Implicit and Indirect: Computational Identification of Ambiguous Conversational Actions in Asynchronous Crisis-Related Conversations

Anonymous ACL submission

Abstract

This paper presents a digital conversation analysis based approach to the computational detection of ambiguous actions in asynchronous online conversations. Action detection has been widely studied for synchronous chats. However, models or datasets for asynchronous conversations are scarce, and have not sufficiently considered the special characteristics of asynchronous discussion, most importantly the tendency for comments to involve multiple actions and multiple valid interpretations of actions. We provide a theory-driven annotation scheme for crisis-related asynchronous conversations, and an annotated dataset for Finnish. We show that considering the multi-action characteristics of asynchronous data statistically improves classification performance, and that an ensemble of best models can represent the ambiguity of actions, which is especially characteristic of face-threatening actions in controversial conversations.

1 Introduction

Natural Language Processing (NLP) and Machine Learning (ML) methods are becoming increasingly popular for analyzing textual content on social media, e.g. discourse signals (Ferracane et al., 2021; Zhang et al., 2017). Recent studies have shown that structural or turn-by-turn analysis of online conversations can allow rich automated linguistic analyses of interaction (Zhang et al., 2018; Sudhahar et al., 2015; Twitchell and Nunamaker, 2004), which are crucial for identifying misbehaviors like manipulative trolling (Paakki et al., 2023), or antisocial behavior (Zhang et al., 2018; Garimella et al., 2018). Asynchronous forum conversations can have wide impacts on public opinion, being persistent online and thus reaching large audiences (Zhang et al., 2018). Especially crises make online discourses vulnerable to trolling, manipulation and disinformation (Di Mascio et al., 2021). Thus, there is need for enhanced computational methods for analyzing controversial crisis conversations from the perspective of action-taking (Paakki et al., 2023; Zhang et al., 2018).

However, although synchronous online conversations have been studied extensively (Clark and Popescu-Belis, 2004; Forsyth and Martell, 2007; Fuscone et al., 2020; Stolcke et al., 2000), there are few models or datasets for analyzing action-taking in asynchronous arenas. Existing resources do not fully consider the specific characteristics of asynchronous conversation, including messages' potential for multiple meanings, or their tendency to include more than one action. Further, they do not include some actions important for analyzing controversial conversations (e.g., *accusations*) (Paakki et al., 2023). In this paper, the concept of *action* refers to what functions a turn has in conversation, i.e. what it does in relation to other turns. Our most important overarching goal is to investigate how to best computationally analyze actions in controversial asynchronous conversations, considering their special contextual characteristics. We argue that considering more than one action per comment and multiple possible interpretations of actions will lead to better performance and model fit with the empirical phenomenon. To this end, we have two research questions:

- RQ1. Does considering more than one prevalent action in a comment lead to better classification performance in contrast to selecting only one likeliest action?
- RQ2. What approach leveraging annotator disagreements best reflects multiple valid interpretations relevant to asynchronous data?

It is important to consider the special characteristics of asynchronous conversations, because they notably differ from both face-to-face and synchronous chats (Virtanen et al., 2021; Xiao et al.,

Comments	Actions
A: There are two powerful presidential candidates in the US; One has done, already years ago, powerful deeds together with God, such which many Presidents in the States have not dared to do, but we found a brave man respecting the Father's will, Donald Trump. (8 laugh emojis, 2 likes)	*statement* *appreciation*
B: A, hallelujah! (1 laugh emoji, 1 like)	*statement* OR *appreciation*

Table 1: Extract from Ukraine war discussion, under YLE Facebook page news 2022.

2020). For example, as seen in Table 1, many comments have a tendency to include more than one prevalent action, with potential for multiple interpretations, e.g. due to semantic ambiguity (Virtanen et al., 2021; Paakki et al., 2021; Stommel and Koole, 2010; Herring, 1999). B's turn in Table 1, given the context, is likely a sarcastic *statement* but could be interpreted as a genuine *appreciation*. On the other hand, at least two actions overlap in A's turn. These considerations are relevant to analyzing crisis conversations, as manipulation or trolling of such political or societal discussions tend to frequently involve covert, ambiguous or indirect strategies of influence (Paakki et al., 2020).

We utilize a digital conversation analysis (CA) based theoretical framework to build an annotation scheme (Clark and Schaefer, 1989; Enfield et al., 2010; Herring et al., 2005) for identifying actions in asynchronous conversations in Finnish, in a low-resource setting. The unit of analysis is one comment in a conversation thread. We employ a 7-point Likert scale annotation format, and compare models assuming only one action and one interpretation per comment *vs.* models considering several actions and different approaches to leveraging annotator disagreements. We show that considering multiple actions and valid interpretations for each comment allows higher model performance. We show that an ensemble model consisting of 2-3 individual annotator based models (or an averaged model) can best represent the ambiguity of actions in our data. We make our annotation guidelines and annotated dataset in Finnish available through an application process, to protect data privacy[1].

[1]Detailed annotation guidelines and models, to the extent that does not compromise any individual's privacy, will be provided on our GitHub upon paper publication.

