Implicit and Indirect: Detecting Face-threatening and Paired Actions in Asynchronous Online Conversations

Henna Paakki, Aalto University, Espoo; University of Helsinki, henna.paakki@helsinki.fi

Pihla Toivanen, University of Helsinki

Kaisla Kajava, Aalto University, Espoo, kaisla.kajava@aalto.fi

Abstract This paper presents an approach to computationally detecting face-threatening and paired actions in asynchronous online conversations. Action detection has been widely studied for synchronous chats. However, there are fewer models or datasets for asynchronous conversations, and they have not included some of the face-threatening actions central to online conversations involving misbehavior like trolling. We examine asynchronous crisis news related online conversations in Finnish, providing an annotation scheme for identifying central actions used in this conversational context. An important contribution is to include face-threatening actions in the scheme, and training computational classifiers for their detection with improved performance compared to prior work. We illustrate that face-threatening actions are important for analyzing conversations related to crisis news. We show that for computational action detection, it is essential to be able to represent how multiple actions may be performed within one comment, and how ambiguity in the expression of actions often leads to multiple possible label interpretations. Annotating actions using scores helps to reflect these characteristics. We also find that an ensemble of models trained on individual annotators' annotations can best represent multiple potential interpretations of action labels. These are especially relevant for face-threatening actions.

1 Introduction

Natural Language Processing (NLP) and Machine Learning (ML) methods are popular for analyzing textual content on social media, e.g. discourse signals (Ferracane et al., 2021; Zhang et al., 2017). These methods combined with examining the structural features of conversation, like comment-response relations, allow rich automated analyses of interaction (Zhang et al., 2018; Sudhahar et al., 2015), crucial for identifying online misbehavior like trolling (Paakki et al., 2024), or aggression (Zhang et al., 2018; Garimella et al., 2018). While computational approaches to detecting online manipulation have focused on revealing content sharing networks (Giglietto et al., 2020) and disinformation campaign content (Shu et al., 2017; Zhou and Zafarani, 2020), fewer have investigated conversational interactions on online forums in more detail. Asynchronous news comment and forum conversations are sites where people often seek to influence others' opinions. The impacts may be widespread, especially for crisis news discussions, as asynchronous discussions are persistent online and thus potentially reach large audiences (Zhang et al., 2018). Thus, this is an important context to investigate.

Recent research has highlighted the importance of investigating actions in conversational online interaction to reveal manipulative behavior like trolling (Paakki et al., 2024). Computational models can help accomplish this at scale, which is why computational methods for detecting actions in this context need more attention. Although actions in 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 fewer models or datasets for analyzing actions in asynchronous arenas. Moreover, existing annotation schemes or datasets do not include face-threatening actions (e.g., *accusations* and *challenges*) that are important for analyzing trolling and disinformation (Paakki et al., 2024; Bellutta et al., 2021). This type of interaction often strategically utilizes ambiguity in expression (Paakki et al., 2021). In this paper, the concept of *action* refers to what functions a turn has in conversation in relation to other turns. We understand face-threatening actions as having the potential to spoil or threaten the face of their addressee by showing a negative orientation toward them or acting against their wishes (Brown and Levinson, 1987).

Our most important overarching goal is to investigate how to computationally approach face-threatening and paired actions (e.g. *question-answer*) in

https://doi.org/10.3384/nejlt.2000-1533.2025.5980

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, 2 likes)	*statement* AND *appreciation*
B: A, hallelujah! (1 laugh emoji, 1 like)	*statement* OR *appreciation*

Table 1: Extract from Ukraine war discussion, 2022.

asynchronous conversations. To this end, we develop an annotation scheme and computational models for classifying face-threatening and paired actions that are central for this context. This will allow analyzing, at scale, how participants use and respond to common actions. Moreover, to represent how actions unfold in this type of conversation, we deem it important to be able to account for multiple actions per comment and multiple possible interpretations of actions. We argue that this will help to better align computational classification with how actions are performed in these interactions, while also leading to better classification performance. We have two research questions:

- RQ1. Does considering multiple actions per comment better reflect how actions are performed in asynchronous conversation in contrast to a single-label approach?
- RQ2. What approach can best represent all possible interpretations of action labels?

We anticipate that considering multiple actions in comments and multiple possible label interpretations is important for better representing what comments do in asynchronous conversation. For example, as seen in Table 1, comments often 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.

We utilize a Conversation Analysis (CA) driven theoretical framework based on Paakki et al. (2021) to build and refine an annotation scheme for identifying actions (8 actions in total) in asynchronous conversations in Finnish – a low-resource setting. The unit of analysis is one comment in a conversation thread. Annotators rate the likelihood of each action on a 7-point Likert scale. Such an approach conveys how face-threatening actions like *challenges*, for instance, are mostly not categorically present in a comment, but instead often implicitly expressed and thus better annotated using scalar values. We compare the performances of models assuming only one action *vs.* models allowing multiple actions per comment. We experiment with different approaches to leveraging annotator disagreements to investigate which approach can best predict all possible action labels for each comment.

Our contributions include an annotation framework for face-threatening and paired actions in asynchronous conversations, and action classifiers for Finnish. We showcase annotation disagreement in the case of asynchronous conversation, illustrating how face-threatening actions tend to involve higher levels of disagreement than other actions. We find that actions are frequently performed in an implicit or indirect manner in this context. Also, comments often include more than one action. We show that considering multiple actions per comment and multiple valid interpretations of actions allows higher classification performance, and that an ensemble combining few-shot learning based ML models trained separately with individual annotators' annotations can best represent all possible labels.

This approach is important for computationally analyzing how people perform actions in comments to asynchronous conversation, and how they respond to previous actions. Identification of face-threatening actions – not included in prior computational models – is especially crucial for crisis related conversations, and for detecting misbehavior like trolling online (Paakki et al., 2024). We provide our annotation guidelines and related materials on our GitHub page[1], and models on Huggingface[2].

2 Identifying Online Actions

This section introduces earlier work relevant to this paper: we introduce 1.) the theoretical foundations of our work (CA), 2.) prior work on computational action detection, and 3.) research highlighting how tasks like action detection may involve meaningful disagreements.

2.1 CA and Asynchronous Conversation

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

[1] Detailed annotation guidelines and models, to the extent that it does not compromise any individual's privacy, are provided on our GitHub page: https://github.com/henniina/Detecting-paired-actions.

[2] https://huggingface.co/Finnish-actions

and interpret its role (Sacks et al., 1974). CA has a robust theoretical foundation for studying the dynamics between paired actions (i.e. adjacency pairs) in turn-by-turn interactions (Schegloff, 2007), as well as face-threatening actions like accusations and their expected responses (Antaki et al., 2008; Dersley and Wootton, 2000). Their characteristics have been well established in CA (Dersley and Wootton, 2000; Koshik, 2003; Turowetz and Maynard, 2010). Action pairs allow us to coordinate interaction through actions and responses that match their expectations (Stivers and Rossano, 2010). They reflect accountability, which is crucial for cooperation (Enfield and Sidnell, 2017).

What differentiates CA's understanding of actions from, e.g., Speech Acts, is that interpretations are based on the analysis of interactions between turns rather than judging the intent behind a turn in conversation (Sacks et al., 1974). Speech Act theory, although much used in computational approaches to actions (or acts), has been criticized for its speaker centric perspective (Linell and Marková, 1993; Savolainen, 2020). It has also been found that turns in asynchronous conversations cannot very well be classified into Speech Act categories (Qadir and Riloff, 2011).

Due to these concerns, we root our approach in CA. Studies have shown it to be well suited for analyzing online interactions (Giles et al., 2015; Meredith and Stokoe, 2014). Digital CA research stresses the need to consider the specific characteristics of different types of online interaction (Virtanen et al., 2021; Meredith, 2017). In asynchronous conversation, a real-time back-and-forth is not expected, and participants often tend to include several actions in one message to accomplish more at one go (Virtanen et al., 2021). Participants may enter and exit the discussion whenever, and adjacency pairs may be disrupted by other messages in between (Herring, 1999). Others have also shown face-to-face or synchronous chats to differ from asynchronous interactions (Taniguchi et al., 2020; Xiao et al., 2020).

