Complaints with Target Scope Identification on Social Media

Kazuhiro Ito[1], Taichi Murayama[2], Shuntaro Yada[1], Shoko Wakamiya[1] and Eiji Aramaki[1]
[1]Nara Institute of Science and Technology, Nara, Japan
[2]SANKEN, Osaka University, Osaka, Japan
E-mail: ito.kazuhiro.ih4@is.naist.jp, s-yada@is.naist.jp, wakamiya@is.naist.jp, aramaki@is.naist.jp, taichi@sanken.osaka-u.ac.jp

Keywords: complaint, dataset, Twitter, social media, annotation

Received: March 22, 2023

A complaint is uttered when reality fails to meet one's expectations. Research on complaints, which contributes to our understanding of basic human behavior, has been conducted in the fields of psychology, linguistics, and marketing. Although several approaches have been implemented to the study of complaints, studies have yet focused on a target scope of complaints. Examination of a target scope of complaints is crusial because the functions of complaints, such as evocation of emotion, use of grammar, and intention, are different depending on the target scope. We first tackle the construction and release of a complaint dataset of 6,418 tweets by annotating Japanese texts collected from Twitter with labels of the target scope. Our dataset is available at `https://github.com/sociocom/JaGUCHI`*. We then benchmark the annotated dataset with several machine learning baselines and obtain the best performance of 90.4 F1-score in detecting whether a text was a complaint or not, and a micro-F1 score of 72.2 in identifying the target scope label. Finally, we conducted case studies using our model to demonstrate that identifying a target scope of complaints is useful for sociological analysis.*

Povzetek: Raziskava se osredotoča na analizo pritožb iz 6.418 tvitov z več metodami strojnega učenja.

1 Introduction

[1]A complaint is "*a basic speech act used to express a negative disagreement between reality and expectations for a state, product, organization, or event*" [23, p.195‐208]. An analysis of complaints contributes not only to linguistically [30] and psychologically [1, 18] interesting but also beneficial for marketing [17].

Understanding why people are dissatisfied can help improve their well-being by analyzing the situation of their complaints. The methods required to deal with complaints vary greatly depending on whether the target scope of complaints is him/herself, other people, or the environment (e.g., in the workplace, the way of improvement differs when employees are complaining about their own skills or about their work environment). The categorization presented above, regarding the target scope, aligns with James' three psychological categories for the Self as the object of reference [13]: the spiritual Self, the social Self, and the material Self, respectively.

In the field of natural language processing (NLP), there are some studies on how to determine whether a text is a complaint or not [26, 9, 14], or how to identify its severity [15], but no studies have been conducted yet to identify a target scope of complaints, which means the object toward which/whom the complaint is directed. Our study is an attempt to apply a computational approach focusing on a target scope of complaints on social media. More specifically, we emphasize the importance of identifying whether the complaints are intended for the complainer him/herself, for an individual, for a group, or for the surrounding environment.

This paper introduces a novel Japanese complaint dataset collected from Twitter that includes labels indicating the target scope of complaints [2]. We then investigated the validity of our dataset using two classification tasks: a binary classification task (shortly binary task) that identifies whether a text is a complaint or not, and a multiclass classification task (shortly multiclass task) that identifies the target scope of complaints. Furthermore, we apply our target scope classification model to case studies: COVID-19, office work, and the 2011 off the Pacific coast of Tohoku earthquake (we call Tohoku earthquake), aiming to analyze social phenomena.

Our contributions are as follows:

- We constructed a dataset of complaints extracted from Twitter labeled with the target scope of complaints.
- We conducted an experiment with identifying the target scope of complaints and achieved an F1 score of 90.4 in detecting whether a text is a complaint or not, and a micro-F1 score of 72.2 in identifying the target scope label.

[1] This paper is extended version of our study [12] presented in The 11th International Symposium on Information and Communication Technology (SOICT2022)

[2] Our dataset is available at `https://github.com/sociocom/JaGUCHI`

Table 1: Counts and examples of complaint tweets per target scope label in our dataset

Target Scope Label	# of Tweets	Example Tweet
SELF	468	しかしたぶん全部顔とか行動に出ちゃってるから最低なのは自分なんだよね向こうには落ち度はないし勝手に苛ついてるだけだしね (Maybe I'm the one who's the worst because it's all showing on my face and in my actions. It's not the other person's fault, I'm just irritated by myself.)
IND	3,866	わたしが居ないとミルクしまってある場所すらわかんないのかよ (You do not even know where the milk is stored without me?)
GRP	648	価値観の違いかもしれないけど物買うのは3千円でもしぶるのにギャンブルに平気で金突っ込むひとの気持ちがわからない (Maybe it's a difference in values, but I do not understand people who are reluctant to spend even 3,000 yen to buy something, but do not mind excessively spending money on gambling.)
ENV	1,436	保育士の給料上がらないかな～手取り15～18じゃやってけないよな (...) 政治家の給料とかより保育士に回してほしいわ、切実に (I wonder if childcare workers' salaries will go up. I can not make it on 15 to 18 take-home pay. (...) I'd really like to see more money spent on childcare workers than on politicians' salaries.)

– We conducted three case studies to demonstrate the usefulness of identifying a target scope of complaints for sociological analysis.

2 Related work

In pragmatics, a complaint is defined as "*a basic speech act used to express a negative disagreement between reality and expectations for a state, product, organization, or event*" [23, p.195-208]. What makes complaints different from negative sentiment polarity is that complaints tend to include expressions of the breaches of the speaker's expectations [26], and include reasons or explanations [31].

The dataset construction is actively conducted to analyse the substance of complaints. A previous study collected complaints about food products sent to governmental institutions and built an automatic classification model according to the nature of the complaint [9]. The classification classes were set up taking into account the use of customer support, the type of economic activity related, the priority of the treatment, and whether it is under the responsibility of the authority or not. Another study has created complaints dataset with labels for service categories (e.g., foods, cars, electronics, etc.) collected from reply posts to company accounts on Twitter [26]. Another study has also constructed a complaint dataset with four labels [15]: (1) No explicit reproach: there is no explicit mention of the cause and the complaint is not offensive, (2) Disapproval: express explicit negative emotions such as dissatisfaction, annoyance, dislike, and disapproval, (3) Accusation: asserts that someone did something reprehensible, and (4) Blame: assumes the complainee is responsible for the undesirable result. These four categories follow the definitions of the standard in pragmatics [29]. [7] has assigned the intensity of complaints as a continuous value using the best-worst scaling method [20] by crowdsourcing. Another corpus based on the data accumulated by *Fuman Kaitori Center* collects Japanese complaints about products and services [22]. The corpus includes labels about a target of complaints such as product or service names, which is different in granularity from our study.