2 Digital CA Based Analysis of Actions

(Digital) CA has potential for computational operationalization due to its tendency to pay attention to distribution and generalizable features of interaction (Stivers, 2015). CA interpretations of actions arise from what a turn does in a conversation, based on the utterance itself and the next turns – how other turns relate to the utterance and interpret its role (Sacks et al., 1974). What differentiates CA based understanding of actions from, e.g., speech acts, is that interpretations are based on the next-turn-proof procedure rather than judging the intent behind a turn in conversation (Sacks et al., 1974). Digital CA is of interest here, as it has potential for analyzing the dynamics between actions in interrelated turns, the expected responses to relevant actions like accusations, as well as their ambiguity or indirectness.

CA is well suited for analyzing online interactions (Giles et al., 2015; Meredith and Stokoe, 2014), as people treat actions and their norms online similarly to face-to-face conversation (Meredith, 2017; Paakki et al., 2021). Here our interest centers around most common actions in online discussion (Herring et al., 2005), central in analyzing sequential organization (Clark and Schaefer, 1989; Schegloff, 2007) and adherence to conversational norms (Paakki et al., 2021).

Digital CA research stresses the need to consider the specific characteristics of different types of online interaction (Virtanen et al., 2021; Meredith, 2017), as face-to-face or synchronous chats differ from asynchronous interactions (Xiao et al., 2020), where participants e.g. tend to commit several actions in one message (Paakki et al., 2021; Virtanen et al., 2021). However, most computational modeling of actions relates to customer chat bots (Casanueva et al., 2020; Ghosh and Ghosh, 2021), telephone conversations (Godfrey et al., 1992; Fuscone et al., 2020), recorded face-to-face dialogue (Clark and Popescu-Belis, 2004) and synchronous chats (Forsyth and Martell, 2007; Moldovan et al., 2011). The above studies represent a context different from casual, anonymous, and asynchronous online conversation (Herring, 1999), for which there are much fewer models (Bracewell et al., 2013; Zhang et al., 2017; Joty and Hoque, 2016). Also, existing resources for asynchronous data do not include actions like *accusations* or *challenges* relevant to analyzing harmful or manipulative online behaviors (Paakki et al., 2021, 2023). The char-

acteristics of these actions have been well established in CA research (Dersley and Wootton, 2000; Koshik, 2003; Turowetz and Maynard, 2010).

3 Multiple Interpretations of Actions

Recent studies show that for many NLP tasks there is no single ground truth (Jiang and de Marneffe, 2022; Plank, 2022; Uma et al., 2022), due to uncertainty in text meaning, leading to different interpretations of label distribution (Jiang and de Marneffe, 2022), constituting meaningful systematic disagreement (Jiang and de Marneffe, 2022; Nie et al., 2020). Thus, relying on a single ground truth ignores the possibility of multiple valid interpretations. Multilabel models offer more expressive results (Jiang and de Marneffe, 2022), and including disagreement into models can improve performance (Passonneau et al., 2012; Plank, 2022). Multiple interpretations are also relevant to identifying actions (Passonneau et al., 2012; Thomas, 1995).

Most existing action detection models rely on one ground truth (e.g. Zhang et al., 2017). An exception is Ferracane et al.'s (2021) study, which incorporated multiple interpretations into action modeling, aiming to classify all valid interpretations. Another study by Taniguchi et al. (2020) predicted both utterance-level and message-level interpretations of actions. However, the former used live congressional hearings and the latter emails as data. These approaches thus differ from crisis-related asynchronous forum conversations, which involve ambiguous use of actions in medium length texts, with frequent use of controversial actions (e.g. *accusations*). Thus, we investigate how to represent the ambiguity (Jiang and de Marneffe, 2022; Uma et al., 2022) of actions on such arenas.

4 Data

To answer our RQs, we collected asynchronous conversation data related to reader comments on crisis news about the COVID-19 Pandemic and Ukraine war. We manually annotated the comments using a digital CA based framework, which enables the comparison of different identification models. Our data comes from public Facebook (FB) pages owned by two Finnish news media, YLE and Helsingin Sanomat (HS)[2]. Our interests beyond this paper relate to computationally analyzing trolling and manipulation in crisis conversations, so we collected data including controversial actions central to trolling (Paakki et al., 2021).

We used Facepager v.4.5.3 (Jünger and Keyling, 2019) (MIT License)[3] to scrape FB posts in the two pages' feed, and their threaded comments, between 1 Dec. 2019–10 Feb. 2023. All posts included a news title, description and link to a piece of news. This resulted in the raw non-annotated dataset in Table 2. To select a subset of random conversations for manual annotation, we shuffled the non-annotated dataset per news posts to keep the comments in the same conversation together. We divided it into three parts, one for each annotator (the authors)[4]. Each annotator manually collected comments for annotation, selecting the first 400 crisis-related comments from their sample. This meant that we had to manually read the titles and descriptions of approx. 100-150 news posts to find comment sections on COVID or Ukraine war.

We collected both comments and their replies in comment section threads if the conversations fit our inclusion criteria. They had to be related to Ukraine war or COVID-19, commenting allowed, and including at least one comment with two or more replies, as we were interested in conversational interaction. Due to the restrictions of Facepager, we had to retrieve some missing comments manually. We excluded examples we had already seen during annotation scheme development (see section 5). To achieve greater variation in comment topics, we included max. 30 comments from the same section. We finally annotated all comments and replies following our guidelines (see section 5). The resulting datasets are described in Table 2. The number of comments to a news post (when comments were allowed) ranged from 14 to 684, with mean 90.56 and median 84.5, and comment mean length 151.4 characters, median 105.0.