Despite the differences between asynchronous conversations and more traditional contexts of CA research (e.g., face-to-face conversation), action pairs are a fruitful CA concept for the analysis of online conversation. This is because people have been shown to treat actions and their norms online quite similarly to what has been established in CA research on face-to-face interaction (Meredith, 2017; Paakki et al., 2021; Salonen et al., 2022). Despite disrupted turns, people tend to utilize sequentially organized actions to maintain coherent interactions online (Stommel and Koole, 2010; Meredith and Stokoe, 2014). Conversational norms dictate that an action that expects an answer should be answered in accordance to its normative expectation (e.g. a question requires an (informative) answer) (Sacks et al., 1974; Schegloff, 2007; Enfield et al., 2010). Unexpected answers are seen to signal a problem, for instance a misunderstanding, a gap in presumed shared knowledge, or reluctance (Clark and Schaefer, 1987; Pomerantz, 1984). They can also be used to purposefully direct and disrupt a conversation (Paakki et al., 2021).

2.2 Computational Approaches

Most prior research on computational action detection 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). These, however, represent a context different from casual, anonymous, and asynchronous online conversation (Herring, 1999), for which there are fewer resources, and no large annotated datasets like the Switchboard corpus (Jurafsky, 1997).

Asynchronous conversations have been researched increasingly in recent years. Annotation schemes, for example, have been developed for different asynchronous conversational contexts (Duran and Battle, 2018; Herring et al., 2005; Jeng et al., 2017; Kim et al., 2010; Savolainen, 2020; Stivers, 2015; Wang et al., 2014). Some earlier work on computational classification has focused on email interaction (Carvalho and Cohen, 2005; Cohen et al., 2004; Taniguchi et al., 2020). Others have looked into online forums, newsgroups and question answering sites (Bhatia et al., 2014; Joty and Hoque, 2016; Kim et al., 2010; Feng et al., 2006; Fortuna et al., 2007). Prior approaches have most often depended on Dialogue Act (DA) driven classification (Bhatia et al., 2014; Carvalho and Cohen, 2005; Cohen et al., 2004; Joty and Hoque, 2016; Kim et al., 2010; Taniguchi et al., 2020), based on Speech Act Theory (Austin, 1975), and researchers have developed both message-level and sentence-level classification of DAs. Joty and Mohiuddin (2018) studied the effect of modeling conversational dependencies on DA classification. Taniguchi et al. (2020) trained both sentence-level and message-level DA detection models for emails using BERT (Devlin et al., 2019). A large dataset using asynchronous Reddit conversations was published by Zhang et al. (2017), annotating and classifying discourse relations based on earlier work on Rhetorical Structure Theory (Mann and Thompson, 1988) and online DAs (Feng et al., 2006; Fortuna et al., 2007).

Existing resources for asynchronous conversation, however, mostly do not include actions like *accusations* or *challenges* relevant to analyzing manipulative online behavior (Bellutta et al., 2021; Paakki et al., 2021, 2024). Though action pairs have been investigated in question-answer modeling (of DAs) (Joty and Mohiuddin, 2018; Taniguchi et al., 2020), not much attention has been paid to face-threatening actions. The only computational approaches we are aware of that have included face-threatening actions relevant to our interests are Paakki et al. (2024) (*challenge* and *accusation*), Bracewell et al. (2012; 2013) (*challenge credibility*, *disrespect*, *relationship conflict* and *task conflict*), Feng et al. 2006 (*criticize*), Zhang et al. (2017) (*negative reaction*) and Zakharov et al. (2021) (e.g. *complaint*). Bracewell et al. (2012) concentrate on acts reflecting psychological study of power and leadership, Zhang et al. (2017) on discourse relations, and Zakharov et al. (2021) on a wide selection of discursive categories based on Bakhtinian theories of discourse (e.g. related to tone and style) (Bakhtin, 1981). The listed studies, however, were based on different theoretical premises and thus did not concentrate on paired actions. Paakki et al. (2024) used a set of 10 paired actions, but they utilized 0-shot Natural Language Inference (NLI) for classification instead of providing a model trained specifically for this task.

2.3 Actions and Disagreement

We consider it relevant to be able to reflect the multiple possible interpretations of action labels in our data. Ambiguity is a natural part of human interaction, and it is often used strategically in manipulative interaction (Paakki et al., 2021). Actions, as well, are sometimes expressed in an ambiguous manner, leading to different interpretations (Thomas, 1995). Similarly, for many NLP tasks there is no single ground truth (Jiang and de Marneffe, 2022; Plank, 2022; Uma et al., 2022; Yang, 2021), due to uncertainty in text meaning. This leads to different annotator interpretations of label distribution, constituting meaningful systematic disagreement (Jiang and de Marneffe, 2022; Nie et al., 2020). Including annotator disagreement into models has also been shown to improve model performance on some NLP tasks (Passonneau et al., 2012; Plank, 2022). Models leveraging annotator disagreements are thus needed for better representing actions in online interaction.

Most existing action detection models rely on one ground truth (e.g. Zhang et al. (2017); Joty and Mohiuddin (2018)). An exception is Ferracane et al.'s (2021) model, which sought to predict all valid interpretations of actions in their data. Another study by Taniguchi et al. (2020) predicted both sentence-level and message-level interpretations of actions, which provides insight into how readings of parts versus the whole message contributed to interpretations of labels – though not further drawing from annotator disagreements. However, the former used live congressional hearings and the latter emails as data. These approaches notably differ from our context of asynchronous forum conversations, which involves ambiguous use of actions in medium length texts (in contrast to the formerly mentioned types of data), with frequent use of face-threatening actions (e.g. *accusations*).

Topics	#Comments	#Annotations
Covid: 675	1,204	1,204 (single annotations)
War: 529		3,612 (multiple annotations)

Table 2: Dataset description. An annotation refers to one set of 8 scores, one score per action.

3 Data

To answer our RQs, we collected asynchronous conversations under news regarding the COVID-19 Pandemic and Ukraine war posted on Facebook. These were of interest, as crises like these have societal, economic and environmental force, and drive forward change and renewal due to the need for novel courses of action; they also create uncertainty, making online discourses vulnerable to trolling, manipulation and disinformation (Di Mascio et al., 2021). Our data comes from public Facebook (FB) pages owned by Finnish news media: Yleisradio (Yle) and Helsingin Sanomat (HS).[3] Our dataset is described in Table 2. To conduct our experiments, we manually annotated news comments using a digital CA based framework, informed by earlier research (Clark and Schaefer, 1989; Herring et al., 2005; Paakki et al., 2021; Stivers, 2013) and data-driven insights (see section 4).

We used Facepager[4] (v.4.5.3) (Jünger and Keyling, 2019) (MIT License) 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 news article. To select a subset of random conversations for manual annotation, we shuffled the scraped data per news posts, keeping all comments part of the same conversation together. In this study, we had three annotators (the authors). All annotators are native Finnish speakers, living in Finland, and have training in linguistics and/or computational linguistics and data science[5]. We divided the data into three parts, one for each annotator. Each annotator manually selected the first 400 crisis-related comments from their sample for annotation. This meant that we had to manually read the titles and descriptions of 100-150 news posts to find conversations on COVID or the Ukraine war.

We annotated 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

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

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

[5] Two of us have MA level training in linguistics/computational linguistics and further doctoral training in computer science and computational linguistics. One of us has MA level training in data science, and further doctoral training in digital humanities and social sciences.

one comment with two or more replies, as we were interested in conversational interaction. While annotating, we referred to the original FB page for better readability and to validate the data.[6] We excluded any comments we had already seen during annotation scheme development (see section 4). To achieve greater variation in comments, we included a maximum of 30 comments related to the same piece of news. The number of comments to a news post (when comments were allowed) ranged from 14 to 684, with a mean of 90.56 and median 84.5, and comment mean length being 151.4 characters, median length 105.0.

We annotated all comments and replies following our refined annotation guidelines (see section 4). To compare the performances of models trained on single annotations against models leveraging multiple annotations (and thus annotation disagreements), we used two different versions of the data: data with **single annotations**, where each comment has only one annotation, and data with **multiple annotations**, where all comments have three annotations, one per annotator. This enabled us to investigate how to best represent multiple possible interpretations of actions in our data (RQ2).

