that this word is a dialect.

---

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

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

[13] http://taku910.github.io/mecab/learn.html#seed

The results of morphological analysis using this dictionary are shown in Table 6. The dialect "chosu(ちょす)" is a word meaning "to touch(触る)" in standard Japanese. In MeCab without the dialect dictionary installed, it is analyzed as the noun "cho(ちょ)" and the verb "su(す)". In DialectMeCab, it correctly recognizes "chosu(ちょす)" as a single word and understands that it is a dialect.

Table 5 Sample of Dialect Dictionary

surface form(word itself)	Left Context Id	Right Context Id	Cost	Feature	Feature	Feature	Feature	Feature	Feature	Feature	Feature	Feature
～んやて	0	0	0	接続詞	*	*	*	*	*	～んやて	～んやて	方言
おーぼ	0	0	0	名詞	一般	*	*	*	*	おーぼ	おーぼ	方言
かざく	0	0	0	動詞	自立	*	*	五段・ガ行	基本形	かざく	かざく	方言
かてる	0	0	0	動詞	自立	*	*	五段・カ行イ音便	基本形	かてる	かてる	方言
かんぱ	0	0	0	名詞	固有名詞	一般	*	*	*	かんぱ	かんぱ	方言
きとろ	0	0	0	名詞	一般	*	*	*	*	きとろ	きとろ	方言
きなんば	0	0	0	名詞	一般	*	*	*	*	きなんば	きなんば	方言

Table 6 Example of the results of morphological analysis of phrases including dialect.

Sentence for Morphological Analysis	with Dialect Dictionary	Result
スマホを**ちょす** ※「**ちょす**」は北海道の方言で「触る」	without dictionary	スマホ 名詞,固有名詞,一般,*,*,*,スマホ,スマホ,スマホ を 助詞,格助詞,一般,*,*,*,を,ヲ,ヲ **ちょ** 名詞,動詞非自立的,*,*,*,*,ちょ,チョ,チョ **す** 動詞,自立,*,*,サ変・スル,文語基本形,する,ス,ス
	with dictionary	スマホ 名詞,固有名詞,一般,*,*,*,スマホ,スマホ,スマホ を 助詞,格助詞,一般,*,*,*,を,ヲ,ヲ **ちょす** 動詞,自立,*,*,五段・ラ行,基本形,ちょす,ちょす,方言

3.3.4. Further Pre-training

The further pre-training model employs the BERT model provided by Inui et al. [15] as the base model. The architecture of this model is identical to the original BERT implementation by Devlin et al. [13], with 12 Transformer layers and 768 dimensions in the hidden layers. The training data comprises Japanese Wikipedia articles, representing a dataset of approximately 17 million sentences that were used for pre-training[14]. We further pre-train this model using the dialect corpus created in this study and the DialectMeCab described in the previous section. Since the base model was pretrained with Masked Language Modeling (MLM), the same MLM approach is adopted in this study. MLM involves masking and predicting words, a concept illustrated in Figure 6. The dialect corpus is tokenized using DialectMeCab, and the model is pre-trained with randomly masked data. The parameters used in this further pre-training are presented in Table 7.

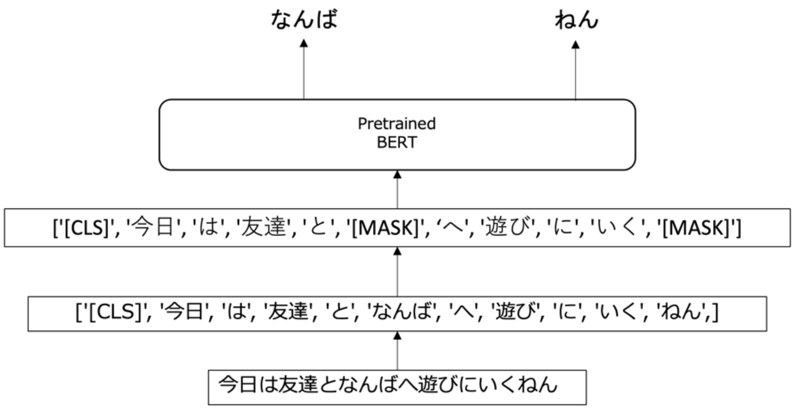

Figure 6 MLM process

14 https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

Table 7 The parameters of further pre-training

		Futher Pretraining	Fine tuning
data size	training	258,657	30,000
	evaluation	32,332	2,500
	test	32,332	2,500
batch size		32	32
epoch		15	3
learning rate		2.0E-05	2.0E-05
optimizer		AdamW	
loss function		CrossEntropy	
vocabulary size		32,000	32,000

3.3.5. Fine-tuning

We fine-tune the model pre-trained in 3.3.4 for sentiment analysis. We make nine copies of the pretrained model, fine-tuning each one for the eight emotions labeled with WRIME. The remaining model is fine-tuned for emotion polarity. The fine-tuning process follows the same methodology as Devlin et al. [15]. At the end of the BERT model, we add a Linear layer with the number of output labels (four for sentiment analysis, five for emotion polarity prediction). For training, we utilize the output from the special [CLS] token placed at the beginning of the input tokens. This output is connected to the Linear layer and trained to match the correct output. This process is illustrated in Figure 7. In this way, we create eight sentiment analysis models and one emotion polarity analysis model.