To enable comparison of how models trained on annotations by one annotator perform against data where each example has several annotations by different annotators, we decided to produce two versions of the to-be-annotated data: the **single annotations dataset**, where each comment will have one annotation and the **multiple annotations dataset**, where all comments will have three annotations. In the latter, each annotator will annotate all 1,200 comments. This will enable us to study

[2]These are among most followed news outlets in Finland, YLE being the national public broadcasting company, and Helsingin Sanomat Finland's largest subscription newspaper.

[3]https://github.com/strohne/Facepager

[4]Annotators' demographic information will be reported after reviews not to risk anonymity.

Dataset	Comments	Annot.	News
Single annotations	1,200 675 Covid 529 War	1,200	46 26 HS 20 YLE
Multiple annotations	1,200 675 Covid 529 War	3,600	46 26 HS 20 YLE

Table 2: Dataset descriptions. Annot.= Annotations. An annotation is one set of 8 scores for each action.

how to best represent multiple interpretations of actions (RQ2).

5 Annotation Methods

We base our annotation scheme on typologies of actions in (digital) CA and Computer-Mediated Communication (CMC) literature (Clark and Schaefer, 1989; Enfield et al., 2010; Herring et al., 2005; Paakki et al., 2021; Schegloff, 2007; Stivers, 2015). The final scheme (Table 3) resulted from empirical insights during our incremental development of the scheme and annotation guidelines. It includes key actions for asynchronous conversation (Herring et al., 2005; Paakki et al., 2021), representing most central rhetorical and interactive functions of comments. We include both responsive actions and ones initiating a paired action, which expect specific responses (e.g. *denial* to *accusation*).

Due to the excess of possible actions in CA and CMC (Schegloff, 2007), we aimed at a simplified scheme as very fine-grained tag sets like DAMSL (Allen and Core, 1997) can suffer from sparseness and complexity reducing annotator agreement (Savy, 2010). We wished to limit actions only to ones observed in our data. This risks oversimplification of our theoretical framework (Stivers, 2015), so we relied on theoretical support and data-driven insights. Originally we considered 15 actions (*rejection*, *admission*, *announcement*, *answer to question*, *evaluation*, *proposal* in addition to ones in Table 3), reducing it to 8.

We aimed to reduce the effects of underspecification of guidelines (Aroyo and Welty, 2015) by iteratively developing the scheme and guidelines, to reduce unnecessary disagreement. We developed our scheme before data scraping: annotating, negotiating, and analyzing practice data based on the guidelines. This involved 19 iterations with specification of guidelines, and adaptation of the scheme. The data used in the development was manually

Action	Description	%	r_{WG}
Question *(initiating)*	Asks smdb for information.	0.94	0.88
Request *(initiating)*	Requests, proposes or tells smbd to commit an action.	0.93	0.90
Statement *(initiating/ responding)*	Asserts an opinion, information, wish, neutral/negative evaluation.	0.59	0.58
Challenge *(initiating)*	Refutes epistemic claims made by an actor.	0.91	0.65
Accusation *(initiating)*	Points out a reprehensible act committed by an actor.	0.86	0.74
Appreciation *(initiating/ responding)*	Positive evaluation or comment about actor, event/object.	0.97	0.94
Acceptance *(responding)*	Agrees or accepts a comment (e.g. request) or admits an accusation.	0.97	0.89
Denial *(responding)*	Rejects or denies an action (e.g. accusation).	0.95	0.87

Table 3: Action annotation scheme and inter-annotator agreements (% agreement and r_{WG}) with separate test set (N=100).

selected as screen captures and links from approx. first 300 crisis news posts in YLE and HS FB pages, between August–December 2022. This data was not included in the final annotated datasets. During annotation, we always read the conversations in their original FB context.

We stopped development when reaching sufficient agreement, and good scheme applicability for our data. See Table 5 for label distributions for final annotated datasets.

Our task required expertise, training and familiarity with our theoretical premises, so expert annotation was chosen. Crowdsourcing was not used as non-expert annotation involves concerns related to reliability, misinterpretation or misuse of labels (Duran et al., 2022), shown to be ineffective when analysis requires consideration of context, in-depth reading and domain expertise (Eickhoff, 2018; Rezapour et al., 2020).

We decided to use 7-point Likert scale scores (0: action not present – 3: maybe or partly present – 6: action very strongly present) for annotation (Peterson et al., 2019). This was due to many com-

Nro actions	Count (mean)	Percentage
0	3	0.25%
1	551.33	45.79%
2	443.33	36.82%
3	161	13.37%
4	39.33	3.27%
> 4	6	0.50%

Table 4: Number of labels assigned on average to single messages in multiple annotations data.

Action	Count (s)	Count (m)
Question	243	276
Request	130	186
Statement	857	999
Challenge	138	381
Accusation	155	305
Appreciation	36	66
Acceptance	113	171
Denial	97	151

Table 5: Distribution of annotated actions in the final ground truth datasets. (s = single annotations dataset, m = multiple annotations dataset)

ments including more than one prevalent action, and we judged that forcing annotators to label only one would reduce annotation quality (see Table 4). Confidence scores are useful in achieving improved inter-rater agreement, and highlighting difficult cases (Weber et al., 2018; Troiano et al., 2021). Scores were coded for each action separately per comment, allowing multiple interpretations of labels (Barnhurst and Mutz, 1997), signaling confidence through the same score.