4 Annotation Methods

We based our initial annotation scheme on earlier research regarding which actions emerge as relevant in conversations involving possible online trolling (Paakki et al., 2021). Speech Act, DA and CA-related online research alike have emphasized that for different contexts or events different actions matter. For example, crisis (or emergency) related online discussions are more geared toward orders and suggestions than celebrity discussions (Laurenti et al., 2022; Zhang et al., 2011). For conversations including potential trolling or manipulation, different forms of assertive and face-threatening actions are relevant: *accusations*, *challenges*, and *statements* (Paakki et al., 2021). Also, investigating actions that form pairs is important in this context, e.g., *accusation-denial* (Paakki et al., 2021). Paired actions are well-established in CA and Computer-Mediated Communication (CMC) literature (Clark and Schaefer, 1989; Enfield et al., 2010; Schegloff, 2007; Stivers, 2015).

The final scheme (Table 3) was formulated based on data-driven empirical insights during our incremental development of the scheme and annotation guidelines, to ensure that it represents the most central rhetorical and interactive functions of comments for our context of investigation. We include both responsive actions and ones initiating a paired action, which expect specific responses, marked with an asterisk in Table 3 (based on Paakki et al. (2021) and above-listed CA literature). These are important for analyzing the dynamics of user-to-user interaction in conversations with potential trolling, e.g., whether actions expecting a specific type of response are responded to in a normatively expected manner.

Action	Description	Expected response
Question* *(initiating)*	Asks someone for information (request for information)	statement
Request* *(initiating)*	Requests, proposes or tells someone to perform an action.	denial, acceptance
Challenge* *(initiating)*	Refutes someone's epistemic claim or authority.	acceptance, denial, statement
Accusation* *(initiating)*	Points out a reprehensible act performed by someone.	acceptance, denial
Statement *(initiating/ responding)*	Asserts an opinion, information, wish, neutral or negative evaluation, or answer (to question).	
Appreciation *(initiating/ responding)*	Positive evaluation or comment about an actor, event or object.	
Acceptance *(responding)*	Agrees, or accepts a request, statement or challenge, or admits an accusation.	
Denial *(responding)*	Rejects or denies an action.	

Table 3: Annotation scheme (*=expects response).

There is a large number of possible actions, studied in CA and CMC (Schegloff, 2007). We aimed at a simplified annotation scheme, because 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), but finally reduced them to 8 actions. Merging of categories was based on actions having similar functions (e.g., *proposals* and *requests*), and difficulty faced in distinguishing them from each other (e.g., *denial*, *rejection*). Moreover, different tasks often require different annotation schemes de-

[6] Due to the restrictions of Facepager, we had to retrieve some missing comments manually.

#Actions	#C(s)		#C(m)					
	A1-3	%	A1	%	A2	%	A3	%
0	17	1.3%	56	4.7%	1	0.08%	4	0.3%
1	722	60%	885	73.5%	477	39.6%	676	56.1%
2	360	30%	238	19.8%	456	37.9%	417	34.6%
3	95	7.9%	23	1.9%	217	18%	94	7.8%
4	10	0.8%	2	0.2%	50	4.2%	12	1%
> 4	1	0.08%	0	0%	3	0.3%	1	0.08%

Table 4: Number of actions labeled per comment (C) in single (s) and multiple (m) annotations (A1=Annotator1, A2=Annotator2, A3=Annotator3).

Action	Count(s)	Count(m)	%	r_{WG}
Question	243	276	0.94	0.88
Request	130	186	0.93	0.90
Statement	857	999	0.59	0.58
Challenge	138	381	0.91	0.65
Accusation	155	305	0.86	0.74
Appreciation	36	66	0.97	0.94
Acceptance	113	171	0.97	0.89
Denial	97	151	0.95	0.87

Table 5: Distribution of annotated actions in our data (s = single annotations, m = multiple annotations), and inter-annotator agreements (% agreement and r_{WG}) with a separate test set (N=100).

pending on forum type: e.g. question-answer forums (Savolainen, 2020; Jeng et al., 2017) and trolling conversations (Paakki et al., 2021) require focus on different classes, so empirical insights were important here.

We aimed to avoid under-specification and unnecessary disagreement by iteratively developing our annotation scheme and annotation guidelines. We developed our scheme before data scraping: annotating, negotiating, and analyzing practice data based on our guidelines at each step. The development involved 15 iterations, during which we honed the guidelines and adapted the scheme in a data-driven manner. The 4 first iterations started with single-label annotation, which we felt did not align with the empirical phenomenon at hand; we found score-based multilabel annotation much more fitting. The data used in the development was manually selected as screen captures and links from the first 300 crisis news posts in Yle's and HS's FB pages we found, between August–December 2022. This data was not included in the final annotated datasets. During annotation, we read the conversations in their original FB context. We stopped when we had reached sufficient agreement (upper boundary of medium agreement or strong agreement) in annotation, and a framework that we felt had a meaningful separation between classes and a scheme that was applicable to our data. Using the finalized scheme, we annotated so far unseen data selected for manual annotation (see section 3).[7] See label distributions in Table 5, and action-specific distributions in Appendix C.

Our task required expertise, training and familiarity with the theoretical premises of (digital) CA, including what constitutes each action, so expert annotation was chosen, meaning that the annotators had linguistics, computational linguistics and social science expertise. 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 chose a message-level approach – comment as the unit of analysis – as it has been found useful when computationally identifying question-answer pairs (Taniguchi et al., 2020). Since we focus on face-threatening actions and action pairs, we found this a suitable option instead of sentence-level detection. Research on Speech Acts and CA in offline as well as digital contexts has highlighted that messages may often perform several actions in one message to accomplish more at one go, e.g. to respond to some action and to initiate a new one (Goffman, 1974; Levinson, 2013; Stommel and Koole, 2010; Virtanen et al., 2021). Although there are many differences in how turn-taking unfolds in asynchronous conversation in contrast to face-to-face, like the fact that turns cannot be interrupted, people often treat comments as turns in online conversation (Meredith, 2019; Virtanen and Kääntä, 2018), see also section 6.1. However, actions are not always marked with clear boundaries (e.g., line breaks) but might overlap, as in Table 9. CA researchers have emphasized that for interpreting what action a turn commits, a holistic assessment of the speaker's conduct or purposive, goal oriented behavior is needed; no single feature in a turn necessarily signals a specific action (Enfield and Sidnell, 2017; Rossi, 2018). Many factors might influence the interpretation of actions: e.g., the relevance of an action might influence interpretations of expected responses (Stivers and Rossano, 2010). In our data, participants tended to orient to comments as having some main pair-initiating action, even when the comment included several actions, as seen in Table 9. These considerations supported our choice of message-level detection.

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, inspired by previous work on annotation disagreement by e.g. Peterson et al. (2019) and work that has illustrated the usefulness of Likert scores in annotator ratings (Barnhurst and Mutz, 1997). See example in Table 9. This

[7] We provide our detailed annotation guidelines on our GitHub: https://github.com/henniina/Detecting-paired-actions.

Metric	Annotator	Action							
		question	request	statement	accusation	challenge	acceptance	denial	appreciation
Average	Annotator1	5.26	4.33	5.01	4.04	3.64	4.54	4.61	3.80
	Annotator2	4.88	4.27	4.88	3.70	3.60	4.06	4.03	3.31
	Annotator3	4.74	3.96	4.87	4.04	3.59	4.10	3.50	4.20
	Overall	4.96	4.19	4.92	3.93	3.61	4.23	4.05	3.77
Correlation	Annotator1&2	0.86	0.64	0.50	0.38	0.22	0.66	0.38	0.41
	Annotator2&3	0.92	0.75	0.60	0.61	0.44	0.73	0.60	0.62
	Annotator3&1	0.87	0.74	0.56	0.42	0.27	0.72	0.46	0.45

Table 6: Label score averages by annotator and overall, and score correlations per action between annotators.

Annotator	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 ground truth labels.

was due to many comments including more than one prevalent action, and we judged that forcing annotators to label only one would reduce annotation quality (see Table 4). We saw that the signal for an action could be present at different scales of strength, which Likert scores could reflect. Confidence scores, for example, have been found useful for 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 label distribution. We set score 0 as designating the absence of an action, as annotators found this most intuitive, meaning 6 is the highest score.

4.1 Annotation Quality and Agreement

We measured our annotation quality comparing the 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 individual action (Table 5), 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).

For more detailed insights, we evaluated the annotators' annotation performances against ground truth labels (see section 5.1), in Table 7. These demonstrate some differences among annotators, some achieving higher performances on specific actions than others. We also investigated score means across actions, and annotation score correlations between pairs of annotators (see Table 6). Overall, correlations between Annotator2 and 3 tended to be higher.