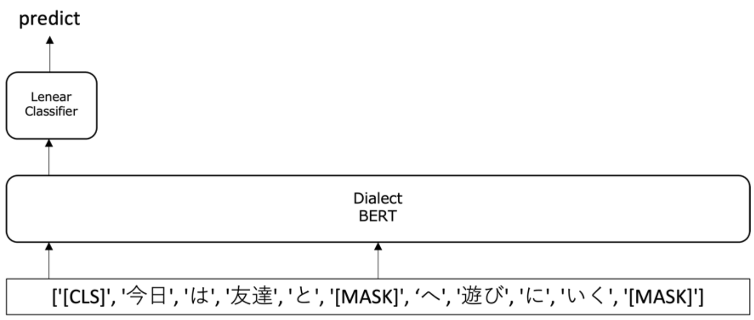

Figure 7 Fine-tuning image.

3.4 Results and Evaluation

The results obtained using the models acquired in the previous section are presented. For comparison, we use the results from the LSTM model as the base value. In addition to the base value, we conduct analysis with four combinations: with and without the use of DialectMeCab and DialectBERT, comparing their respective accuracies. For the evaluation of the analytical models fine-tuned with the eight emotions, we use the Mean Absolute Error (MAE). For the analytical models fine-tuned for emotion polarity, we use Accuracy as the evaluation metric.

3.4.1. Long Short-Term Memory (LSTM)

The LSTM model is a type of Recurrent Neural Network (RNN). The LSTM introduces a mechanism called a memory cell, which is designed to avoid the vanishing gradient problem that occurs in other RNN models. It does this by controlling the state of the memory cell using functions known as the input gate, forget gate, and output gate. The network configuration is illustrated in Figure 8. A simple configuration was implemented with one layer of LSTM.

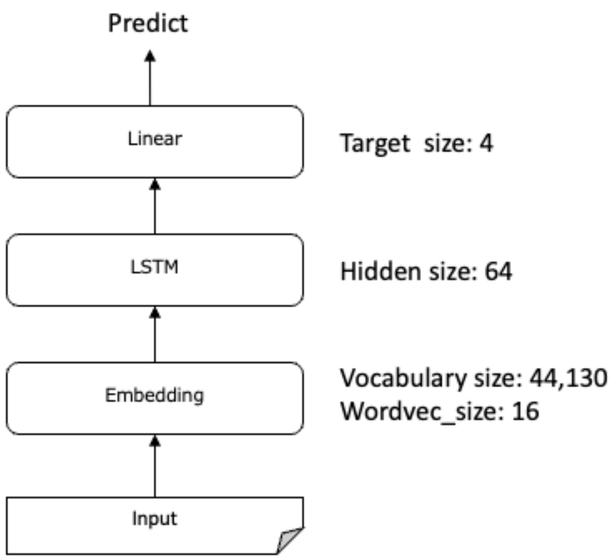

Figure 8 Structure of LSTM base model

3.4.2. MAE and Accuracy

MAE, or Mean Absolute Error (Equation 3.1), is an evaluation method that calculates the average of the absolute differences between the predicted and actual values produced by a model. Accuracy (Equation 3.2), on the other hand, is a measure of the proportion of correct predictions made by the model. Here, TP represents True Positive, TN represents True Negative, FP stands for False Positive, and FN denotes False Negative.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{3.1}$$

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{3.2}$$

3.4.3. Results of Sentiment Analysis

Figure 9 shows the loss value for each epoch during pre-training. Table 8 presents the analysis results for each of the eight emotions, while Table 9 shows the results for emotion polarity. Among the eight emotions, the pattern utilizing both DialectBERT and DialectMecab achieved the highest accuracy for Joy, Anticipation, Surprise, and Anger. Additionally, the pattern only using DialectBERT yielded the best results for Sadness and Disgust. For these six emotions, there was an average difference of 0.163 from the least accurate pattern. Also, there was a difference of 0.214 from the results of Kajiwara et al. These results suggest that models understanding dialects are beneficial for comprehending these emotion intensities. On the other hand, for Fear and Trust, the base model yielded the highest accuracy. It is thought that this might be due to situations involving Fear and Trust being less likely to include dialectal expressions, meaning that the model's understanding of dialect did not affect the results. Regarding emotion polarity, the pattern using both DialectBERT and DialectMeCab demonstrated a better ability to accurately discern sentiment polarity. Next, Tables 10 and 11 display the results comparing accuracies by data size. For determining Joy, Sadness, Fear, Disgust, and Trust, results were more accurate than those obtained with the full dataset of about 320,000 instances. The remaining emotions - Anticipation, Surprise, and Anger - showed only a slight average difference of 0.009 when compared to the results using 150,000 instances. While it's common in deep learning model training to assume that more data is always better, our experiment suggests that for sentiment analysis, a sufficient level of accuracy can be achieved with a data size of about 100,000 to 150,000 instances. As the full dataset yielded the highest accuracy for emotion polarity, it appears necessary to validate the model with an increased amount of data.