We measured annotation quality comparing observed and expected variances of scores between annotators. For this purpose, r_{WG} score has been deemed helpful for evaluating score-based agreement within a group (Castro, 2002): $r_{WG} = 1-$(Observed Group Variance/ Expected Random Variance) (Lindell and Brandt, 1999). We used r_{WG} score for each action (Table 3) and $R^*_{WG(J)}$ score for overall agreement (Lindell and Brandt, 1999; O'Neill, 2017). The first score is for single-item scale and the second for multi-item scale, measuring the variance between annotations when random variance is eliminated. $R^*_{WG(J)}$ score was 0.72. Agreement strength boundaries are defined for the r_{WG} measurement family: our $R^*_{WG(J)}$ score is over the lower bound of strong agreement $(0.71 - 0.90)$ (O'Neill, 2017).

6 ML Models

Recent action identification models rely on sentence transformers, other neural networks (Ghosh and Ghosh, 2021), or few-shot learning (Casanueva et al., 2020). Many studies emphasize the relevance of linguistic lexical and collocational features (Stolcke et al., 2000; Ferracane et al., 2021; Zakharov et al., 2021). Thus, we chose a similar approach, allowing comparability to earlier work. As separate classifiers for sub-tasks have proven effective (Ferracane et al., 2021; Zakharov et al., 2021), we used a separate model for each action following Ferracane et al. (2021), to predict whether a text contains an action or not.

We used BERT, having proved its capacity in text classification (Devlin et al., 2019; Arabadzhieva-Kalcheva and Kovachev, 2022), finetuning the Finnish pretrained FinBERT (Virtanen et al., 2019). Our data is highly imbalanced, with less data mostly in the positive class, so we used class weighting in training. Besides standard fine-tuning we used Setfit, based on fine-tuning a pre-trained model with sentence pairs, then training a classifier based on fine-tuned embeddings (Tunstall et al., 2022). SetFit creates more variety to the minority category training samples with sentence pairing, generating in total $k(k-1)/2$ different pairs from training data, k being the size of training set. With FinBERT (Virtanen et al., 2019) as a base for SetFit, both models used a pretrained Finnish language model to provide embeddings for comments. We used Optuna for hyperparameter optimization[5] (Akiba et al., 2019). We used 240 GPU hours.

To allow comparison of our results to earlier research, we included SVM – widely used in action modeling – similarly to previous studies on asynchronous data (Cohen et al., 2004; Zhang et al., 2017). We used 1-grams and TF-IDF for feature extraction, with SVD for dimensionality reduction, balanced class weighting, and Grid Search for hyperparameter optimization, utilizing sci-kit learn (Pedregosa et al., 2011). We applied preprocessing (tokenization and lemmatization with spaCy (Haverinen et al., 2014)), leaving in stop words as their removal negatively impacted performance.

We divided our data into train, validation (for hyperparameter tuning) and test sets using sci-kit

[5]batch size 14, 4 epochs, 7 iterations, learning rate 3.0191843531454982e-05.

learn train test split twice, with respective set sizes 60%, 20% and 20%. Model evaluation included accuracies and macro-F1 scores – we report the latter due to class imbalances.

To build a set of ground-truth labels, we mapped annotation scores back to binary labels. For single annotations, the label was 1 if the annotation score was ≥ 3 . For multiple annotations, we use *conservative* (if even one annotator had given a score ≥ 3, the label was 1), and *relaxed ground truth labels* (if at least two annotators had given a score ≥ 3, the label was 1).

To answer RQ1, we trained FinBERT, SetFit and SVM classifiers considering only one action vs. multiple actions. We compared three approaches: a.) a *single action model (1-act)*, b.) *multilabel single-annotation model (MS)*, and c.) *averaged multilabel multiannotation model (Avg.)*.

The 1-act model used single annotations, setting for each comment the action with highest score among all labels as positive if score ≥ 3, other actions as negative. In cases where two actions (or more) had equal scores, we assigned a positive label randomly between them. In our view, this corresponds to annotator decision-making [6]. We will compare other models to the 1-act model to test whether considering more than one action will statistically improve performance. All other models allow multiple actions labeled.

The MS model used single annotations, allowing multiple actions labeled. The averaged model (Avg.) utilized multiple annotations to decide an average annotation score for label decision for training data, $score_{action} = (a_1 + a_2 + a_3)/3$, with a a reference to $AnnotationScore_i$ for the action. For statistical tests, we used the Nemenyi test, utilizing the scikit-posthocs implementation for python (Terpilowski, 2019), recommended for comparing classifier performances (Derrac et al., 2011).

To answer RQ2, using SetFit, our best model, we compared three approaches to leveraging annotator disagreements to find the best approach for representing multiple interpretations of actions: the a.) *averaged model (Avg.)* (Uma et al., 2022), b.) *positive/negative/complicated model (PNC)* (Jiang and de Marneffe, 2022), and c.) *individual annotator models (Annotator1-3)* (Ferracane et al., 2021).