As mentioned above, although some studies have constructed datasets that collect complaints, they have not yet constructed them that are labeled with a target scope to which complaints are directed.

3 Dataset

3.1 Collection

We constructed a Japanese complaint dataset using Twitter. For our dataset, we collected 64,313 tweets including "# 愚痴 (/gu-chi/)" (a hashtag of a Japanese term for complaints) from March 26, 2006 to September 30, 2021 using the Twitter API[3]. We excluded URLs, duplicates, and retweets, and extracted only those tweets with a relatively low possibility of being a bot. Specifically, we extracted only those tweets for which the posting application was *Twitter for iPad, Twitter for iPhone, Twitter Web App, Twitter Web Client, or Keitai Web*. All hashtags were removed from the text. Tweets with less than 30 characters were excluded. We extracted tweets for each month through a stratified sampling and finally obtained 7,573 tweets, which are

[3] https://developer.twitter.com/

of similar size with datasets recently released for NLP for social media [16, 24, 5, 3, 21].

3.2 Annotation

We annotated the 7,573 tweets with the target scope label. The tweets were divided into three sets (2,524, 2,524, and 2,525 tweets in each set), and three trained external annotators annotated each set.

First stage: Whether the tweet is a complaint or not is identified. Because most of the tweets are complaints owing to the inclusion of "# 愚痴", we remove tweets identified as non-complaints. Following Olshtain's definition [23, p.195-208], we identified tweets that expressed a negative disagreement between the tweeter's expectations and reality as complaints. Examples of non-complaints tweets removed by this process is shown below.
"If a company is violating the Labor Standards Act, gathering evidence is critical to remedy the situation."
"It's easy to complain, so I'm going to shift my thinking to the positive and creative."
"I came home exhausted again today. But I saw Mt. Fuji for a bit on the train on the way home, and it kind of loosened me up. I thought I was going to cry."

Second stage: We identify the target scope of complaints. We assigned one of four labels, SELF, IND, GRP, and ENV. Although our labels broadly follow James' theory of Self [13], we separate IND (individual) and GRP (group) because we believe that the nature of the complaints differs depending on whether the target is an individual or a group. In the case of individuals, it is associated with abuse, while in the case of groups, it is associated with hate speech. When the target scope was not determined uniquely or was unclear, it was removed from the dataset. We show definitions and examples of labels below.

`SELF`: A target scope includes the complainer.
e.g., "I have said too much again."

`IND`: A target scope does not include the complainer, which is one or several other persons.
e.g., "I hate that my boss puts me in charge of his work!"

`GRP`: A target scope does not include the complainer and has a group.
e.g., "I cannot be interested in people who only think about money."

`ENV`: A target scope is not human.
e.g., "It's raining today, so I do not feel like doing anything."

As a result of the annotation, among the 7,573 texts, 6,418 were considered as complaints. Among the complaint tweets, the number of labels per target scope is 468 for SELF, 3,866 for IND, 648 for GRP, and 1,436 for ENV. As a result, we collected 6,418 tweets. The agreement ratio (Kappa coefficient) between the annotators and an evaluator was measured to be 0.798 for the binary identification and 0.728 for the four-label classification. Agreement values are between the upper part of the substantial agreement [2]. Figure 1 presents the confusion matrix of human agreement on four classes normalized over the actual values (rows). Examples of text for each target scope label and number of tweets are shown in Table 1.

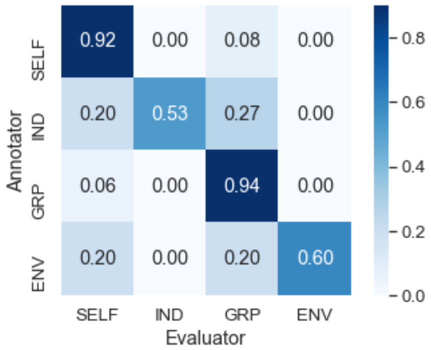


Figure 1: Confusion matrix of annotator agreement on four target scope of complaints.

Table 2: Statistics on the number of characters per label. The label with the highest mean number of characters in the texts is GRP, whereas the label with the lowest mean number of characters in the texts is SELF.

Target Scope Label	Mean	Median	Std
SELF	76.8	74.0	32.2
IND	83.2	83.0	32.4
GRP	87.8	89.0	32.5
ENV	77.8	74.0	33.8
ALL	82.0	81.0	32.8

3.3 Data analysis

We conducted two types of analysis for the contents of the dataset to gain linguistic insight into this task and the data: the number of characters and the emotions. The results of each analysis are shown below.

3.3.1 Number of characters

The average number of characters in the entire dataset is 82.0, and the median is 81.0. The label with the most characters is GRP (mean of 87.8 and median of 89.0), and the label with the fewest characters is SELF (mean of 76.8 and median of 74.0). This suggests that while descriptions of other groups tend to be detailed, those of him/herself have

Table 3: Results of emotion analysis using JIWC. We investigated the average score for each emotion per label. The highest results are in **bold**.

Target Scope Label	Sadness	Anxiety	Anger	Disgust	Trust	Surprise	Joy
SELF	**0.448**	**0.502**	0.774	0.858	**0.591**	0.467	0.459
IND	0.424	0.425	0.846	0.904	0.568	0.457	0.451
GRP	0.407	0.431	**0.861**	**0.954**	0.564	**0.477**	0.444
ENV	0.434	0.490	0.773	0.824	0.545	0.464	**0.482**
ALL	0.426	0.445	0.826	0.888	0.564	0.461	0.458

relatively not in detail. The statistics of the number of characters per label are shown in Table 2. Note that we removed tweets of less than 30 characters in Section 3.1.