5 Methods

In this section, we will describe how we set our ground truth labels, which ML models we used in our experiments, how we evaluated our models, and how we designed our experiments to answer our RQs.

5.1 Ground Truth

For computational modeling, we needed to define a meaningful ground truth against which to compare our models. Thresholding has been found useful in earlier studies when using confidence scores in annotation (Xu et al., 2017), and was relevant here due to our use of scores. We set a threshold $\theta = 3$ based on theoretical insights, and our empirical interests. Although our goal was to represent annotators' confidence in labels as well as their assessment of label strength as closely as possible, scores of 1-2 were rarer (see Appendix C), and based on theoretical considerations (Stivers and

Rossano, 2010) and empirical insights (see section 6.1), we considered scores 3-6 as representing a medium to strong signal that online discussion participants would likely consider relevant, requiring a response if the action usually involves normative expectations. Thus, for analyzing responding, we consider this a useful threshold. For a deeper understanding of how thresholding might affect classification, we conducted an additional (suggestive) threshold test: see Appendix B.

To build a set of ground truth labels, we mapped annotation scores back to binary labels (1: action present, 0: action not present). For single annotations, the label was 1 if the annotation score was $\geq \theta$. For multiple annotations, we use *conservative* (if even one annotator had given a score $\geq \theta$, the label was 1), and *relaxed ground truth labels* (if at least two annotators had given a score $\geq \theta$, the label was 1).

5.2 ML Models

Recent action classifiers rely on sentence transformers, other neural networks (Ghosh and Ghosh, 2021; Taniguchi et al., 2020; Joty and Mohiuddin, 2018), or few-shot learning (Casanueva et al., 2020). Studies emphasize the relevance of linguistic features (e.g., lexical and collocations) (Stolcke et al., 2000; Ferracane et al., 2021; Zakharov et al., 2021), and the use of pretrained word embeddings (Joty and Mohiuddin, 2018). Language Models (LM) like BERT have provided promising results in recent research (Taniguchi et al., 2020), and Generative LMs are receiving a lot of attention. Transfer learning techniques, then again, have proven useful for low-resource settings (Pamungkas et al., 2020). We thus chose to compare the performances of 1.) SVM, 2.) FinBERT, 3.) SetFit (with word embedding FinBERT), 4.) SetFit (sentence embedding FinBERT), 5.) Llama2 0-shot, 6.) Llama2 (finetuned), and 7.) FinGPT 0-shot.

As using separate classifiers for sub-tasks has proven effective in recent related work (Ferracane et al., 2021; Zakharov et al., 2021), we decided to train a separate classifier for each action following Ferracane et al. (2021), instead of training one multi-label ML classifier (Wu et al., 2019; Xu et al., 2017). The models predict whether a text contains an action or not.

To allow a comparison of our results to earlier research, we trained SVM models – often used in earlier action detection – similarly to previous models used for 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 (10 dimensions), balanced class weighting, and Grid Search for hyperparameter optimization[8] with sci-kit learn (Pedregosa et al., 2011). We applied preprocessing (tokenization and lemmatization with spaCy, following Haverinen et al., 2014), leaving in stop words as their removal negatively impacted performance. We were interested in whether weaker models, like SVM, would also show improved action classification performance if allowing multiple actions labeled per comment.

We used FinBERT, as BERT-derivatives haved proved their capacity in text classification (Devlin et al., 2019; Arabadzhieva-Kalcheva and Kovachev, 2022), also for asynchronous CMC (Davidson et al., 2020; Guo and Sarker, 2023), and DA classification (Taniguchi et al., 2020). We finetuned the Finnish pretrained FinBERT (bert-base-finnish-cased-v1)[9] (Virtanen et al., 2019) based on cased word embeddings (henceforth FinBERT1). Our data is highly imbalanced, so we used class weighting.

Besides standard fine-tuning, we used Setfit, a transfer learning technique that can also be initiated with Finnish LMs. It is based on fine-tuning a pretrained model with sentence pairs, then training a classifier based on finetuned 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 the training set. We initialized SetFit both with FinBERT1 (Virtanen et al., 2019) and FinBERT based on cased sentence embeddings (sbert-cased-finnish-paraphrase)[10] (henceforth FinBERT2) as underlying models. Both thus used a pretrained model to provide embeddings for comments.

We used Optuna (Akiba et al., 2019) for hyperparameter optimization, including learning rate, number of epochs, and batch size.

We tested the performance of Llama2 (Touvron et al., 2023) and FinGPT (Finnish GPT-3 large)[11] (Luukkonen et al., 2023) due to the recent popularity of Generative LMs. We used an uncensored version of Llama2 (llama2-7b-chat-uncensored)[12], because it is accessible and free to use, and as many comments included e.g. offensive language use. First, we used it without finetuning, i.e., as a 0-shot model. We prompted the models to decide whether each comment included a specific action or not. On our GitHub page, we provide a summary of prompts and how generated output was interpreted, matching key words in the generated output to extract predictions. With Llama2, we used the PyPi translators library (v5.9.2, GPLv3 license) to translate our data to English, because Llama models have shown to work much better for English (Wendler et al., 2024), and as Finnish is a low-resource language for Llama. We sought to reduce problems resulting from

[8] Optimizing kernel, regularization parameter C, and degree.

[9] https://huggingface.co/TurkuNLP/bert-base-finnish-cased-v1

[10] https://huggingface.co/TurkuNLP/sbert-cased-finnish-paraphrase

[11] https://turkunlp.org/gpt3-finnish

[12] https://huggingface.co/georgesung/llama2_7b_chat_uncensored

Model ID	Single label	Multi-label	Single annotation	Multiple annotation
SVM-1-act	✓		✓	
SVM-MS		✓	✓	
SVM-Avg		✓		✓
FinBERT1-1act	✓		✓	
FinBERT1-MS		✓	✓	
FinBERT1-Avg		✓		✓
SetFit-FinBERT1-1-act	✓		✓	
SetFit-FinBERT1-MS		✓	✓	
SetFit-FinBERT1-Avg		✓		✓
SetFit-FinBERT1-PNC		✓		✓
SetFit-FinBERT1-A1		✓	*	
SetFit-FinBERT1-A2		✓	*	
SetFit-FinBERT1-A3		✓	*	
SetFit-FinBERT2-1-act	✓		✓	
SetFit-FinBERT2-MS		✓	✓	
SetFit-FinBERT2-Avg		✓		✓
FinGPT-0shot-1-act	✓		✓	
FinGPT-0shot-MS		✓	✓	
FinGPT-0shot-Avg		✓		✓
Llama2-0shot-1-act	✓		✓	
Llama2-0shot-MS		✓	✓	
Llama2-0shot-Avg		✓		✓
Llama2-finetuned-1-act	✓		✓	
Llama2-finetuned-MS		✓	✓	
Llama2-finetuned-Avg		✓		✓

Table 8: Model IDs with related configurations (*=these utilize the full annotated dataset, but only the annotations made by one single annotator).

hallucination by using the outlines library[13] (Willard and Louf, 2023), which forces the model to choose between selected classes. We also finetuned a quantized Llama2 for our task using Adapters (Poth et al., 2023), the QLoRA approach[14] by Dettmers et al. (2024) and the bitsandbytes library[15] (MIT license), similarly to Dettmers et al. (2024). For Finnish, we used FinGPT without further finetuning (0-shot).

5.3 Data Splits and Evaluation

We divided our data into train, validation (for hyperparameter tuning) and test sets using sci-kit learn train test split twice, with respective set sizes 60%, 20% and 20%. Input data included individual comments for all models and labels.

Model evaluation included accuracy and Macro-F1 – we report F1 due to class imbalances. We report Jaccard coefficient scores when evaluating which computational approach might best predict all possible labels.

5.4 Experiments

To answer RQ1, we compared all models using 3 configurations to compare the effects of considering one vs. multiple actions: a.) a *single action (1-act)*, b.) *multilabel single-annotation (MS)*, and c.) *averaged (Avg)* approach. See a depiction of the configurations and model IDs in Table 8. Although model simplicity is often preferred, we wished to test for possible gains in performance with increased model complexity.