Averaging was included as it often been used successfully with multiple annotations (Uma et al., 2022), as was the case with the PNC model (Jiang and de Marneffe, 2022). Following Jiang and de Marneffe (2022), we divided comments into three classes by counting an average cross-entropy score between annotations: two classes where annotators agreed they belong (here positive/negative), and a class where there was significant disagreement (here complicated). We included a comment in the "complicated" class if $C_i > \frac{\sum_i(C)}{N} + 2SD(C)$, C a reference to average cross-entropy[7]. We treated the task as a 3-way classification problem. Also individual annotator models have been fruitful for representing multiple interpretations, by predicting all different annotations (Davani et al., 2022; Ferracane et al., 2021). These were trained with only each specific annotator's annotations.

We compared how different ensembles of best performing models fared in predicting all possible labels, using Jaccard Coefficient to compare each ensemble's predicted set of labels to a set of all possible annotations by annotators for each text: $JaccardSimilarityJ(A, B) = \frac{|A \cap B|}{|A \cup B|}$.

7 Results

We will next discuss our results to answer our RQs: whether considering more than one prevalent action in comments in asynchronous conversation affects model performance (RQ1), and what approach to leveraging annotator disagreements best matches the multiple valid interpretations relevant to asynchronous data (RQ3).

7.1 RQ1: One Action or Multiple Actions

First, we compared 1-act models to multi-label models, in Table 6, to discover whether predicting more than one action would improve performance.

Considering multiple actions vs. one, SVM achieves statistically improved performances for only two classes at best. With FinBERT, Avg. models achieve notably higher macro-F1 scores for many actions, although the results are statistically significant for only one action. SetFit fares much better: MS models achieve statistically higher performance in classifying three action classes. Despite higher macro-F1s for the other actions, there

[6] As in the first rounds of the annotation process, starting with single action labeling, we most often had to randomly choose the primary action if having to decide only one most important action between actions with the same action strength.

[7] This decision boundary differs from Jiang and de Marneffe 2022, but as they had data with a hundred crowdsourced annotations per example, and we only three, we could not use a similar method. We judged that an outlier boundary of $mean(C) + 2SD$ would be strict and similar enough.

Model	Feature	Data	Action							
			question	request	statement	accusation	challenge	acceptance	denial	appreciation
SVM	1-act	single	0.60	0.53	0.62	0.52	0.52	0.55	0.53	0.56
	MS	single	0.66	0.58	0.63	0.62*	0.52	0.69	0.59	0.51
	Avg.	multiple	0.63	0.60	0.62	0.62*	0.53	0.69**	0.59	0.51
FinBERT	1-act	single	0.86	0.63	0.70	0.57	0.54	0.65	0.57	0.62
	MS	single	0.87	0.69	0.73	**0.76**	0.61	0.66	0.56	0.65
	Avg.	multiple	0.85	0.70	0.77**	0.63	**0.68**	0.62	0.51	0.75
SetFit	1-act	single	0.81	0.60	0.66	0.52	0.50	0.59	0.55	0.50
	MS	single	0.94	**0.82***	0.77	0.73*	0.64	**0.86***	**0.68**	**0.78**
	Avg.	multiple	**0.97***	0.80**	**0.78****	0.65*	0.63	0.79**	0.58	0.72

Table 6: 10-fold cross-validated macro-F1 scores for models comparing single vs. multiple annotations data. Statistically significant differences are indicated with $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Model	Ground truth	Action							
		question	request	statement	accusation	challenge	acceptance	denial	appreciation
Annotator1	conservative	0.95	0.88	0.78	0.61	0.51	0.86	0.72	0.80
Annotator2	conservative	0.98	0.93	0.93	0.95	0.96	0.95	0.94	0.86
Annotator3	conservative	0.98	0.89	0.85	0.87	0.72	0.88	0.86	0.86
Annotator1	relaxed	0.95	0.91	0.85	0.72	0.65	0.92	0.78	0.80
Annotator2	relaxed	0.98	0.91	0.92	0.86	0.72	0.91	0.86	0.86
Annotator3	relaxed	0.99	0.95	0.94	0.93	0.93	0.94	0.95	0.90

Table 7: Macro-F1 scores for annotators comparing annotator-specific annotations to groundtruth labels.

statistical tests do not show p-values < 0.05. The Avg. SetFit model fares better, achieving a statistically significant improvement in performance for five action classes.

Based on the tests, we conclude that considering more than one action in modeling can notably increase model performance (RQ1).

7.2 RQ2: Modeling Ambiguity

To answer RQ2, we utilized multiple annotations and different approaches to leverage them to find the best model for representing the ambiguity related to actions. Model performances can be seen in Table 8. We include annotator performances evaluated against ground-truth labels for comparison, in Table 7.

Annotator2 model performed best overall. Macro-F1s were somewhat higher when predicting relaxed ground truth labels. Also, Individual Annotator model performances differ among annotators, also in statistical tests. Also annotator performances (Table 7 differ notably. The PNC model performs poorly, more so than SVM in Table 6. Statistical tests show that for *questions* and *statements*, results are significantly lower. PNC models performed worst for the Complicated class, similarly to Jiang and de Marneffe (2022), perhaps due to class heterogeneity; other classes' performances are much higher.