3.3.2 Emotion

We examine the relationship between our dataset and emotions, and the differences in emotions between target scope. To do so, we used the Japanese Linguistic Inquiry and Word Count (JIWC) emotion dictionary [4]. This dictionary matches words with seven emotion categories (Joy, Sadness, Anger, Surprise, Trust, Anxiety, and Disgust) based on a translation of Pluchik's emotion wheel [25], obtained from a naturalistic dataset of emotional memories. The scores for each tweet (S_{ij}) were a ratio of the number of emotion terms in each category (W_{ij}), to the total number of terms (tokens; W_i^*) in each tweet:

$$S_{ij} = \frac{W_{ij}}{W_i^*} \log_2(W_{ij} + 1) \quad (1)$$

We used the scores from this emotion dictionary to calculate the emotion score for each tweet in our dataset and investigated the average score for each emotion per label. The results are shown in Table 3.

For SELF, the low value for Anger and high value for Anxiety are consistent with our intuition. When the complainer is him/herself, it can be interpreted that Anxiety is stronger than Anger. Disgust is higher for GRP than for IND. This indicates that feelings of Disgust are stronger for groups than individuals. In the case of Anger, both IND and GRP are high.

3.3.3 Topic

To investigate whether it is possible to extract the detailed contents of complaints in our dataset, we analyzed tweets' topics using the Latent Dirichlet Allocation (LDA), a kind of topic model [4]. The number of topics is set to 8, and LDA is applied only to nouns with two or more Japanese characters. Table 4 shows each topic and assigned words.

The following is an interpretation of the topics. Some of the topics are work-related (Topics 1, 3, 4, and 5), suggesting that work is the majority of complaints posted on Twitter. Among work-related topics, there were topics related to mental health (Topic 3), including "mood," "stress," and "hospital," and topics related to family (Topic 1), including "husband" and "children," which were divided into several tendencies. The other topic focused on COVID-19 (Topic 8), which includes "COVID-19" and "mask." Although only recent tweets are relevant to this topic, it is suggested that many such complaints had been posted intensively.

4 Experiment

4.1 Settings

In this section, we demonstrate the validity of the dataset using two types of classification tasks: a binary task (2-way) that identifies whether a text is a complaint and a multiclass task (4-way) that classifies the target scope of complaints. These tasks correspond to the first and second stages of annotation, respectively.

We employ two types of machine learning models: Long Short-Term Memory (LSTM) [11] and Bidirectional Encoder Representations from Transformers (BERT) [6]. The BERT model is a fine-tuned version of a model pretrained on the Japanese version of Wikipedia published by Tohoku University[5].

Before training, the dataset was preprocessed into lowercase, and all numbers were replaced with zeros. We split the dataset, into training, validation, and test sets (7:1.5:1.5). When we split the dataset the label distribution was maintained.

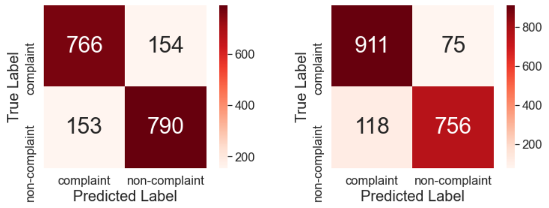
We set each parameter of the LSTM model as follows: the number of dimensions of the word embedding representation is 10, the number of dimensions of the hidden layer is 128, cross-entropy is used as the loss function, a Stochastic Gradient Descent (SGD) was applied as the optimization method, the learning rate is 0.01, and 100 epochs are used. We also set each parameter of the BERT model as follows: The maximum number of tokens per tweet is 128, the number of batches is 32, Adam is used as the optimization method, the learning rate is 1.0×10^{-5}, and 10 epochs are used. After examination of the validation data, we used the above parameters. Then, for the binary task, we added

[4] https://github.com/sociocom/JIWC-Dictionary

[5] https://github.com/cl-tohoku/bert-japanese

Table 4: The top 5 words per topic (translated from Japanese). Some of the topics are work-related (Topics 1, 3, 4, and 5), suggesting that work is the majority of complaints posted on Twitter. The other topic focused on COVID-19 (Topic 8), which includes "COVID-19" and "mask".

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
husband	child	human	company	really	why	without saying	angry
child	movie	workplace	husband	stupid	friend	adult	COVID-19
boss	parents' house	mood	world	vacation	everyday	money	cry
mood	block	stress	place	word	child	senior member	forbidden word
senior member	article	hospital	mother	company	meal	staff	mask



(a) LSTM model (b) BERT model

Figure 2: Confusion matrices of the binary task (2-way).

6,000 tweets to the dataset that were randomly sampled and removed complaints according to our annotation method.

4.2 Metrics

We report predictive performance of the binary task as the mean accuracy, macro-F1 score, and ROC AUC as well as existing complaints study [26]. On the other hand, we report predictive performance of the multiclass task as the micro-F1 score and macro-F1 score.

4.3 Results

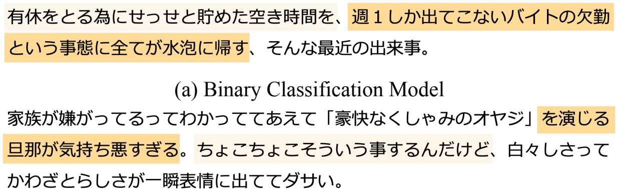
4.3.1 Binary task (2-way)

The results of the binary task reach an accuracy level of 83.5, an F1 score of 83.7, and an AUC of 83.5 for the LSTM model, and a level of accuracy of 89.6, an F1 score of 90.4, and an AUC of 89.4 for the BERT model (as shown in Table 5). The confusion matrix of the BERT model has a True Positive rate of 0.92, False Positive rate of 0.14, False Negative rate of 0.08, and True Negative rate of 0.86. For the BERT model, false negatives were reduced in number in comparison to the LSTM model. Figure 2 (a) and (b) show the confusion matrices for the LSTM and BERT models, respectively.

Table 5: Results of the binary and multiclass tasks. The BERT model outperformed Major Class and the LSTM model for each metric. The bold font indicates the best score for each evaluation metric.

Task	Metric	Major Class	LSTM	BERT
Binary	Accuracy	51.7	83.5	**89.6**
	F1 score	69.3	83.7	**90.4**
	AUC	50.0	83.5	**89.4**
Multiclass	micro-F1 score	62.1	51.7	**72.2**
	macro-F1 score	19.2	30.1	**54.5**

We are interested in what types of tokens our complaint model tries to capture. To interpret the behavior of the model, we used LIME [28], a method for explaining machine learning models, to create a visualization. We visualize the attention weights extracted from BERT model for the following example (translated from Japanese): "Recently, I had an encounter where all the free time I worked hard to make for a paid vacation was wasted because of the absence of a part-time worker who comes to work only once a week." We observed that the model paid attention to the expression "wasted because of the absence of a part-time worker who comes to work only once a week" for classification (as shown in Figure 3). In this example, the reason was the cause of the complaint, suggesting that our model pays attention to the same part as human intuition.