The *1-act* configuration used single annotations, setting for each comment the action with highest score among all labels as positive if score $\geq \theta$, other actions as negative. In case two actions (or more) had equal scores, we assigned a positive label randomly. In our view, this corresponded to annotator decision-making, as in annotation we often had to randomly choose the primary action if having to decide only one most important action between actions with the same signal strength. We compared the *MS* and *Avg* configurations to the *1-act* model to test whether allowing more than one action will statistically improve performance. Other configurations allow multiple actions labeled.

The *MS* configuration used single annotations, allowing multiple actions. *Avg* utilized multiple annotations to decide an average annotation score for label decision, $score_{action} = (a_1 + a_2 + a_3)/3$, with a a reference to $AnnotationScore_i$ for the action. It was included as it has been one of the most popular approaches to dealing with disagreements.

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).

In relation to RQ2, using SetFit-FinBERT1, which had achieved the highest number of best performances in relation to RQ1, we compared three configurations for leveraging annotator disagreements to find the most suitable approach for representing multiple interpretations of actions: a.) *averaging (Avg)* (following Uma et al. (2022), b.) *positive/negative/complicated (PNC)* (Jiang and de Marneffe, 2022), and c.) *individual annotator (Annotator1-3)* approach (Ferracane et al., 2021).

Avg and *PNC* were included as the approaches have often been used with multiple annotations (Uma et al., 2022; Jiang and de Marneffe, 2022). With *PNC*, per 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. This decision boundary differs from Jiang and de Marneffe (2022), but as they had 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. We treated the task as a 3-way classification problem. Individual annota-

[13] https://github.com/dottxt-ai/outlines

[14] https://github.com/Adapter-Hub/adapters/blob/main/notebooks/QLoRA_Llama_Finetuning.ipynb

[15] https://github.com/TimDettmers/bitsandbytes

tor models have been fruitful for representing multiple interpretations, by predicting all possible annotations (Davani et al., 2022; Ferracane et al., 2021). These were trained using each individual annotator's annotations only: Annotator1 (*A1*), Annotator2 (*A2*), and Annotator3 (*A3*). All configurations were tested against conservative and relaxed ground truth.

We wished to find a model that could predict all possible annotations. Thus, we compared how different model ensembles 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 comment: $JaccardSimilarityJ(A, B) = \frac{|A \cap B|}{|A \cup B|}$.

Finally, we compared differences between actions by investigating average label scores per action, and annotation and class correlations by using scipy.stats package with Kendall rank correlation (Schober et al., 2018), visualized with heatmaps using the seaborn library. Kendall correlation was used as normal distribution could not be assumed and as it is often considered more robust in contrast to Spearman's (Schober et al., 2018).

6 Results

Next, we will discuss our results to answer our RQs: RQ1.) whether a multi-label approach better aligns with how actions are performed in our empirical context, and RQ2.) what approach can best represent the multiple valid interpretations relevant to how people express actions in asynchronous conversations.

6.1 RQ1: One Action or Multiple Actions

We asked whether a multi-label approach would better align with how actions are performed in our asynchronous data in contrast to a single-label approach. This was addressed by both 1.) comparing these approaches during the annotation scheme development, and negotiating and examining annotation interpretations, and 2.) testing whether adopting a single-label approach in contrast to a multi-label approach would significantly affect classification performances.

First, turns in our asynchronous conversation data tended to utilize actions somewhat differently as compared to face-to-face conversation, and synchronous online conversation.

In Table 9, we see a conversation extract from our data. Examining comments 2-4, we can see that participants in our online conversation data utilized paired actions to build coherence across turns. In (2), user A asks a question that can be interpreted as *challenging*. In (3), B does not provide information in response to the *challenging question*, but can be seen to *challenge* its grounds. Annotators 2 and 3 interpret the response, as well as comment (2), as containing a *challenge*. The extract illustrates how turns may often perform multiple actions at the same time at different scales of strength – they might contain an action responding to a previous turn (completing an action pair), and other actions, potentially also initiating new action pairs. Comment (4) is a case in point. Also, actions are not neatly organized into separate sentences or paragraphs. For example, in comment (2), the interpretation of the comment containing a *challenge* arguably cannot be reduced to a single sentence, but is derived from a more holistic reading. This was common for comments in our data as users often formed actions that overlapped. For example, this could be seen in our data when one action was used as a vehicle for another action, one part of the comment (e.g., a sentence) performed multiple actions (e.g., stating and accusing) (cf. double-barreled actions or composite actions: Rossi, 2018; Levinson, 2013; Schegloff, 2007: pp. 73–78), or the whole comment included multiple (overlapping) actions.

Users often tended to treat a previous turn as having some main action(s), and they did not necessarily respond to all actions present in a previous turn. This phenomenon has also been described in CA, though in face-to-face conversations with shorter turns (e.g., Rossi, 2018; Schegloff, 2007: pp. 73–78). Based on our data, it seems that with comments that contain multiple actions, users might treat 1-2 actions as the 'main job(s)' of a previous turn. For example, in comment (5), B does not address the accusations (that had lower scores) made in comment (4) by A, but chooses to respond to the *question* or *request* instead.

These insights illustrate how modeling actions by adopting a multi-label approach better aligns with how actions in this context are used, in contrast to the single-label approach. They also emphasize how score-based annotation helps to account for overlapping actions, and to represent signal strength, i.e. how some actions are more prevalent in a turn than others.

Next, to test how limiting predictions to a single action per comment would affect classification, we compared *1-act* models to models allowing multiple actions per comment: see Table 10.

Considering multiple actions vs. one, SVM achieves statistically improved performances for only two classes at best. With *FinBERT1-Avg* the increase in performance is statistically significant for only one action. *SetFit-FinBERT1* fares much better: *SetFit-FinBERT1-MS* achieves statistically higher performance in classifying three actions. For the remaining actions, despite higher F1s, statistical tests do not show $p < 0.05$. *SetFit-FinBERT1-Avg* fares better, achieving a statistically significant improvement in performance for five actions. *SetFit-FinBERT2-MS* achieves

User	Comment	Annotations		
		Annotator1	Annotator2	Annotator3
(1) A	*B*, The country hasn't attacked. I'm speaking from personal experience. I moved to Russia for a year when I was 17. In other occasions I've been there hundreds of times, both for fun, and driving a bus. Nothing has ever happened, the people are friendly and you can always get help. It's a different story in Finland.	statement:6 denial:6 apprec.:2	statement:5 challenge:5 denial:3	challenge:5 statement:4 denial:3
(2) A	*C*, war is in no way acceptable in any form. But these issues aren't so black and white. How much more often have you and *B* visited Ukraine and spent time there compared to myself?	question:6 statement:3	question:6 statement:6 challenge:4	question:6 statement:6 challenge:4
(3) B	*A*, Dear *A*. How many times each of us has traveled to or lived in Russia, well that has no bearing on whether Russia is a rogue state ruled by a dictator. It is, believe it or not. It's that black-and-white. I hope you'll get over your gullibility someday. Your actions only support Putin's dictatorship.	accusation:6 statement:3	statement:6 request:4 challenge:4	statement:6 request:4 challenge:3
(4) A	*B*, Well tell me about your own experiences, traveling in Ukraine, have you got an impression that it's somehow a more virtuous country? Surely you too have seen those Nazi arguments in the media there. Children have been killed by Ukraine for 8 years. For example, Ukrainian authorities did not let me into the Donetsk region at all. The answer to my question 'why' was simply 'because' and that's that. They were probably afraid I was a journalist. Still, it's a wonderful country, too, and the citizens. After all, Finland is also a country that welcomes all kinds of scum and yet we also love our own homeland.	accusation:6 question:6 apprec.:4	statement:6 question:5 request:5 accusation:2	statement:6 question:5 request:5 accusation:1
(5) B	*A*, Again, my travel experiences aren't in any way related to the actions of any government. Parroting it won't change anything. If your next claim is that it's right for the Ukrainians that the wonderful country of Russia invaded Ukraine, then good luck with that.	statement:6	statement:5 challenge:5 denial:3 accusation:2	statement:5 challenge:5 denial:4 accusation:1

Table 9: Annotated extract from a Ukraine war related conversation.