In the comparison of ensembles for predicting all possible annotations for each text, the Annotator2 model, Averaged model or combinations of two or three Individual Annotator models fared best according to Jaccard coefficient scores. For relaxed ground truth label predictions, the Averaged model fairs best. These provide best insights matching human annotations and multiple interpretations.

8 Discussion

We investigated how comments in asynchronous crisis conversations could be best approached to account for their contextual characteristics, showing that predicting multiple actions for comments helps statistically improve model performance.

Likert score annotation allowed us to consider the presence of actions on different scales of strength, rather than categorically (Glickman and Dagan, 2005). We showed that an ensemble model with 2-3 individual annotator models with relaxed ground truth labels could best predict all possible annotations relevant to our data. We also illustrated that the results differ somewhat between actions.

Model	Feature	Ground truth	Action							
			question	request	statement	accusation	challenge	acceptance	denial	appreciation
SetFit	PNC	cross-e.	0.71	0.55	0.57	0.50	0.50	0.59	0.47	0.39
	Avg.	conservative	0.96	0.74	0.74	0.65	0.54	0.73	0.58	0.72
	Annotator1	conservative	0.92	0.75	0.75	0.51	0.43	0.77	0.55	0.65
	Annotator2	conservative	0.95	0.77	**0.84**	0.72	**0.76**	0.78	**0.68**	0.66
	Annotator3	conservative	**0.97**	0.71	0.78	0.66	0.52	0.74	0.63	0.69
	Avg.	relaxed	**0.97**	**0.80**	0.76	0.65	0.63	0.79	0.56	0.65
	Annotator1	relaxed	0.95	0.77	0.79	0.60	0.52	**0.82**	0.62	0.52
	Annotator2	relaxed	**0.97**	0.76	0.79	**0.75**	0.61	0.78	**0.68**	**0.73**
	Annotator3	relaxed	**0.96**	0.71	0.78	0.66	0.52	0.74	0.63	0.68

Table 8: 10-fold cross-validated macro-F1 scores for best performing learner utilizing different approaches to leveraging multiple annotations.

Ground truth	Ensemble														
	A1	A2	A3	Avg.	A1+A3	A1+A2	A2+A3	A1+A2 +A3	A1 +Avg.	A2 +Avg.	A3 +Avg.	A1+A2 +Avg.	A1+A3 +Avg.	A2+A3 +Avg.	A1+A2+ A3+Avg.
conserv.	0.52	**0.66**	0.59	**0.66**	0.57	**0.67**	**0.65**	**0.65**	0.35	0.43	0.38	0.43	0.40	0.44	0.43
relaxed	0.42	0.42	0.42	**0.69**	0.42	0.43	0.42	0.41	**0.62**	0.54	0.58	0.53	0.57	0.52	0.52

Table 9: Jaccard coefficient scores for ensemble models using best SetFit models. A1=Annotator1 model, A2=Annotator2, A3=Annotator3, Avg.=Averaged model.

We provided an annotation scheme and annotated dataset for identifying actions in asynchronous crisis conversations in Finnish. This is important as there are no such resources for Finnish yet. We feel that future work analyzing manipulative behaviors including controversial and ambiguous actions will benefit from utilizing our scheme and dataset. The scheme will also enable easier implementation of novel models for languages as well. We provide English translations for support. However, although some actions in our scheme are common across languages (Enfield et al., 2010), contextual differences should be considered: e.g. *apologies* were not found in our data, but have been relevant elsewhere (Paakki et al., 2021). Furthermore, as (adapted) models are often needed in low-resource settings, we showed that even with a relatively small annotated dataset we can reach good performance using few-shot learning.

From a CA based pragmatics perspective, it is challenging to systematically identify actions in asynchronous conversations due to action-taking being context-dependent, implicit or indirect. Interpretation is not a product but a process: meanings of actions are interpreted by participants collaboratively and on-line (Clark and Schaefer, 1989; Jurafsky, 1992). Participants might alter interpretations of comments across turns in conversation. More specifically, our results show that face-threatening actions (*challenge*, *denial*), especially, are more difficult to annotate and/or model than others. This is in line with theoretical views on action-taking: people tend to express these more implicitly or indirectly to avoid face-threats (Brown and Levinson, 1987), which might lead to uncertainty in their interpretation. In their case, we consider it crucial to be able to model their related interpretative ambiguity. Ensemble models and multiple annotations can be helpful to accomplish this.

9 Conclusions and Future Work

We showed how a simplified action identification model with theoretical support can reflect the multi-action properties and ambiguity of turns in controversial asynchronous conversation. Although annotator disagreements have been studied increasingly in NLP, there is still room for exploring how to utilize them in the pragmatic analysis of actions. Future work could develop the identification of some difficult categories involving high levels of ambiguity (e.g. *challenges*), and further utilization of annotation scores and contextual information in more fine-grained models. We conclude that digital CA based modeling of actions in asynchronous data can be fruitful for analyzing the ambiguity related to controversial crisis discussions online.