有休をとる為にせっせと貯めた空き時間を、週１しか出てこないバイトの欠勤という事態に全てが水泡に帰す、そんな最近の出来事。

(a) Binary Classification Model

家族が嫌がってるってわかっててあえて「豪快なくしゃみのオヤジ」を演じる旦那が気持ち悪すぎる。ちょこちょこそういう事するんだけど、白々しさってかわざとらしさが一瞬表情に出ててダサい。

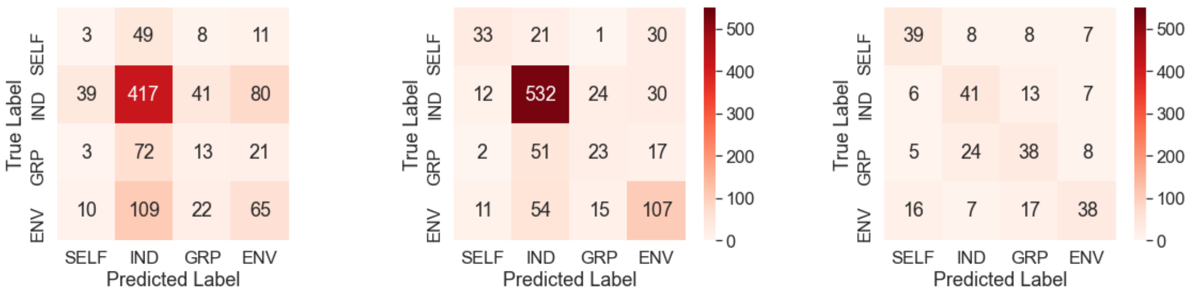
(b) Multi Classification Model

Figure 3: Visualization of the attention weights for the sample sentences in our binary (a) and multi (b) classification models. The orange line highlights the cue of classification. For (a), highlighted words are "wasted because of the absence of a part-time worker who comes to work only once a week." For (b), highlighted words are "The husband who plays the role of ... too disgusting."

4.3.2 Multiclass task (4-way)

The results of the multiclass classification task are a micro-F1 score of 51.7 for the LSTM model, and a micro-F1 score of 72.2 for the BERT model. Figure 4 (a) and (b) show the confusion matrices for the LSTM and BERT models, respectively.

In the LSTM model, a relatively large number of tweets are classified as either IND or ENV, reflecting the bias in the number of tweets in the dataset. Although the BERT model mitigates the effect of label bias in the dataset in



(a) LSTM model

(b) BERT model

(c) BERT with down sampling

Figure 4: Confusion matrices of the multiclass task (4-way). The LSTM model classified a relatively large number of tweets as IND or ENV. The results likely reflect the bias in the number of tweets in the dataset. The BERT model mitigates the effect of label bias in the dataset in comparison to the LSTM model. The BERT model with down sampling results show little bias among the labels.

Table 6: Examples of error cases in the binary task.

ID	Complaint Label		Tweet
	True	Predicted	
(1)	non-complaint	complaint	お仕事終わり! 定時で上がれたけど、フィットネスに行くかヤフオクの発送か....。明日は遅番だからジム行くのが得策。来週まで行けないし。(I finished the work! I was able to leave work on time, but I don't know if I should go to the fitness center or ship the Yahoo Auction... I have a late shift tomorrow, so going to the gym is in my best interest. I can't go to there until next week.)
(2)	non-complaint	complaint	何か作りたいなーという気分が出て来ただけマシかなーと思う昨今。風邪の熱に浮かされてるだけかもしれないが。フォトショ起動するのもめんどくさいモードだけど。うん。(I think it's better that I feel like making something these days. I may just be suffering from a fever from a cold. Although I'm too lazy to start up Photoshop right now.)
(3)	non-complaint	complaint	今日は寝坊して大変だったから早め（でももう 0 時;）に寝よう。お休みなさい! (I overslept and had a hard time today, so I'll go to bed early (but it's already midnight;). Good night!)
(4)	complaint	non-complaint	今、カラオケに行ってるらしい。職場にコロナ持ち込まないでねー!! 感染者出たら、あなたの責任ですから！ (Now they are going to karaoke, I heard. Don't bring coronavirus into the workplace! If anyone gets infected, it's your fault!)
(5)	complaint	non-complaint	感情豊かですねって、その状況、人に合わせて自分を作ってんだよ (People tell me I'm very emotional, but I make myself fit the situation and the people around me.)

comparison to the LSTM model, the accuracy per label shows that SELF tend to be misclassified as ENV. This reflects the fact that it is difficult to classify SELF and ENV because they have the common tendency to omit the target scope in statements about themselves. The accuracy of GRP is relatively low because when a complainer refers to a group that does not include him/herself, the complainer does not always use words that explicitly express that targets are multiple. In short, the LSTM model greatly outperformed the major class results in macro-F1, and the BERT model somewhat mitigated the bias in the number of labels that affected the LSTM classification results, further improving the macro-F1.

As well as binary task, we show the visualization of what types of tokens our complaint model tries to capture for the following example (translated from Japanese): "The husband who plays the role of "a man sneezing boldly" even though he knows his family doesn't like it is too disgusting. He does it occasionally, and it's so dull because it's so artificial and it shows on his face". This tweet was identified

Table 7: Examples of error cases in the multiclass classification task.

ID	Target Scope Label		Tweet
	True	Predicted	
(6)	SELF	IND	あー、でも休みの日とか、歩いてる時とか、ショッピングの時にアイディア浮かぶかも。もう、おっちゃんアイディア出ないから、もっと若い人に頑張って欲しいなぁ。(Maybe ideas happen when I'm on vacation, or walking, or shopping. As an old man, I can't come up with any more ideas so I wish more young people would try their best.)
(7)	SELF	ENV	頑張っても報われないし人間関係でいつもとん挫するしどうすりゃいいのかわかんないな、もう (I don't know what to do because my hard work is not rewarded and I always fail in personal relationships.)
(8)	GRP	IND	とある it 企業のデバッガーとして勤めてますが、今日だけは言わせてください。デバッガーを馬鹿にするな。(I work as a debugger for an IT company, and let me say this today. Don't mock debuggers.)
(9)	ENV	GRP	ニキビ死ねーーーーーーーーっっっ!!!!!!!! お前のせいでブスさ倍増すんだよクソ野郎!!!!!!!! (Pimples go away!!!!!!!!!!!!!!!! You make me look twice as ugly, damn you !!!!!!!!)

as IND by our model. The model paid the most attention to the words "The husband who plays the role of ... too disgusting" for classification (as shown in Figure 3). These words clearly illustrate the target of the complaint, "husband", and the feeling of "too disgusting" for that person, thus the cues to which the model assigned the labels are clearly interpretable to us.