Model	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
SVM-MS	single	0.66	0.58	0.63	0.62*	0.52	0.69	0.59	0.51
SVM-Avg	multiple	0.63	0.60	0.62	0.62*	0.53	0.69**	0.59	0.51
FinBERT1-1-act	single	0.86	0.63	0.70	0.57	0.54	0.65	0.57	0.62
FinBERT1-MS	single	0.87	0.69	0.73	**0.76**	0.61	0.66	0.56	0.65
FinBERT1-Avg	multiple	0.85	0.70	0.77**	0.63	**0.68**	0.62	0.51	0.75
SetFit-FinBERT1-1-act	single	0.81	0.60	0.66	0.52	0.50	0.59	0.55	0.50
SetFit-FinBERT1-MS	single	0.94	**0.82***	0.77	0.73*	0.64	**0.86***	**0.68**	**0.78**
SetFit-FinBERT1-Avg	multiple	**0.97***	0.80**	0.78**	0.65*	0.63	0.79**	0.58	0.72
SetFit-FinBERT2-1-act	single	0.76	0.59	0.66	0.52	0.55	0.57	0.49	0.50
SetFit-FinBERT2-MS	single	0.95	0.69	0.76	0.56	0.64*	0.76*	0.66*	0.60
SetFit-FinBERT2-Avg	multiple	0.95*	0.67	**0.81***	0.58	0.54	0.71	0.61	0.53
FinGPT-0-shot-1-act	single	0.29	0.25	0.48	0.21	0.23	0.24	0.24	0.19
FinGPT-0-shot-MS	single	0.34	0.31	0.51	0.33	0.40	0.29	0.30	0.25
FinGPT-0-shot-Avg	multiple	0.37	0.24	0.51	0.31	0.40	0.30	0.30	0.20
Llama2-0-shot-1-act	single	0.47	0.49	0.29	0.48	0.49	0.49	0.46	0.55
Llama2-0-shot-MS	single	0.47	0.54	0.37	0.46	0.48	0.50	0.48	0.50
Llama2-0-shot-Avg	multiple	0.50	0.52	0.32	0.43	0.41	0.51	0.47	0.52
Llama2-finetuned-1-act	single	0.62	0.51	0.51	0.55	0.48	0.52	0.51	0.49
Llama2-finetuned-MS	single	0.71	0.61	0.80	0.56	0.47	0.81**	0.62	0.57
Llama2-finetuned-Avg	multiple	0.57	0.67*	**0.81***	0.65	0.64	0.69	0.39	0.52

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

significantly higher performances for 3 actions, *SetFit-FinBERT2-Avg* for 2. Generative LMs do not perform very well: only *Llama2-finetuned* achieves notably higher performances when allowing multiple actions in contrast to *Llama2-finetuned-1-act*, but only for *acceptances* (*Llama2-finetuned-MS*), and for *requests* and *statements* (*Llama2-finetuned-Avg*).

To answer RQ1, we conclude that considering multiple actions – and using score-based annotation – to classify actions much better aligns with how people utilize actions in asynchronous online conversation in contrast to the single-label or categorical annotation approach. This is supported by our insights gained during annotation scheme development, examination of example cases, and model performance tests where our best models achieved significantly better performance for many actions with the multi-label approach.

6.2 RQ2: Modeling Ambiguity

To answer RQ2, we utilized multiple annotations and three different approaches to leverage them. We used our best-performing model from previous experiments, SetFit-FinBERT1, in order to discover how to best represent the ambiguity related to actions in our data. To this end, we compared different modeling configurations for leveraging multiple annotations, analyzed label correlations in more detail, and examined extracts from our data to see how models performed in comparison to manual annotations for these cases.

Performances for models where multiple annotations were utilized can be seen in Table 11. Except for *PNC*, performances are evaluated using both conservative and relaxed ground truth.

SetFit-FinBERT1-PNC performs quite poorly. *PNC* models performed worst for the Complicated class, similarly to Jiang and de Marneffe (2022), perhaps due to class heterogeneity; other classes' performances are higher. *A2* models performed best overall if considering which models achieved most of the highest F1 scores. Macro-F1s were higher for some actions when predicting relaxed ground truth labels. However, here results differ according to action. Also, *A1-A3* model performances differed according to annotator.

Overall, for most actions, the models utilizing multiple annotations, or individual annotators' annotations, achieved higher performances than the models using only single annotations (see Tables 10 and 11) – except for *requests* and *acceptances*, for which *SetFit-FinBERT1-MS* had higher F1s.

Comparing the performances against earlier work modeling similar actions (Paakki et al., 2024), we achieved higher performances for all actions, except for *appreciations*. We also achieved higher F1s for some actions modeled in prior work on asynchronous data: message-level detection of *request* and *question* trained with BERT (Taniguchi et al., 2020), *question* detection (and overall performance) in contrast to Zhang et al. (2017), and better agreement for *challenges* in comparison to Bracewell et al. (2012). However, the comparability to the latter three works is limited due to differences in theoretical frameworks, data, annotation practices or modeling choices.

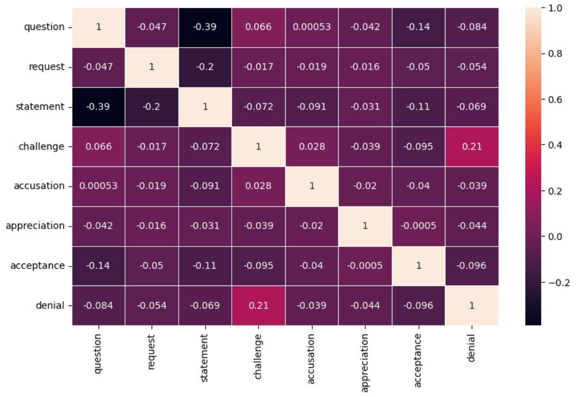

Figure 1: Correlation scores between all 8 actions.

Differences between actions deserve some attention: overall, classifiers for predicting *accusations*, *challenges*, *denials* and *appreciations* seem to have lower performance compared to other actions. The latter three also have notably lower Micro-F1 scores when investigating performances for the positive class for the *SetFit-FinBERT1-A2* model (see Appendix A). Inspecting annotator performances against ground truth (Table 7), performances for face-threatening actions – *accusations*, *challenges* and *denials* – are notably lower, although there are differences when using conservative *vs.* relaxed ground truth. Label score averages, in Table 6, also showed that *accusations*, *challenges*, *denials*, and *appreciations* have lower label scores overall. Measuring score correlations between annotators, by action, we see (in Table 6) that *accusations*, *challenges*, and *denials* have some of the lowest correlations. *Appreciations* also have low correlations between Annotator1 and 2, and Annotator1 and 3. There were very few *appreciations* in the data, which might affect the results. Also, *statements* have surprisingly low annotation score correlations. These might be partly due to *statements* being much more frequent in comments in contrast to other actions (see Table 5): present in 71% of all comments (in single annotations; 83% in multiple). Trained model performances for *statements*, on the other hand, are quite good.

Investigating the correlations between all annotated action labels for each comment (see Figure 1), there is a weak positive correlation (Schober et al., 2018) with $p < 0.001$ between *challenges* and *denials*, and a weak positive correlation between *questions* and *chal-*

Model	**Ground truth**	**Action**							
		question	request	statement	accusation	challenge	acceptance	denial	appreciation
SetFit-FinBERT1-PNC	cross-e.	0.71	0.55	0.57	0.50	0.50	0.59	0.47	0.39
SetFit-FinBERT1-Avg	conservative	0.96	0.74	0.74	0.65	0.54	0.73	0.58	0.72
SetFit-FinBERT1-A1	conservative	0.92	0.75	0.75	0.51	0.43	0.77	0.55	0.65
SetFit-FinBERT1-A2	conservative	0.95	0.77	**0.84**	0.72	**0.76**	0.78	**0.68**	0.66
SetFit-FinBERT1-A3	conservative	**0.97**	0.71	0.78	0.66	0.52	0.74	0.63	0.69
SetFit-FinBERT1-Avg	relaxed	**0.97**	**0.80**	0.76	0.65	0.63	0.79	0.56	0.65
SetFit-FinBERT1-A1	relaxed	0.95	0.77	0.79	0.60	0.52	**0.82**	0.62	0.52
SetFit-FinBERT1-A2	relaxed	**0.97**	0.76	0.79	**0.75**	0.61	0.78	**0.68**	**0.73**
SetFit-FinBERT1-A3	relaxed	0.96	0.71	0.78	0.66	0.52	0.74	0.63	0.68