Epoch	Training Loss	Validation Loss
1	3.50	3.40
2	3.32	3.22
3	3.24	3.14
4	3.16	3.08
5	3.09	3.04
6	3.02	3.01
7	2.98	2.99
8	2.95	2.96
9	2.91	2.94
10	2.85	2.92
11	2.86	2.93
12	2.84	2.89
13	2.81	2.87
14	2.80	2.87
15	2.84	2.88

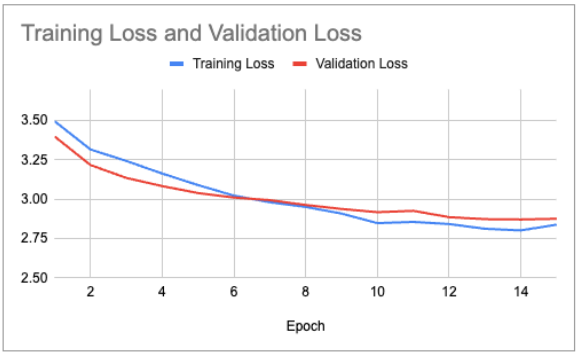

Figure 9 Training and validation loss

Table 8 Predictions by emotion

	MAE							
	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust
Kajiwara et al,. (as reference)	0.734	0.666	0.899	0.684	0.218	0.344	0.443	0.432
LSTM	0.773	0.481	0.722	0.569	0.180	**0.246**	0.265	**0.428**
BERT + MeCab	0.693	0.442	0.699	0.578	0.179	0.275	0.264	0.468
BERT + DialectMeCab	0.691	0.440	0.700	0.578	0.178	0.268	0.258	0.476
DialectBERT + MeCab	0.658	**0.433**	0.658	0.562	**0.170**	0.260	**0.252**	0.468
DialectBERT + DialectMeCab	**0.646**	0.448	**0.652**	**0.552**	**0.170**	0.265	0.255	0.481

Table 9 Predictions for emotion polarity

	Accuracy
Suzuki et al,.(as reference)	39.10%
LSTM	22.84%
BERT + MeCab	39.76%
BERT + DialectMeCab	40.88%
DialectBERT + MeCab	43.00%
DialectBERT + DialectMeCab	**43.12%**

Table 10 Evaluation of sentiment analysis by data size

	size of data	MAE							
		Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust
DialectBERT + DialectMecab	5k	0.666	0.453	0.684	0.571	0.173	0.273	0.249	0.476
	10k	0.658	0.442	0.667	0.561	0.174	0.263	**<u>0.248</u>**	**<u>0.464</u>**
	50k	0.659	0.431	0.656	0.562	0.171	0.263	0.251	0.475
	100k	0.654	**<u>0.430</u>**	0.675	0.570	0.174	0.265	0.254	0.483
	150k	**<u>0.644</u>**	0.434	0.665	0.564	0.173	**<u>0.259</u>**	0.258	0.475
	full	0.646	0.448	**<u>0.652</u>**	**<u>0.552</u>**	**<u>0.170</u>**	0.265	0.255	0.481

Table 11 Evaluation of emotion polarity by data size

	size of data	Accuracy
DialectBERT + DialectMecab	5k	41.12%
	10k	40.84%
	50k	41.56%
	100k	40.92%
	150k	42.56%
	full	**<u>43.12%</u>**

Chapter 4
Flaming and Cyberbullying Detection Using Sentiment Analysis Models

4.1 Objectives

We examine a method for detecting conversations where flaming and cyberbullying are occurring, using the sentiment analysis model created in Chapter 3. If a conversation involves flaming or cyberbullying, it is presumed that the texts of the conversation contain a significant number of emotions such as anger, disgust, and sadness, while there are less of emotions like joy and trust. It is believed that if these emotions can be predicted more accurately, it would be possible to detect conversations where flaming and cyberbullying are occurring with a higher degree of accuracy.

4.2 Flaming and Cyberbullying Conversation Data

In this section, we explain the method for creating the data to be used in determining flaming and cyberbullying. In this study, we obtain conversation data using Twitter's API, separate from the dialect corpus used in the previous chapter. Therefore, when collecting the data, we do not include the presence or absence of dialects or location information of user posts in the search keywords or conditions. The overall picture of data creation is shown in Figure 10.

In Twitter, you can notify the original poster to your post by posting a reply that includes "@username" in the text part of your post. You can then reply to that reply, and by continuing this process, you can have a conversation between the posters. In addition, it is possible to send multiple replies to a single post or for one person to reply to multiple times. We use these conversation data in this experiment. It is believed that conversations where flaming or cyberbullying are occurring include words that lead to abuse or threats (such as "die(死ね)", "gross(キモい)", "kill(殺す)", and so on). This time, we listed words that are generally considered to be slander in Japanese (Table 12) using websites[15][16][17] that introduce words that are defamatory, abusive, or threatening. We searched and collected data containing these words on the Twitter. The number of conversations obtained was 7,547, and the total number of posts were 183,516. The data obtained by this method was reviewed one by one, and classified into categories 1,2,3,4 below.

1. Flaming or Cyberbullying is occurring.
2. Flaming or Cyberbullying is occurring on political topics.
3. Normal conversations.
4. Unrelated

For cyberbullying, the criteria were whether the conversation contains elements of bullying, and if it consists of consecutive replies insulting or threatening a specific individual. We defined flaming as cases where a large number of unspecified individuals are using words that lead to insults or threats towards an organization, incident, accident, or event, not a specific individual. In flaming on political topics, the target of the content of the post is specifically a politician or political group, and the discussion is excessively aggressive. For the unrelated label, data such as only replies that include advertising or cases where the poster is the only one replying to their own posts (replying to one's own post), i.e., although there are consecutive replies, they do not actually constitute a conversation, were classified. For data that does not fall into any of the above classifications, the classification was set as "ordinary conversation". No.1 to 3 are the classifications considered to represent flaming or cyberbullying in this study. Out of the 6,830-conversation data checked, 190 were labeled as conversations involving flaming or cyberbullying. Table 13 shows samples of data classified as cyberbullying.