4.4 Downsampling

Because the error in our multiclass task might be highly influenced by the unbalanced labels of the dataset, we experimented with a dataset with down sampling. We negatively sampled the number of data for labels other than SELF to approximately equal the number of labels for SELF, which has the fewest number of labels. For this experiment, we employ the BERT model and the settings are equal to Section 4.2. The result is a micro-F1 score of 55.3 and a macro-F1 score of 55.5. The results, as illustrated in Figure 4(c), indicate little bias among the labels. This result still shows a relatively high level of confusion between IND and GRP, suggesting that these pairs of labels tend to be similar languages. In addition, there were relatively many cases where ENV tweets were classified as SELF, suggesting that this error may be due to the omission of the target to which the complaint is directed (See Section 4.5).

4.5 Error analysis

4.5.1 Binary task (2-way)

Although the BERT model showed a high score of F1 score of 90.4, the model could not classify tweets correctly in some cases. The examples of error cases are shown in Table 6.

(1), (2), and (3) in Table 6 show the results of False Positive. In the example of (1), although the tweeter writes an expression that is not sure about the choice, it is labeled as NEGATIVE in the true data because It does not contain any negative emotions related to the complaint. In the example of (2), although the word "lazy", which is closely related to complaints, appear in the sentence, the expression "I think it's better" is the intent of the entire sentence. In the example of (3), the word "overslept" indicates an unfavorable situation, but the whole sentence is not a complaint because it is simply a tweet indicating the intention to go to bed early. In all of these cases, although negative elements are used in some parts of the tweets, the purpose of the tweet is other than just complaining. These tend to be False Positive.

On the other hand, in the case of (4) and (5) in Table 6, the results are False Negative. The example of (4), syntactically, it is a tweet indicating a kind of request to the target scope, but semantically it is a sentence accusing the target of going out to play. The tweet in (5), tweeter corrects an error in the target's perception and intends to express that he/she is feeling uncomfortable. As in these examples, there are often cases in which there is no explicitly complaint language or syntax in the tweets, but words appear that semantically imply a complaint.

4.5.2 Multiclass task (4-way)

We use the results of the BERT model with high accuracy to analyze error cases. The examples of error cases are shown in Table 7.

In many cases, the model predicts tweets as IND or ENV whose true labels are SELF. For example, in (6) in Table 7, there are two possible error factors: first, if the model focused on the sentence "I want more young people would try their best" and recognized "young people" as the target, it would be a false identification because the tweeter him/herself is the target scope for the purpose of the tweet. The second is that the tweeter, who is the true target scope, is paraphrased as “old man,” and thus this word is perceived as if he were a third party. Example (7) is a tweet that targets him/herself, which the model predicts as a label for ENV, since the scope of the tweet is not explicitly stated. Also, the model predicts tweets as IND or ENV whose true labels are GRP. In example of (8), although it can be inferred from the context that there is more than one person who is the target scope of the complaint, it is difficult to determine from the text whether the number is singular or plural, because there is no noun specified that indicates the target scope of the complaint. In example of (9), the use of the expression “go away” for a non-living target, commonly used to call out to a human, results in the incorrect identification of the target as a human being. Overall, the model tended to misclassify tweets that implied the target scope, which could only be inferred from extra-textual knowledge or the tone of the comments.

5 Case studies

We apply the constructed classification model of a target scope of complaints to tweets related to COVID-19, office work, and Tohoku earthquake to show that it is useful for sociological analysis.

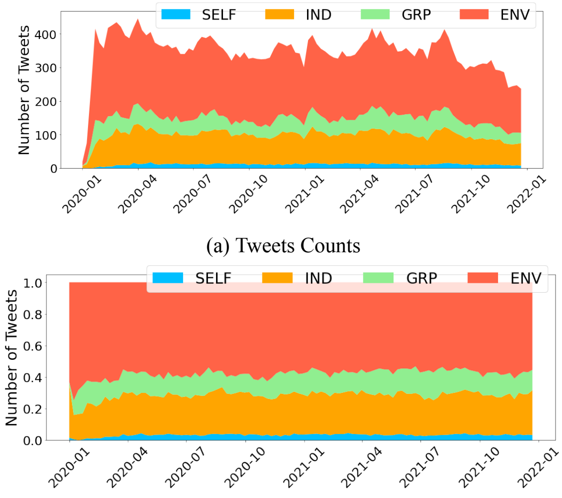
5.1 Case 1: COVID-19

We obtained 698,950 Japanese tweets including “コロナ (/ko-ro-na/)” which is a Japanese word for COVID-19 from January 1, 2020 to December 31, 2021 using the Twitter API.

The time series data presented in Figure 5 show that ENV accounted for a large ratio of cases during the early stages of the pandemic, and that this ratio decreased over time. In the tweets classified as IND or GRP, there were many complaints for others whose views on COVID-19 were different from those of the complainer, whereas in the tweets classified as ENV, there were many complaints for SARS-COV-2 and life during the pandemic. The examples of tweets labeled as each label is shown in Table 8.

In addition, To confirm our hypothesis that a content of complaints varies depending on a target scope, we analyzed the topics of the tweets using the Latent Dirichlet Allocation (LDA), a kind of topic model [4]. The number of topics is set to 16, and LDA is applied only to nouns and adjectives. Table 9 shows the five characteristic topics and five words extracted from the top 10 words per topic. The words that appear in topics about tweets labeled SELF include a number of adjectives such as “afraid,” “happy,” and “sad,” expressing their state of mind. IND is closely related to the tweeter’s personal relations, such as “girlfriend,” “family,” and “parents’ house.” Complaints about GRP tend to target public things, such as “government,” “politics,” “Olympics,” and “celebrity.” ENV frequently contains words related to the services of their customers, such as “lesson,” “movie,” “vaccine,” and “news.”