Table 11: 10-fold cross-validated macro-F1 scores for SetFit-FinBERT1, utilizing different configurations for 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.
conservative	0.56	0.69	0.63	0.39	0.67	0.70	**0.72**	**0.72**	0.51	0.60	0.56	0.61	0.58	0.62	0.62
relaxed	0.54	0.50	0.52	**0.62**	0.52	0.48	0.48	0.48	0.55	0.46	0.50	0.45	0.50	0.45	0.44
	A1 +PNC	A2 +PNC	A3 +PNC	Avg. +PNC	A1+A3 +PNC	A1+A2 +PNC	A2+A3 +PNC	A1+A2 +A3 +PNC	A1 +Avg. +PNC	A2 +Avg. +PNC	A3 +Avg. +PNC	A1+A2 +Avg. +PNC	A1+A3 +Avg. +PNC	A2+A3 +Avg. +PNC	A1+A2+ A3+Avg. +PNC
conservative	0.46	0.55	0.51	0.35	0.52	0.55	0.56	0.56	0.46	0.54	0.50	0.54	0.51	0.62	0.55
relaxed	0.48	0.40	0.44	0.45	0.45	0.40	0.40	0.40	0.47	0.40	0.43	0.39	0.43	0.44	0.39

Table 12: Jaccard coefficient scores for ensemble models using SetFit-FinBERT1. A1=Annotator1 model, A2=Annotator2, A3=Annotator3, Avg=Averaged, PNC=Positive/Negative/Complicated.

lenges, only slightly above the lower bound of weak correlation ($p < 0.001$). For most actions, there is a weak or negligible negative (or positive) correlation magnitude according to the statistic with no statistical significance. There is a moderate negative correlation between *questions* and *statements* with $p < 0.001$.

Finally, comparing model ensembles for predicting all possible annotations for comments, in Table 12, an ensemble of *A1+A2* or *A1+A2+A3* fared best for conservative ground truth, according to Jaccard coefficient scores. *Avg* fared best with relaxed ground truth labels. However, for relaxed ground truth Jaccard similarity drops notably in contrast to conservative ground truth.

A further investigation of difficult cases and comparison between human annotations and model predictions provides more insight. Tables 13–14 illustrate what annotation scores human annotators gave to some of the comments with annotation disagreement, and what our best models predicted as actions present in the comments. For the second column, we only list actions that at least one annotator deemed to be present in the comment, and for the third column, actions predicted as being present by at least one model. For model predictions, we always used a model trained on a data split that did not include the to-be-classified comment.

There was disagreement between humans and models in some cases, e.g. with *challenges* and *accusations*, but model disagreements correspond with human disagreements quite well if investigating them side-by-side with our best models. Many comments required interpretation of implicit meaning: e.g. ambiguity, contextual knowledge, proverbs or even poetic expression. This seemed to often result in multiple potential interpretations. A case in point is B's comment in Table 13, which was interpreted by human annotators as either sarcastic, a *challenge* (or *accusation*) presented together with a *statement*, or just *statement*. A less likely interpretation would be sincere *accepting statement*, not chosen by human annotators. Similarly, ”hallelujah” in Table 1 is ambiguous, interpreted by human annotators as either sincerely appreciative, or sarcastic, a *challenge* merged with a *statement*. Models predicted it as a *statement* (A1, A2, A3, Avg.), a *challenge* (A2), but also as an *appreciation* (A2). Another ambiguous example is B's comment in Table 14. It depends on subjective interpretation whether hindsight can be seen as morally repre-

Comments	**Annotations (Annotator: score)**	**Predictions (Model IDs)**
A: What other nation could be responsible for Russia's invasion of Ukraine besides the Russian people? All else is just about making excuses. (27 likes)	*question* (A1: 5, A2: 4, A3: 3) *challenge* (A2: 5, A3: 5) *statement* (A1: 0, A2: 0, A3:1) *accusation* (A1: 3, A2: 0, A3: 0)	*question* (A1-3, Avg) *challenge* (A2, A3) *statement* (A3)
B: yea the people did attack there on their own initiative, without any government authority and orders. (2 likes)	*statement* (A1: 0, A2: 0, A3: 3) *accusation* (A1: 0, A2: 0, A3: 2) *challenge* (A1: 4, A2: 5, A3: 5)	*statement* (A1-3, Avg) *accusation* (A2)

Table 13: Extract from Ukraine war discussion on Russian citizens' role in the war, HS 10/2022. Second column lists annotation scores by annotators, third column the SetFit-FinBERT1 model IDs predicting that listed action exists for the comment in first column.

Comments	**Annotations (Annotator: score)**	**Predictions (Model IDs)**
A: Would it have been worth taking an interest in things before these men came to power? (21 likes, 2 angry emojis)	*question* (A1: 3, A2: 5, A3: 3) *accusation* (A1: 2, A2: 4, A3: 3)	*question* (A1-3, Avg.) *accusation* (A3)
B: *A*, Well here we have a real master of hindsight. (Dog sticker facepalming) (15 likes)	*statement* (A1: 0, A2: 3, A3: 5) *accusation* (A1: 0, A2: 3, A3: 5) *challenge* (A1: 3, A2: 0, A3: 0)	*statement* (A1-3, Avg.) *accusation* (A2)
A: *B*, hindsight is a sweet thing… (3 likes, 1 angry)	*statement* (A1: 3, A2: 4, A3: 5) *accusation* (A1: 0, A2: 0, A3: 3) *appreciation* (A1: 0, A2: 3, A3: 0)	*statement* (A1-3, Avg.) *accusation* (A2)

Table 14: Extract from Ukraine war discussion on how Russian soldiers are treated, YLE 12/2022. Second column lists annotation scores by annotators, third column the SetFit-FinBERT1 model IDs predicting that listed action exists for the comment in first column.

hensible. Not all annotators agreed here. Model predictions corresponded quite well with human annotators' interpretations, and their disagreements. However, for some cases, like Tables 13 and 14, where human annotators disagreed on whether a comment should be labeled as an *accusation* or a *challenge*, models did not predict the ambiguous comment as a challenge.

Overall, an ensemble of individual annotator-based models seems to best align with the human annotations, also quite well representing many of the disagreements. Based on the Jaccard coefficient scores and the above empirical insights, we conclude in response to RQ2 that an ensemble of individual annotator based models best represents the multiple possible interpretations of actions in our data.

7 Discussion

We investigated how paired actions in comments to asynchronous crisis-related conversations could be computationally modeled, especially face-threatening actions central to misbehavior like trolling. We illustrated that, in this context, it is important to take into account multiple possible label interpretations, and multiple actions often performed in one comment. The conversational context we studied is largely assertive in nature, statements being notably more common than other actions. As we assumed, face-threatening actions like *accusations* and *challenges* are also common, which highlights the need for including them in computational models used for identifying actions in these conversations.

We showed that allowing multiple actions predicted per comment helps to align models with the empirical phenomenon studied, and statistically improves model performance. SetFit models (initialized with FinBERT + word embeddings) performed best overall. We also demonstrated improved classification performance in contrast to earlier action detection in asynchronous conversations. Likert score annotation allowed us to consider the presence of actions on different scales of strength, rather than categorically (cf. Glickman and Dagan, 2005). We illustrated that this was important for representing the main functions of a comment. Based on our empirical insights, signal strength affected how paired actions and their norms were treated by discussion participants: in the case of multiple (pair-initiating) actions, participants tended to orient to (1-2) main actions as having the strongest normative expectations for responding (see Table 9).

We demonstrated that for predicting all possible labels, with conservative ground truth (positive label if even one annotator gave a score ≥ 3), an ensemble of models trained separately on individual annotators' annotations performed best. Based on our results, we highlight that although averaging has been popular in accounting for annotation disagreements, it is not always the best option as it might lead to losing important information on multiple possible interpretations of labels. Generative LMs performed quite poorly on the action detection task, except for specific classes (*statements* and *acceptances*) with Llama2-finetuned. In other words, these models seemed to recognize a limited amount of categories even when finetuned for the downstream task.

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

We illustrated some differences between actions: e.g., for *questions*, our models reached high performances, but for some others performances were lower. An interesting result found was that face-threatening actions (*challenge*, *denial*, *accusation*) were more difficult to annotate and/or model than others. They also involved more disagreement between annotators than other actions. This is in line with theoretical views on social actions: 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. *Appreciations*, surprisingly, also portrayed a lower agreement. This might have been due to the low number of appreciations in our data; or they might have been expressed in an ambiguous manner. Whether this is a context-specific tendency (e.g. platform or cultural context) remains an open question.