---

[15] https://sakujo.izumi-legal.com/column/chishiki/insult-jirei

[16] https://amata-lawoffice.com/deletion-request/types-of-slander/

[17] https://best-legal.jp/slander-slander-6888/

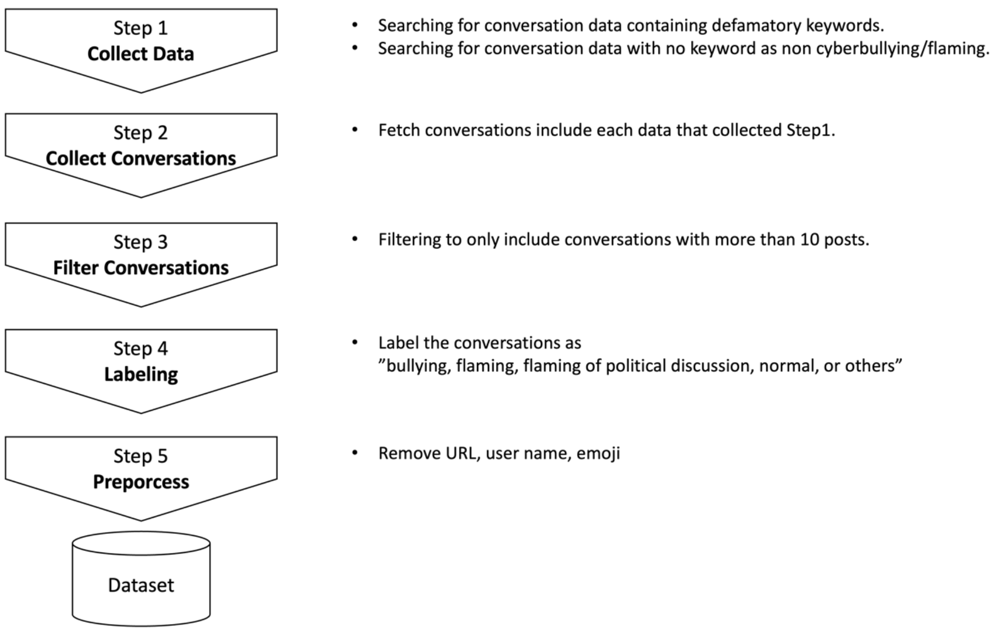

Figure 10 An overview of flaming and cyberbullying data creation

Table 12 Defamatory word list

Defamatory Word	Meaning of words
死ね	Die
しね	
氏ね	
タヒね	
ばか	Idiot, Fool
馬鹿	
バーカ	
ばーか	
バカ	
うざい	Annoying
あほ	Idiot
阿呆	
でぶ	Fat
デブ	

Defamatory Word	Meaning of words
きもい	Gross
去れ	Go away, Disappear
消えろ	
きえろ	
消え去れ	
きえされ	
くたばれ	Drop dead
くそやろう	Bastard
消すぞ	Kill
殺す	
殺害する	
殺す	
ころす	
コロス	

Defamatory Word	Meaning of words
根性なし	Weakling
ヘタレ	
あたおか	Crazy
性格悪	Bad personality
せいかくわる	
チビ	Short
ビッチ	Bitch
ボケ	Coward
逝っ	(Kind of) Die
うるせ	Shut up
負け犬	Loser

Table 13 Sample conversations cyberbullying is occurring.
(User ID has been removed for privacy protection)

Bullying conversation #1	Bullying conversation #2	Bullying conversation #3
ググッても出てこんやろ笑	お前が死ね	そうそうキモいよな。でもな、キミは死ねって言っちゃダメだ。は→（てめぇは裸になりゃぁ 俺たちに悪口言われたくないならお前らに謝れ）
Googleに聞けwww	なるほどな夜遊びもほどほどにな……	変態ゴリゴリだな
なんでそんなに変態なのかを教えて	道連れにすんぞいつ帰ってくるんや	だよなー。ホントに
4ね	最高の兄弟すからね笑	そうだそうだてめぇは勝手に全裸のままでいれば？
死ね	お前ら兄弟仲良すぎきっしょ	そういうことだ。お仕置きはオレに任せとけ。
うわ💧　キモスギィ	ダルいっすわー笑笑	そうそう
にっこり笑顔	おまえオカマなめんなよ掘り殺すぞ	可愛くないね
待ってきも	きっしょ	そうだそうだ貴様は消えろと言いたくないならエ○本とか女子裸の写真集をよこせ。
よんごーよんごー	絡みダル！殺すぞ	こわいよねこの人。悪い子じゃね？こいつ
何語、	殺すぞ	てめぇはオレの仲間じゃねぇ可愛くねぇしバカ女だ。だから全裸になって全部脱いで写真を取ってツイートすれば？そして全裸で生活すれば？てめぇに言われたくないならみんなに謝れよ。
545	ダルいっすわー笑笑	きしょいきしょいきしょいきしょいきしょいきしょい笑笑も一辞めて笑笑！
4	お前が死ね	ゴリラやんけ
怖いね〜	なるほどな夜遊びもほどほどにな……	きも
殺すゴ	道連れにすんぞいつ帰ってくるんや	頭平気ー？
うるさいモ	最高の兄弟すからね笑	お前がブサイク過ぎて面白いからフォローされてんだよｗｗｗ気づけｗｗｗ
おはようゴ	お前ら兄弟仲良すぎきっしょ	きもいゴリラ
おー、分かってくれるかwww嬉しいモ	ダルいっすわー笑笑	きゃわたんなら困らんやろブスやから困っとるんやわ気づけ。
似てないことも無い笑	おまえオカマなめんなよ掘り殺すぞ	んーとね0点くらいかな!!本当は -1000000000000000000000000000000点にしたいけど♥
違うわ。クレヨンしんちゃんの園長先生やしwwwwwwwww嘘	きっしょ	きーも　きーも帰れしねー！
きざる	絡みダル！殺すぞ	冗談は顔だけにしてくれwwwwwきもいぞ
しばく	殺すぞ	おいブス
は、死ねゴキブリ	暇な子テレフォンしましょー♡	しね
Following data were omitted		**Following data were omitted**