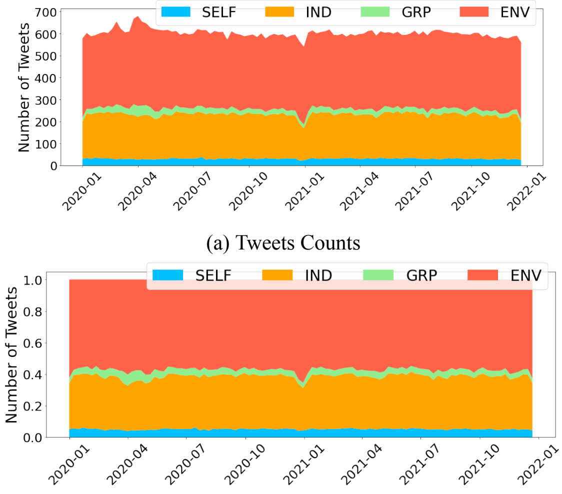
The differences in topics per label showed a certain interpretability, suggesting that automatic classification of a target scope of complaints at the granularity of our dataset also contributes to a categorization of the content of complaints.



(a) Tweets Counts

(b) The ratio of tweets labeled with each label

Figure 5: Time series data on the number of tweets per target scope of complaints related to COVID-19. ENV accounted for a large proportion of cases during the early stages of the pandemic, and this proportion decreased over time.



(a) Tweets Counts

(b) The ratio of tweets labeled with each label

Figure 6: Time series data on the number of tweets per target scope of complaints related to office work. There were few changes in the number of complaints per target scope over time.

Table 8: The examples of tweets related to COVID-19 labeled as each label

Target Scope Label	Tweet
IND	旦那ね、色んなところで営業回ってる人だからよく風邪ひいたり熱出たりすんの。手洗いうがいしてねって言ってもしねぇの。こいつのことこれからコロナさんって呼ぶことにした。(My husband is a salesman who goes around to various places so he often catches a cold or gets a fever. I tell him to wash his hands and gargle, but he doesn't. I've decided to call him Mr. COVID from now on.)
	一生、平行線なんでもういいんじゃないですか。あなたは、コロナは大したことないと思ってる、私は違う。これでいいですよ。(All along, it's failed to reach an agreement, so I think we're done. You think COVID-19 is no big deal, I don't. I'm fine with this.)
ENV	コロナが長引くと永遠に子供に会えなくなります子供はその環境に馴染んでしまうからうちは何とか line で繋げようとしてるけど、もう手遅れなんでそれは悲しいこと (If the situation with COVID-19 is prolonged, we won't be able to see our child forever ... We are trying to connect with them via LINE so that they don't get used to that environment, but it's too late now, and that's sad ... .)
	ホント疲れちゃったし、我慢してることも多いから辛いよコロナ禍じゃなきゃとっくに東京とかも行ってるし、何よりライブ出来てただろうしね (It's hard because I'm really tired and I have to endure so much ... . If it wasn't the situation with COVID-19, I would have been in Tokyo by now, and more importantly, I would have been able to go to live shows.)

Table 9: Five characteristic topics and five words extracted from the top 10 words per topic (translated from Japanese). SELF contains many adjectives such as "afraid," "happy," and "sad," expressing their state of mind. IND is closely related to the tweeter's personal relations, such as "girlfriend," "family," and "parents' house." Complaints about GRP tend to target public things, such as "government," "politics," "Olympics," and "celebrity." ENV frequently include words related to the service for which the tweeter is a customer, such as "lesson," "movie," "vaccine," and "news."

Target Scope Label	Words extracted from the top 10 words per topic				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
SELF	afraid	hobby	natural	meal	a lot
	happy	ruin	stress	really	complex
	painful	symptoms	dislike	word	surprised
	timing	vaccine	tough	patience	result
	sane	wedding	cheerful	sad	life
IND	part-time job	concert	mask	stupid	afraid
	stress	child	parents' house	money	really
	travel	hospital	test	family	you
	disturbed	aftereffect	fool	mother	friend
	promise	girlfriend	afraid	please	bad
GRP	treatment	covering up	Olympics	vaccine	player
	new type	doctor	report	young man	prejudice
	government	politics	afford	governor	train
	success	opinion	slander	criticism	citizen
	demonstration	civil servants	media	celebrity	trash
ENV	lesson	movie	vaccine	news	infection
	cancellation	ticket	afraid	money	pain
	postponement	gym	time	metropolis	universal
	hospitalization	patience	positivity	summer vacation	like
	return to country	really	insurance	dead	closing down

5.2 Case 2: office work

We obtained 731,000 Japanese tweets including a word "仕事 (/shi-go-to/)", which is related to office work from January 1, 2020 to December 31, 2021 using the Twitter API. Note that among the tweets collected in Case 2, 12,626 tweets overlapped with those collected in Case 1.

The time series data presented in Figure 6 show few changes in the ratio of complaints per target scope over time. This suggests that complaints regarding office work tended to be consistent regardless of the social situation. During the year-end and New Year's periods, the overall number of complaints tended to decrease, while the tweets classified as ENV did not decrease during this period.

As in Case 1, we analyzed the topics of the classified tweets in Case 2. Table 10 shows the five characteristic topics and five words extracted from the top 10 words per topic. The same tendency as in Case 1 was observed for all labels except ENV, with higher weights given to adjectives such as “nervous,” “anxious,” and “sad” for SELF, words indicating personal relations such as “boss,” “you,” and “husband” for IND, and words indicating public targets such as “idol,” “company,” and “voice actor” for GRP.

With regard to ENV, while in Case 1, words indicating services to which the tweeter is a customer appeared, in Case 2, words indicating workload or vacation were common, suggesting that the environment in which complaints target varies greatly depending on the domain.