9.1 Limitations

Crowd-sourcing is often seen to provide heterogeneous, arguably more valid annotations from a large population (Weber et al., 2018). In expert annotation, annotators adjust their work based on expectations regarding outcomes, thus reaching higher agreements with annotation reliability maximized to reflect the desired categories (Weber et al., 2018) – a possible limitation of our work. However, CA analysis requires contextual in-depth reading, which is why non-expert annotation would have been unreliable and unsuitable here (Eickhoff, 2018). We had only three annotators; some suggest that a higher number can lead to better results (Pavlick and Kwiatkowski, 2019). Due to resource limitations, we considered our current scope of annotators and data sufficient. These could be extended in future research.

Likert scale annotation allows fluidity in interpretation, but subjectivity, confidence, signal strength, and understandings of scores and categories might be melted into one metric. Also, we assigned per-comment action scores; another option would be to label segments of the message. We considered this challenging as sometimes the boundaries of actions were unclear or overlapped. Future research could explore this option further.

Crisis related discussions may involve sensitive information, despite the fact we are dealing with openly available social media data. We have translated and modified examples, and anonymized and de-identified the data so that the content is conveyed without privacy concerns. Personal data including names or user IDs will not be stored or used (in models or other purposes). We have published a privacy notice according to University policy regarding data collection on our website. This will be reported at article publication. The data will not be shared publicly: only anonymized data where sensitive information has been removed will be shared with researchers through an application process. This will, of course, limit the possibilities of data sharing.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

James Allen and Mark Core. 1997. Damsl: Dialog act markup in several layers. draft. *Draft of manual (31 March 1997)*.

Neli Arabadzhieva-Kalcheva and Ivelin Kovachev. 2022. Comparison of BERT and XLNet accuracy with classical methods and algorithms in text classification. In *2021 International Conference on Biomedical Innovations and Applications (BIA)*, volume 1, pages 74–76. IEEE.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Kevin G. Barnhurst and Diana Mutz. 1997. American journalism and the decline in event-centered reporting. *Journal of Communication*, 47(4):27–53.

David B. Bracewell, Marc Tomlinson, and Hui Wang. 2013. Semi-supervised modeling of social actions in online dialogue. In *2013 IEEE Seventh International Conference on Semantic Computing*, pages 168–175. IEEE.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*, volume 4. Cambridge university press.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Stephanie L. Castro. 2002. Data analytic methods for the analysis of multilevel questions: A comparison of intraclass correlation coefficients, rwg(j), hierarchical linear modeling, within- and between-analysis, and random group resampling. *The Leadership Quarterly*, 13(1):69–93. Benchmarking Multilevel Methods in Leadership.

Alexander Clark and Andrei Popescu-Belis. 2004. Multi-level dialogue act tags. In *SIGdial 2004 (5th SIGdial Workshop on Discourse and Dialogue)*, pages 163–170. ACL-Association for Computational Linguistics.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316, Barcelona, Spain. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Joaquín Derrac, Salvador García, Daniel Molina, and Francisco Herrera. 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1):3–18.

Ian Dersley and Anthony Wootton. 2000. Complaint sequences within antagonistic argument. *Research on language and social interaction*, 33(4):375–406.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fabrizio Di Mascio, Michele Barbieri, Alessandro Natalini, and Donatella Selva. 2021. Covid-19 and the information crisis of liberal democracies: Insights from anti-disinformation action in italy and eu. *Partecipazione e conflitto*, 14(1):221–240.

Nathan Duran, Steven Battle, and Jim Smith. 2022. Inter-annotator agreement using the conversation analysis modelling schema, for dialogue. *Communication Methods and Measures*, 16(3):182–214.

Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 162–170, New York, NY, USA. Association for Computing Machinery.

N.J. Enfield, Tanya Stivers, and Stephen C. Levinson. 2010. Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, 42(10):2615–2619.

Elisa Ferracane, Greg Durrett, Junyi J. Li, and Katrin Erk. 2021. Did they answer? Subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics.

Eric Forsyth and Craig Martell. 2007. Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing (ICSC 2007)*, pages 19–26. IEEE.

Simone Fuscone, Benoit Favre, and Laurent Prévot. 2020. Filtering conversations through dialogue acts labels for improving corpus-based convergence studies. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 203–208, 1st virtual meeting. Association for Computational Linguistics.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27.

Souvick Ghosh and Satanu Ghosh. 2021. Classifying speech acts using multi-channel deep attention network for task-oriented conversational search agents. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR'21, page 267–272, New York, NY, USA. Association for Computing Machinery.

David Giles, Wyke Stommel, Trena Paulus, Jessica Lester, and Darren Reed. 2015. Microanalysis of online data: The methodological development of "digital CA". *Discourse, context & media*, 7:45–51.

Oren Glickman and Ido Dagan. 2005. A probabilistic setting and lexical coocurrence model for textual entailment. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 43–48, Ann Arbor, Michigan. Association for Computational Linguistics.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48:493–531. Open access.

Susan Herring. 1999. Interactional coherence in CMC. *Journal of Computer-Mediated Communication*, 4(4).

Susan Herring, A. Das, and S. Penumarthy. 2005. CMC act taxonomy.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Shafiq Joty and Enamul Hoque. 2016. Speech act modeling of written asynchronous conversations with task-specific embeddings and conditional structured models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1756.

Jakob Jünger and Till Keyling. 2019. Facepager. an application for automated data retrieval on the web. *Source code and releases available at https://github.com/strohne/Facepager (Accessed June 16 2023).*

Daniel S. Jurafsky. 1992. *An On-line Computational Model of Human Sentence Interpretation: A Theory of the Representation and Use of Linguistic Knowledge*. University of California, Berkeley.