4.3 Proposed Methods

Using the conversation data and the eight emotion analysis models, we predict the occurrence of flaming and cyberbullying in conversations. It is demonstrated that these predictions can be made based on the combination of the strengths and weaknesses of the eight emotions (Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation). An eight-dimensional vector called an "emotion vector" was created from each conversation data. The accuracy of the prediction using this emotion vector is compared using a method that predicts the occurrence or non-occurrence based on similarity, and a method using machine learning algorithms. The procedure for creating the emotion vector is presented in the following section.

4.3.1. Emotion Vectors

The procedure for creating the emotion vector is shown in Figure 11. Using the eight emotion analysis models obtained from emotion analysis, we evaluate the intensity of emotions of each data. In the example of Figure 11, we first use JoyBERT, which has been fine-tuned to predict Joy emotion, to predict to what extent the speakers of the utterances were feeling joy "こわっ(Scary)", "アイコン見てると呪われそうで怖い(Looking at the icon feels like being cursed, it's terrifying)", "そーゆうのを自分で言う人はブスなんですよ(People who say that to themselves are ugly)", "まじきもい(Seriously gross)". Afterward, the evaluated values for each utterance are averaged, and that value is used as the Joy value of this conversation. Following the same procedure for the other seven emotions, we create the Sadness value, Trust value, Disgust value, Fear value, Anger value, Surprise value, and Anticipation value of this conversation, and put these together to form an eight-dimensional vector. This is called the "emotion vector". The definition of the emotion vector is shown in Equation (4.1).

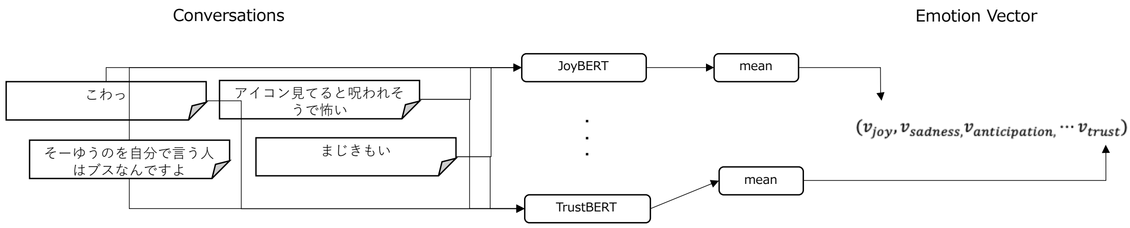

Figure 11 The process of emotion vector creation

$$V_{emotions} = (v_{joy}, v_{sadness}, v_{trust}, v_{disgust}, v_{fear}, v_{anger}, v_{surprise}, v_{anticipation}) \quad (4.1)$$

4.4 Results and Evaluation

In this section, we describe the evaluation experiments of the proposed method. As mentioned in 4.2, the number of flaming and cyberbullying data prepared for this experiment is 190. To this, we randomly sampled 190 of data labeled as "normal conversation" and conducted experiments with a total of 380 of data. In this evaluation experiment, we divided all the data into 80% training data (flaming and cyberbullying conversations 152, normal conversations 152) and 20% (flaming and cyberbullying ones 38, normal ones 38) as test data.

4.4.1. Prediction Using Vector Similarity.

To create a reference vector for calculating similarity, we use only the flaming and cyberbullying data (152 items) out of the 304-training data. We calculate the emotion vector for each conversation data using the method mentioned in 4.3.1. Finally, we obtain the average value for each element of these emotion vectors and obtain an eight-dimensional vector using these values. Since we are only using flaming and cyberbullying data here, we call this vector the "flaming and cyberbullying vector". We then compare this obtained flaming and cyberbullying vector with the emotion vector of the evaluation data using cosine similarity. Cosine similarity is a measure to determine the similarity between two vectors. The formula for cosine similarity is shown in Equation 4.1. By measuring the angle between vectors, it determines how similar they are. Cosine similarity takes a value from -1.0 to 1.0, and the closer it is to 1.0, the more similar the vectors are. Therefore, conversation data whose emotion vector is close to the flaming and cyberbullying emotion vector can be predicted to be a conversation where flaming and cyberbullying are occurring. Figure 12 represents the values of the flaming and cyberbullying vectors when using the BERT model and MeCab, and DialectBERT and DialectMeCab, respectively. Both cases show strong emotions of anger and disgust, with less appearance of emotions such as fear, trust, and joy. Particularly when using DialectBERT and DialectMeCab, these emotions are even more pronounced. Table 14 shows the judgment results using cosine similarity, broken down by similarity (80%, 90%). These are results judged using test data. When conversations with a similarity of 80% or more were judged as flaming or cyberbullying, the judgment accuracy was 92.1%. In either case, the accuracy was higher when DialectBERT and DialectMeCab were used. This demonstrates the importance of understanding dialects in detecting flaming and cyberbullying.

$$\cos\boldsymbol{\theta} = \frac{\boldsymbol{v_1} \cdot \boldsymbol{v_2}}{\|\boldsymbol{v_1}\|\|\boldsymbol{v_2}\|} \qquad (4.1)$$