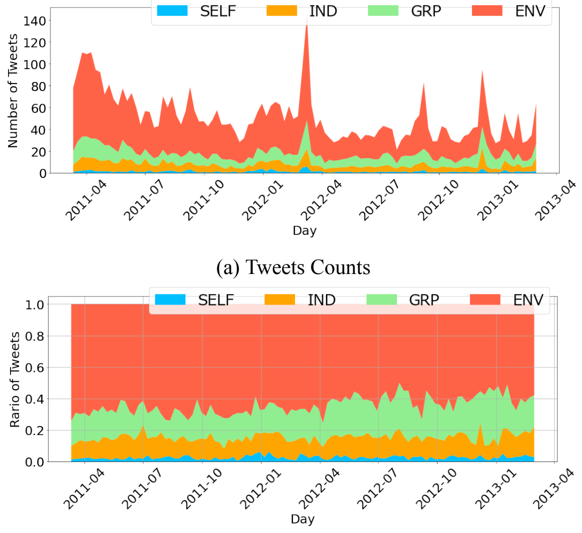
5.3 Case 3: Tohoku earthquake

In Case 1, the time series data show that complaints labeled as ENV accounted for a large proportion of cases during the early stages of the pandemic, but decreased over time, while complaints labeled as IND and GRP are flat over time. This tendency suggests our labels of the target scope of complaints caught phenomenon called “*a paradise built in hell*” [27]. This concept means that victims often exhibit altruistic behavior, engaging in voluntary mutual aid after a disaster. In the case of our classification model, we hypothesize that if the phenomenon of “*a paradise built in hell*” occurs, the ratio of complaints labeled as ENV is high in the early period after the disaster, while the ratio of complaints labeled as IND or GRP increases over time.

We obtained 106,732 Japanese tweets including “東日本大震災 (/hi-ga-shi-ni-ho-n-da-i-shi-n-sa-i/)” which is a Japanese word for Tohoku earthquake from March 11, 2011 to March 10, 2013 using the Twitter API. The time series data presented in Figure 7 show that complaints labeled as ENV accounted for a large ratio of cases during the early period after the disaster and that this ratio decreased over time. In contrast to the complaints labeled as ENV, the ratio of complaints labeled as GRP increased from one year after the disaster. These trends suggest that our classification model for the target scope of complaints can be used to detect the phenomenon of “*a paradise built in hell*” in Tohoku earthquake. The examples of tweets labeled as each label is shown in Table 11.

6 Conclusion & future work

We examined the use of computational linguistics and machine learning methods to analyze the complaints subjects. We introduced the first complaint dataset including labels that indicate a target scope of complaints. We then built BERT-based classification models that achieved F1 score of 90.4 for a binary classification task and micro-F1 score of 72.2 for a multiclass classification task, suggesting the validity of our dataset. Our dataset is available to the research community to foster further research on complaints. While we tried to adjust the unbalanced labels of the dataset by down sampling, it is also possible to adjust it by semi-supervised learning [19, 10] or data augmentation [8]. The validation of methods to improve model performance, including these methods, is our future work.



(a) Tweets Counts

(b) The ratio of tweets labeled with each label

Figure 7: Time series data on the number of tweets per target scope of complaints related to Tohoku earthquake. The complaints labeled as ENV accounted for a large proportion of cases during the early period after the disaster and this proportion decreased over time. In contrast to the complaints labeled as ENV, the ratio of complaints labeled as GRP increased from about one year after the disaster.

Furthermore, from the results of the case studies, we could show the possibility of applying the constructed models to perform sociological analysis. In case study, we applied our model to tweets extracted using queries related to COVID-19, office work, and Tohoku earthquake. In the case of COVID-19, we identified that the ratio of complaints targeting the surrounding environment decreases over time. We found that complaints targeting the surrounding environment and specific individuals were more frequent, with the former being complaints about “others whose views on COVID-19 differ from the tweeter” and the latter being complaints about “the COVID-19 virus and the environment in which infectious disease is spreading.” These results suggest most complaints can be divided into two categories: complaints that divide people and complaints generate empathy and cooperation. In the case of the 2011 off The Pacific Coast of Tohoku Earthquake, we showed the potential of our model to detect the phenomenon of “*a paradise built in hell*.” These viewpoints show the potential of our dataset as a starting point for sociological analysis.

We also experimented with a topic model for each target scope label as a case study using tweets about COVID-19 and office work, respectively. The distribution of words per topic confirms our hypothesis that the content of complaints varies greatly depending on the target scope. In addition, we observed that the complaints classified by our model as environmentally target scope varied greatly depending on the domain. In the future, as attempted through the case

Table 10: Five characteristic topics and five words extracted from the top 10 words per topic (translated from Japanese). Higher weights were given to adjectives such as “nervous,” “anxious,” and “sad” for SELF, words indicating personal relations such as “boss,” “you,” and “husband” for IND, words indicating public targets such as “idol,” “company,” and “voice actor” for GRP, and words indicating the day of the week, busy season, and vacation for ENV

Target Scope Label	Words extracted from the top 10 words per topic				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
SELF	nervous	human	like	sad	lonely
	overtime	really	motivation	depressed	busy
	hard	painful	anxious	dislike	difficult
	bothersome	stress	happiness	despair	weekend
	painful	get a job	patience	adult	beautiful
IND	boss	you	vacation	bath	meal
	every day	son	computer	senior member	husband
	me	plan	mistake	meal	work place
	information	absolutely	salary	tough	bath
	really	fool	husband	friend	time
GRP	idol	recruitment	voice actor	everybody	doctor
	type	salary	politics	tough	crime
	occupation	serious	interesting	professional	The Diet
	stupid	company	government official	on time	last train
	left-wing	woman	knowledge	understanding	really
ENV	tired	busy	vacation	good	game
	go to work	event	tough	a fun thing	tomorrow
	Monday	afraid	tired	refrain	weekend
	Friday	tough	study	end-of-year	happy
	everybody	reservation	nap	dull	sleep

Table 11: The examples of tweets related to Tohoku earthquake labeled as each label

Target Scope Label	Tweet
GRP	今、電車に乗っていますが、みんな暑い服着ていますね。だから、余計な電力が必要なのです。もうすぐ東日本大震災から2年。もう一度、見つめ直しましょう。あぁあ、電車の空調が入っちゃった。(I’m taking the train now, everyone is wearing hot clothes. So we need extra electric power. It will soon be two years since Tohoku earthquake. Let’s look back once again. Ahhh, the air conditioning is on in the train.)
	東日本大震災の被災に関して言えば、未だに復興どころか復旧すら出来ていない所もある。ましてや、福島県の一部県民は、ふるさとへ帰れないままです。選挙をしてる場合でしょうかねぇ。(As for the damage caused by Tohoku earthquake, there are still some areas that have not even been restored, let alone repaired. And some residents of Fukushima Prefecture are still unable to return to their hometowns. I wonder if it’s a matter of time to hold elections.)
ENV	勉強横目に東日本大震災のドキュメンタリー見てるけど、恐すぎる。これ今日寝れないやつだ。やっぱ1人恐い。。(I’m watching a documentary about Tohoku earthquake while studying, it’s too scary. I’m sure I won’t be able to sleep today. I’m afraid of being alone..)
	いつ災害がくるかわかりません。東日本大震災のとき、カセットボンベの買い置きがなくて困ったよ。(You never know when a disaster will happen. When Tohoku earthquake happened, I was in trouble because I didn’t have any cassette cylinders left over.)

study, we won’t only be able to identify a target scope of complaints in a text, but also be able to reveal potential social problems by investigating the temporal change of a target scope of complaints. Furthermore, the analysis results can be applied beyond social media platforms. For example, we are interested in investigating the relationship between workplace well-being and complaints by measuring the number of complaints and their target scope in the daily reports of a particular company. Such applications will be useful for achieving a comfortable life within society.