Irene Koshik. 2003. Wh-questions used as challenges. *Discourse Studies*, 5(1):51–77.

Michael K. Lindell and Christina J. Brandt. 1999. Assessing interrater agreement on the job relevance of a test: A comparison of CVI, T, $r_{wg(j)}$, and $r^*_{wg(j)}$ indexes. *Journal of applied psychology*, 84(4):640.

Joanne Meredith. 2017. Analysing technological affordances of online interactions using conversation analysis. *Journal of pragmatics*, 115:42–55.

Joanne Meredith and Elizabeth Stokoe. 2014. Repair: Comparing facebook 'chat' with spoken interaction. *Discourse & communication*, 8(2):181–207.

Cristian Moldovan, Vasile Rus, and Arthur C Graesser. 2011. Automated speech act classification for online chat. In *Midwest Artificial Intelligence and Cognitive Science Conference (MAICS)*, volume 710, pages 23–29.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on Natural Language Inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Thomas A O'Neill. 2017. An overview of interrater agreement on likert scales for researchers and practitioners. *Frontiers in psychology*, 8:777.

Henna Paakki, Antti Salovaara, and Heidi Vepsäläinen. 2020. Do online trolling strategies differ in political and interest forums: Early results. In *Disinformation in Open Online Media*, pages 191–204, Cham. Springer International Publishing.

Henna Paakki, Heidi Vepsäläinen, and Antti Salovaara. 2021. Disruptive online communication: How asymmetric trolling-like response strategies steer conversation off the track. *Computer Supported Cooperative Work (CSCW)*, pages 1–37.

Henna Paakki, Heidi Vepsäläinen, Antti Salovaara, and Bushra Zafar. 2023. Detecting covert disruptive behavior in online interaction by analyzing conversational features and norm violations. *ACM Trans. Comput.-Hum. Interact.* Just Accepted.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: Evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46:219–252.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626.

Barbara Plank. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 10671–10682. Association for Computational Linguistics.

Rezvaneh Rezapour, Jutta Bopp, Norman Fiedler, Diana Steffen, Andreas Witt, and Jana Diesner. 2020. Beyond citations: Corpus-based methods for detecting the impact of research outcomes on society. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6777–6785, Marseille, France. European Language Resources Association.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. The simplest systematics for the organization of turn-taking for conversations. *Language*, 50(4):696–735.

Renata Savy. 2010. Pr.A.Ti.D: A coding scheme for pragmatic annotation of dialogues. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge University Press, Cambridge; New York.

Tanya Stivers. 2015. Coding social interaction: A heretical approach in conversation analysis? *Research on Language and Social Interaction*, 48(1):1–19.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Wyke Stommel and Tom Koole. 2010. The online support group as a community: A micro-analysis of the interaction with a new member. *Discourse studies*, 12(3):357–378.

Saatviga Sudhahar, Giuseppe A. Veltri, and Nello Cristianini. 2015. Automated analysis of the US presidential elections using big data and network analysis. *Big Data & Society*, 2(1):2053951715572916.

Motoki Taniguchi, Yoshihiro Ueda, Tomoki Taniguchi, and Tomoko Ohkuma. 2020. A large-scale corpus of E-mail conversations with standard and two-level dialogue act annotations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4969–4980, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maksim Terpilowski. 2019. Scikit-posthocs: Pairwise multiple comparison tests in python. *The Journal of Open Source Software*, 4(36):1169.

Jenny Thomas. 1995. *Meaning in Interaction: An Introduction to Pragmatics*. Longman.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2021. Emotion ratings: How intensity, annotation confidence and agreements are entangled. *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, pages 1–14.

Jason J Turowetz and Douglas W Maynard. 2010. Morality in the social interactional and discursive world of everyday life. In Hitlin S. and Vaisey S., editors, *Handbook of the Sociology of Morality*, pages 503–526. Springer, New York.

Douglas P. Twitchell and Jay F. Nunamaker. 2004. Speech act profiling: A probabilistic method for analyzing persistent conversations and their participants. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, volume 5, page 40107c. IEEE.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Mikko T. Virtanen, Heidi Vepsäläinen, and Aino Koivisto. 2021. Managing several simultaneous lines of talk in finnish multi-party mobile messaging. *Discourse, Context & Media*, 39:100460–100474.

René Weber, Michael J. Mangus, Richard Huskey, Frederic R. Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn, and Ron Tamborini. 2018. Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, 12(2-3):119–139.

Yimin Xiao, Zong-Ying Slaton, and Lu Xiao. 2020. TV-AfD: An imperative-annotated corpus from the big bang theory and Wikipedia's articles for deletion discussions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6542–6548.

Stepan Zakharov, Omri Hadar, Tovit Hakak, Dina Grossman, Yifat Ben-David Kolikant, and Oren Tsur. 2021. Discourse parsing for contentious, non-convergent online discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 853–864.

Amy X. Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Eleventh International AAAI Conference on Web and Social Media*, volume 11, pages 357–366.

Justine Zhang, Cristian Danescu-Niculescu-Mizil, Christina Sauper, and Sean J. Taylor. 2018. Characterizing online public discussions through patterns of participant interactions. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–27.