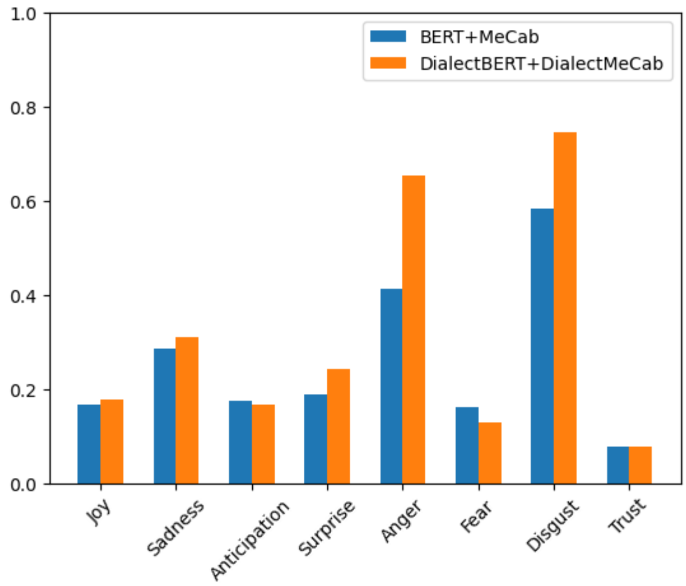

Figure 12 Vector values of flaming and cyberbullying conversations

Table 14 Results of prediction using vector similarity.

	Accuracy of > 80% similarity	Accuracy of > 90% similarity
BERT and MeCab	90.70%	81.50%
Dialect BERT and Dialect MeCab	92.10%	86.80%

4.4.2. Prediction Using Machine Learning Algorithms

We conducted predictive validation using the classification algorithms of machine learning (Support Vector Machine (SVM), AdaBoost, Bagging, ExtraTrees, Gradient, RandomForest, KNeighbors, DecisionTree, ExtraTree) implemented in scikit-learn[18], an open-source machine learning library in Python, using the emotion vector. We trained and evaluated all models using cross-validation (the number of divisions is 5). Cross-validation is a method for checking the generalization performance of a model. In cross-validation, the dataset is equally divided, and training and evaluation are repeated as many times as the number of divisions. In this process, one of the divided datasets is used as test data and the rest as training data. Finally, the accuracy of each round is averaged to evaluate the final accuracy. The results are shown in Table 15. SVM and RandomForest had the highest accuracy, both at 93.42%. In addition to cross-validation, we conducted a search for combinations of hyperparameters. The combinations of hyperparameters and their results are shown in Table 16. The parameters with the highest accuracy in the combination of BERT and MeCab are indicated with an underscore, and those in the combination of DialectBERT and DialectMeCab are in bold. Next, we compared the accuracies by combinations of emotions. We compared the combinations of pairs of emotions mentioned in 3.3.1 (6 combinations in total) (Table 17). The best parameters obtained by GridSearch listed in Table 17 were applied to the hyperparameters of each model. As a result, the combinations of Anger, Fear and Disgust, Trust, and Disgust, Trust and Anticipation, Surprise had the highest accuracy, and the average result of the nine algorithms was 91.08%. This result was 0.29 points better than the results in Table 16. Looking at the prediction results using these machine learning algorithms, similar to the results using cosine similarity, the prediction accuracy was higher when DialectBERT and DialectMeCab were used. It was demonstrated that accurate detection of flaming and cyberbullying is possible using a dialect corpus if there are at least four emotion analysis models.

[18] https://scikit-learn.org/stable/

Table 15 Flaming and Cyberbullying detection accuracy

	Accuracy									
	SVM	**AdaBoost**	**Bagging**	**ExtraTrees**	**Gradient**	**RandomForest**	**KNeighbours**	**DecisionTree**	**ExtraTree**	**Average**
BERT and MeCab	90.79%	85.53%	88.16%	90.79%	89.47%	89.47%	89.47%	86.84%	85.53%	88.45%
Dialect BERT and Dialect MeCab	**93.42%**	88.16%	90.79%	92.11%	92.11%	**93.42%**	92.11%	88.16%	86.84%	**90.79%**

Table 16 Search parameters and best ones in GridSearch

SVM

kernel	**rbf**,linear, poly
gamma	0.1, **0.5**, 1.0
C	0.1, **0.5**, 1.0

AdaBoost

n_estimators	5, 7, **10**
learning_rate	**1**, 1.3, 1.5
algorithm	**SAMME**, SAMME.R

Bagging

base_estimator	**None**, RandomForestClassifier
n_estimators	10, 50, **100**
max_samples	**0.5**, 1.0
max_features	**0.5**, 1.0
bootstrap	**True**, False
bootstrap_features	True, **False**

ExtraTrees

n_estimators	**50**, 100, 200
criterion	**gini**, entropy
max_depth	**None**, 10, 20, 30
min_samples_split	2, 5, **10**
min_samples_leaf	1, 2, **4**
max_features	**sqrt**, log2
bootstrap	True, **False**

ExtraTree

criterion	**gini**, entropy
splitter	best, **random**
max_depth	**None**, 10, 20, 30
min_samples_split	**2**, 5, 10
min_samples_leaf	1, 2, **4**
max_features	**sqrt**, log2

Gradient

n_estimators	100, 200, **300**
learning_rate	**0.01**, 0.1, 1.0
subsample	0.5, **0.7**, 1.0
max_depth	**3**, 5, 7
min_samples_split	**2**, 4, 6