Acknowledgement

This work was supported by JST-Mirai Program Grant Number JPMJMI21J2, Japan.

References

[1] Mark Alicke et al. “Complaining Behavior in Social Interaction”. In: *Personality and Social Psychology*

Bulletin 18 (1992), pp. 286–295. DOI: 10.1177/0146167292183004.

[2] Ron Artstein and Massimo Poesio. "Survey Article: Inter-Coder Agreement for Computational Linguistics". In: *Computational Linguistics* 34.4 (2008), pp. 555–596. DOI: 10.1162/coli.07-034-R2.

[3] Tilman Beck et al. "Investigating label suggestions for opinion mining in German Covid-19 social media". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1–13. DOI: 10.18653/v1/2021.acl-long.1.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022. DOI: 10.5555/944919.944937.

[5] Yi-Ling Chung et al. "CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2819–2829. DOI: 10.18653/v1/P19-1271.

[6] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018). DOI: 10.48550/arXiv.1810.04805.

[7] Ming Fang et al. "Analyzing the Intensity of Complaints on Social Media". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1742–1754. DOI: 10.18653/v1/2022.findings-naacl.132.

[8] Steven Y. Feng et al. "A Survey of Data Augmentation Approaches for NLP". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 968–988. DOI: 10.18653/v1/2021.findings-acl.84.

[9] João Filgueiras et al. "Complaint Analysis and Classification for Economic and Food Safety". In: *Proceedings of the Second Workshop on Economics and Natural Language Processing*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 51–60. DOI: 10.18653/v1/D19-5107.

[10] Akash Gautam et al. "Semi-Supervised Iterative Approach for Domain-Specific Complaint Detection in Social Media". In: *Proceedings of the 3rd Workshop on e-Commerce and NLP*. Seattle, WA, USA: Association for Computational Linguistics, July 2020, pp. 46–53. DOI: 10.18653/v1/2020.ecnlp-1.7.

[11] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[12] Kazuhiro Ito et al. "Identifying A Target Scope of Complaints on Social Media". In: *Proceedings of the 11th International Symposium on Information and Communication Technology*. SoICT '22. Hanoi, Vietnam, 2022, pp. 111–118. DOI: 10.1145/3568562.3568659.

[13] William James. *The Principles of Psychology*. London, England: Dover Publications, 1890.

[14] Mali Jin and Nikolaos Aletras. "Complaint Identification in Social Media with Transformer Networks". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1765–1771. DOI: 10.18653/v1/2020.coling-main.157.

[15] Mali Jin and Nikolaos Aletras. "Modeling the Severity of Complaints in Social Media". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 2264–2274. DOI: 10.18653/v1/2021.naacl-main.180.

[16] Mali Jin et al. "Automatic Identification and Classification of Bragging in Social Media". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3945–3959. DOI: 10.18653/v1/2022.acl-long.273.

[17] Chul-min Kim et al. "The effect of attitude and perception on consumer complaint intentions". In: *Journal of Consumer Marketing* 20 (2003), pp. 352–371. DOI: 10.1108/07363760310483702.

[18] Robin M. Kowalski. "Complaints and complaining: functions, antecedents, and consequences." In: *Psychological bulletin* 119 2 (1996), pp. 179–96. DOI: 10.1037/0033-2909.119.2.179.

[19] Dong-Hyun Lee. "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks". In: 2013.

[20] Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press, 2015. DOI: 10.1017/cbo9781107337855.

[21] Julia Mendelsohn, Ceren Budak, and David Jurgens. "Modeling Framing in Immigration Discourse on Social Media". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 2219–2263. DOI: `10.18653/v1/2021.naacl-main.179`.

[22] Kensuke Mitsuzawa et al. "FKC Corpus : a Japanese Corpus from New Opinion Survey Service". In: *In proceedings of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results*. Portorož, Slovenia, 2016, pp. 11–18.

[23] Elite Olshtain and Liora Weinbach. "10. Complaints: A study of speech act behavior among native and non-native speakers of Hebrew". In: 1987. DOI: `10.1075/pbcs.5.15ols`.

[24] Silviu Oprea and Walid Magdy. "iSarcasm: A Dataset of Intended Sarcasm". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 1279–1289. DOI: `10.18653/v1/2020.acl-main.118`.

[25] Robert Plutchik. "Chapter 1 - A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION". In: *Theories of Emotion*. Ed. by Robert Plutchik and Henry Kellerman. Academic Press, 1980, pp. 3–33. DOI: `10.1016/B978-0-12-558701-3.50007-7`.

[26] Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. "Automatically Identifying Complaints in Social Media". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5008–5019. DOI: `10.18653/v1/P19-1495`.

[27] Solnit Rebecca. *A paradise built in hell: The extraordinary communities disaster*. Penguin, 2010.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. DOI: `10.1145/2939672.2939778`.

[29] Anna Trosborg. *Interlanguage Pragmatics: Requests, Complaints, and Apologies*. De Gruyter Mouton, 2011. DOI: `doi:10.1515/9783110885286`.

[30] Camilla Vásquez. "Complaints online: The case of TripAdvisor". In: *Journal of Pragmatics* 43.6 (2011). Postcolonial pragmatics, pp. 1707–1717. DOI: `10.1016/j.pragma.2010.11.007`.

[31] Guangyu Zhou and Kavita Ganesan. "Linguistic Understanding of Complaints and Praises in User Reviews". In: Jan. 2016, pp. 109–114. DOI: `10.18653/v1/W16-0418`.