RandomForest

n_estimators	100, 200, **300**
criterion	**gini**, entropy
max_depth	**None**, 10, 20, 30
min_samples_split	2, 5, **10**
min_samples_leaf	**1**, 2, 4
max_features	**sqrt**, log2

KNeighbors

n_neighbors	**3**, 5, 7, 9
weights	**uniform**, distance
algorithm	**auto**, ball_tree, kd_tree, brute
p	1, **2**

DecisionTree

criterion	gini, **entropy**
max_depth	**2**, 6, 8, None
min_samples_split	**2**, 5, 10

Table 17 Result by combination of emotion pairs

	SVM	AdaBoost	Bagging	ExtraTrees	Gradient	RandomForest	KNeighbours	DecisionTree	ExtraTree	Average
Anger, Fear, Disgust, Trust										
BERT and MeCab	86.84%	88.16%	89.47%	90.79%	90.79%	88.16%	86.84%	86.84%	84.21%	88.01%
Dialect BERT and Dialect MeCab	**93.42%**	88.16%	92.11%	92.11%	90.79%	92.11%	92.11%	88.16%	90.79%	**91.08%**
Joy, Sadness, Anticipation, Surprise										
BERT and MeCab	82.89%	68.42%	78.95%	80.26%	77.63%	77.63%	82.89%	78.95%	75.00%	78.07%
Dialect BERT and Dialect MeCab	88.16%	81.58%	84.21%	82.89%	80.26%	80.26%	85.53%	78.95%	78.95%	82.31%
Anger, Fear, Joy, Sadness										
BERT and MeCab	88.16%	88.16%	88.16%	85.53%	86.84%	86.84%	86.84%	88.16%	82.89%	86.84%
Dialect BERT and Dialect MeCab	88.16%	88.16%	86.84%	89.47%	89.47%	89.47%	88.16%	88.16%	85.53%	88.16%
Anger, Fear, Anticipation, Surprise										
BERT and MeCab	88.16%	89.47%	90.79%	89.47%	90.79%	90.79%	90.79%	90.79%	86.84%	89.77%
Dialect BERT and Dialect MeCab	86.84%	88.16%	90.79%	88.16%	88.16%	89.47%	92.11%	88.16%	84.21%	88.45%
Disgust, Trust, Joy, Sadness										
BERT and MeCab	85.53%	88.16%	88.16%	86.84%	88.16%	88.16%	85.53%	86.84%	78.95%	86.26%
Dialect BERT and Dialect MeCab	90.79%	89.47%	86.84%	92.11%	**93.42%**	**93.42%**	88.16%	90.79%	88.16%	90.35%
Disgust, Trust, Anticipation, Surprise										
BERT and MeCab	88.16%	89.47%	88.16%	88.16%	89.47%	86.84%	86.84%	86.84%	86.84%	87.87%
Dialect BERT and Dialect MeCab	92.11%	89.47%	89.47%	**93.42%**	**93.42%**	**93.42%**	89.47%	90.79%	88.16%	**91.08%**

Chapter 5
Conclusion

5.1 Summary

In this study, we created a dialect corpus using a dialect dictionary and post data obtained from Twitter, and performed sentiment analysis using DialectBERT, which was pre-trained with this dialect corpus and fine-tuned with sentiment analysis dataset WRIME. We also examined and evaluated a method for detecting flaming and cyberbullying using these models. We confirmed the effectiveness of using DialectBERT for six of the eight emotions (Joy, Sadness, Anticipation, Surprise, Anger, Fear, Disgust, Trust) intensity and these polarities that we evaluated. As a result, it was possible to demonstrate that text containing dialects more strongly reflects the emotions of the writer, and by using this in training, it is possible to construct a model that efficiently understands emotions. We also verified the impact of the amount of training data on accuracy. In training models using deep learning, it is often the case that more data is better, but in this experiment, it was found that if there are about 100,000 to 150,000 instances of data, sufficient accuracy can be achieved in sentiment analysis. In the judgment of conversations where flaming and cyberbullying are occurring, the results using DialectBERT were more accurate, demonstrating that using DialectBERT is effective for detection.

5.2 Future Works

5.2.1. Emotion Analysis

In this study, we focused on dialects as texts that strongly reflect the writer's emotions, but it is presumed that things like trendy words and youth language also contain emotions. Particularly in classification targeting data posted by young people, considering these words is thought to be a method to increase accuracy. Especially when considering cyberbullying that occurs among elementary or junior high school students, these incidents occur not only on SNS like Twitter, but also in conversations within the chat functions of LINE[19](generally used in Japan) or online game platforms. When considering detection in such situations, there is an even greater need to consider the language of young people and trendy words. In addition, we believe that the method used in this research can be applied not only to Japanese dialects but also to dialects that are spoken in other countries.

In the future, we would like to collect words that are considered strongly reflect emotions other than dialects in Japanese and expand our experiments. At the same time, we would like to apply it to foreign languages and advance research in NLP that takes into consideration the words spoken in each country and region.

5.2.2. Flaming and Cyberbullying Detection

The biggest challenge in this field of research is the amount of data. Compared to general conversation data, the amount of conversation data where flaming or cyberbullying occur is extremely small. Even in this study, only 190 out of 6,830 conversation data checked for labeling were judged to be instances where flaming or cyberbullying occurred (2.7%). The reason for this may be that when a real flaming, incident or accident occurs, both the post that caused the flaming and the replies to it are often deleted by the posters. Therefore, continuous data collection and methods to improve prediction accuracy with unbalanced data are required. Next, cyberbullying is not only direct, but can also be indirect, such as ignoring comments or excluding someone from conversations, which does not appear in the conversation data itself. In addition, bullying can exist not only in text but also using images. Therefore, a wide range of learning that considers images, the date and time of the conversations, the timeline of the speakers, etc., is necessary.

---

[19] https://line.me/ja/

Acknowledgement

I would like to express my deepest gratitude to Professor NGUYEN, Minh Le, who has provided many supports and advice for my study and writing of this master's thesis. Without your guidance, I could not be able to complete my whole works. The meetings held at the Tokyo satellite office were excellent opportunities for me to advance my study. I would like to also express my gratitude to Professor Tojo, everyone in the laboratory, and Inada-san for their assistance during my visit to Ishikawa campus. Finally, I am profoundly thankful for my wife, Keiko and daughters, Ai and Mai who have understood my position as a graduate student and supported me throughout this period.

Bibliography

[1] 大向一輝, "電子情報通信学会　通信ソサイエティマガジン," *9 巻, 2 号,* pp. 70-75, 2015.

[2] 総務省, "令和 3 年通信利用動向調査報告書（世帯編）," p. 32, 令和 3 年.

[3] T. Kajiwara, C. Chu, N. Takemura, Y. Nakashima and H. Nagahara, "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*, 2021.

[4] H. Suzuki, Y. Miyauchi, K. Akiyama, T. Kajiwara, T. Ninomiya, N. Takemura, Y. Nakashima and H. Nagahara, "A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022.

[5] F. Toriumi, T. Sakaki and M. Yoshida, "Social emotions under the spread of COVID-19 using social media," *Trans. Jpn. Soc. Artif. Intell.,* vol. 35, no. 4, pp. F-K45_1-7, 2020.

[6] 高橋直樹, 檜垣泰彦, "Twitter における感情分析を用いた炎上の検出と分析," *信学技報,* 2017.

[7] 廣田壮一郎, 笹野遼平,高村大也,奥村学, "方言コーパス収集システムの構築," in *The 27th Annual Conference of the Japanese Society for Artificial Intelligence*, 2013.

[8] 山口真一, "ネット炎上の実態と政策的対応の考察," no. 11 号, 11 月 2025 年.

[9] 文部科学省, "「ネット上のいじめ」に関する 対応マニュアル・事例集（学校・教員向け）," 平成２０年.

[10] 文部科学省, "令和 3 年度 児童生徒の問題行動・不登校等生徒指導上の諸課題に関する調査結果の概要," 令和 4 年度.

[11] Ozawa, Seiichi; Yoshida, Shun; Kitazono, Jun; Sugawara, Takahiro; Haga, Tatsuya, "A sentiment polarity prediction model using transfer learning and

its application to SNS flaming event detection," *2016 IEEE Symposium Series on Computational Intelligence (SSCI),* pp. 1-7, 2016.

[12] Zhang, Jianwei; Otomo, Taiga; Li, Lin; Nakajima, Shinsuke, "Cyberbullying Detection on Twitter using Multiple Textual Features," *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST),* pp. 1-6, 2019.

[13] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

[14] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Lukasz; Polosukhin, Illia, "Attention Is All You Need," 2017.

[15] 東北大学自然言語処理研究グループ, "東北大学自然言語処理研究グループ," [Online]. Available: https://www.nlp.ecei.tohoku.ac.jp/research/open-resources/. [Accessed 5 2023].

[16] Salima Mdhaffar, Fethi Bougares, Yannick Esteve, Lamia Hadrich-Belguith, "Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments," vol. Proceedings of the Third Arabic Natural Language Processing Workshop, p. 55–61, 2017.

[17] Abdaoui, Amine; Berrimi, Mohamed; Oussalah, Mourad; Moussaoui, Abdelouahab, "DziriBERT: a Pre-trained Language Model for the Algerian Dialect," 2021.

[18] Antoun, Wissam; Baly, Fady; Hajj, Hazem, "AraBERT: Transformer-based Model for Arabic Language Understanding," 2020.

[19] Abdul-Mageed, Muhammad; Elmadany, Abdelrahim; Nagoudi, El Moatez Billah, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," 2020.

[20] Abdelali, Ahmed; Hassan, Sabit; Mubarak, Hamdy; Darwish, Kareem; Samih, Younes, "Pre-Training BERT on Arabic Tweets: Practical Considerations," 2021.

[21] Inoue, Go; Alhafni, Bashar; Baimukan, Nurpeiis; Bouamor, Houda; Habash, Nizar, "The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models," 2021.

[22] 久高優也, "琉日機械翻訳のための対訳コーパスの自動拡張について," *北陸先端科学技術大学院大学修士論文,* 2020.

[23] 柴田直由; 横山晶一; 井上雅史, "統計的手法を用いた双方向方言機械翻訳システム," *言語処理学会第 17 回年次大会発表論文集,* 2013.

[24] R. PLUTCHIK, "A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION," 1980.

[25] 宮内 裕人, 鈴木 陽也, 秋山 和輝, 梶原 智之, 二宮 崇, 武村 紀子, 中島 悠太, 長原 一, "主観と客観の感情極性分類のための日本語データセット," *言語処理学会第 28 回年次大会,* 2022.

[26] Bataa, Enkhbold; Wu, Joshua, "An Investigation of Transfer Learning-Based Sentiment Analysis in Japanese," 2019.

[27] Gururangan, Suchin; Marasović, Ana; Swayamdipta, Swabha; Lo, Kyle; Beltagy, Iz; Downey, Doug; Smith, Noah A, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* pp. 8342-8360, 2020.