We saw correlation between *challenges* and *denials*, and to some extent between *questions* and *challenges*. This might be because *challenges* and *denials* were at times difficult to distinguish from each other. We may have needed more detailed clarification of class boundaries here. However, the annotation scheme (and guideline) development process was already very extensive. Also, based on examples seen during the development, *denials* were often followed by *challenges* of epistemic claims (or of interlocutors' positions), both in the same comment. This might explain the correlation. Finally, there seemed to be a significant amount of *challenging questions* in our data, which could explain the correlation between these two actions.

We found that users tended to treat a previous turn as having some main action, not necessarily responding to all actions. Given that in CA action has been viewed as the 'main job' of a turn, as turns in asynchronous conversation are longer and often involve multiple actions, it is interesting how users interpret the 'main job' of such turns. Although there are many factors that may affect the interpretation of actions and required responses (Stivers and Rossano, 2010), based on empirical insights, it seems that the strength of label signal can affect how strong normative expectations pair initiating actions might incur on subsequent turns. Thus, actions that are present with at least moderate strength of signal will likely be fruitful for analyzing responding behaviors in relation to paired actions (and face-threats).

From a (digital) CA perspective, it is challenging to systematically identify actions in asynchronous conversations due to actions often 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. Ambiguity in the expression of (face-threatening and other) actions might also be a strategic choice. Thus, we consider it crucial to be able to model ambiguity in the expression of actions, especially face-threatening actions, when studying crisis-related or manipulative asynchronous conversations.

8 Conclusions and Future Work

To conclude, we investigated how to approach action detection in asynchronous crisis news conversations. Our computational approach was able to reflect how multiple actions were performed within the same comment, and the ambiguity related to actions, with improved classification performance in contrast to earlier action detection for asynchronous conversation. Although annotator disagreements have been studied increasingly in NLP, there is still room for exploring how to utilize them in the analysis of face-threatening and paired actions. The contributions of this paper included 1.) portraying and modeling disagreement in the annotation of actions in asynchronous conversations in Finnish, 2.) a paired action annotation framework and dataset for asynchronous conversation including face-threatening actions, and 3.) models with improved classification performance for many of these actions. Although our study focuses on Finnish, the framework can be applied for other languages as well.

Future work could investigate how to further address the differing nature of actions, and how to further utilize annotation scores and contextual information in more fine-grained models.

We conclude that representing ambiguity in computational modeling is especially relevant to analyzing face-threatening actions. These actions, in turn, are crucial for the study of online misbehavior. Digital CA based approaches, having a robust theoretical understanding of such actions, can be fruitful for analyzing meaningful ambiguity related to how actions are performed in asynchronous conversations online.

8.1 Limitations

Crowdsourcing 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, 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 (Eickhoff, 2018). We had only three annotators; a higher number might lead to better results (Pavlick and Kwiatkowski, 2019). Due to resource limitations, we considered our current scope of annotators and data sufficient for now.

Score-based annotation allows fluidity in class interpretation, but subjective reading, confidence, signal strength, and understandings of scores and categories might be melted into one metric. Also, in contrast to message-level classification, we could have labeled segmented comments. However, we considered this challenging as sometimes the boundaries of actions were unclear or actions tended to overlap (e.g., Table 9). The boundaries of an action did not always correspond to sentence boundaries. Thus, we deemed message-level scoring best. Future research could explore the sentence segmentation option further. The choice of threshold (3) was theoretically and empirically motivated, but in some cases it might be relevant to be able to computationally identify even very weak action signals. We could have further investigated how to model conversational context or sequential dependencies more intricately (e.g. whether a comment responds to a certain action). However, these endeavors are best left for future research.

At this point we did not utilize context information when training models, which is a potential limitation since annotators had access to this information. However, we deemed it interesting to model the potential for different interpretations of a turn considered alone, through scores. We decided to do this because we were concerned that using previous turns (or their actions) as features would confound the computational analysis of responding behaviors. This is because people favor certain types of responses due to conversational norms (e.g. Stivers and Rossano, 2010). However, deviations from these norms can be highly meaningful (e.g. Paakki et al., 2024), so if models would overgeneralize normative responses as predicted actions for responses, this would confound computational analyses of less common deviant responses. Future work could expand on this to develop more contextually sensitive computational models that could also detect contra-normative responses.

Since we aimed at a simplified model, using applied digital CA, we could not strictly follow the analytical practices typical for CA. Restricting the granularity of the annotation scheme limits the interpretation of actions, forcing annotators (and thus models) to ignore some rarer actions. However, this process is important for a computational classification approach like ours. Secondly, it is debatable whether some of the classes in our applied approach are notably different in contrast to some DA classification approaches (e.g., Taniguchi et al., 2020). However, we see some of the added classes (e.g., accusations) as very different from those included in earlier DA models based on Speech Acts. Also, our key literature related to action pairs and normative expectations of face-threatening actions comes from CA (Schegloff, 2007; Turowetz and Maynard, 2010; Dersley and Wootton, 2000), which views these actions and paired actions somewhat distinctly, in our view. We wished to highlight these theoretical foundations in this paper.

Crisis related discussions may involve sensitive information, even when dealing with publicly available social media data. We have translated the examples, and anonymized and de-identified the data so that the content is conveyed without privacy concerns. We published a privacy notice according to Aalto University policy, regarding data collection and management, on our research project's website during the study.[16] We only release models and materials where any possibly sensitive information has been removed.

Acknowledgements

This research has been funded by the Academy of Finland project ”Reconstructing Crisis Narratives for Trustworthy Communication and Cooperative Agency”, project number 339931, and ”Liquid Forms of AI-Infused Disinformation: Improving Information Resilience in the Finnish News Ecosystem” project[17] funded by the Helsingin Sanomat Foundation. We utilized computer resources provided by the Aalto University School of Science ”Science-IT” project. We also used the Finnish CSC – IT Center for Science LTD.'s high-performance computing resources. We would like to thank Antti Salovaara, Minttu Tikka, Antti Ukkonen, Anna-Kaisa Jokipohja, Raul Vazquez and Eetu Mäkelä for their feedback and support during our research.

[16] The project has already ended, so the website is no longer operational, however. For project details, see: https://research.aalto.fi/en/projects/crisissawhney/

[17] https://blogs.helsinki.fi/disinformation-news-media/

References

Akiba, Takuya, 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.

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

Antaki, Charles, Michela Biazzi, Anette Nissen, and Johannes Wagner. 2008. *Accounting for moral judgments in academic talk: The case of a conversation analysis data session*. Walter de Gruyter.

Arabadzhieva-Kalcheva, Neli 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.

Austin, John Langshaw. 1975. *How to Do Things with Words*. Oxford university press.

Bakhtin, Mikhail Mikhaĭlovich. 1981. *The dialogic imagination: Four essays*, volume 1. University of Texas Press, Austin, Texas.

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

Bellutta, Daniele, Catherine King, and Kathleen M Carley. 2021. Deceptive accusations and concealed identities as misinformation campaign strategies. *Computational and Mathematical Organization Theory*, 27:302–323.

Bhatia, Sumit, Prakhar Biyani, and Prasenjit Mitra. 2014. Summarizing online forum discussions – can dialog acts of individual messages help? In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2131, Doha, Qatar. Association for Computational Linguistics.

Bracewell, David, Marc Tomlinson, and Hui Wang. 2012. Identification of social acts in dialogue. In *Proceedings of COLING 2012*, pages 375–390, Mumbai, India. The COLING 2012 Organizing Committee.

Bracewell, David B., 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.

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

Carvalho, Vitor R. and William W Cohen. 2005. On the collective classification of email "speech acts". In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 345–352.

Casanueva, Iñigo, 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.

Castro, Stephanie L. 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.

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

Clark, Herbert H and Edward F Schaefer. 1987. Collaborating on contributions to conversations. *Language and cognitive processes*, 2(1):19–41.

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

Cohen, William W., 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.

Davani, Aida Mostafazadeh, 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.

Davidson, Sam, Qiusi Sun, and Magdalena Wojcieszak. 2020. Developing a new classifier for automated identification of incivility in social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 95–101, Online. Association for Computational Linguistics.

Derrac, Joaquín, 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.

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

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLoRa: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

Devlin, Jacob, 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.

Di Mascio, Fabrizio, 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.

Duran, Nathan and Steve Battle. 2018. Conversation analysis structured dialogue for multi-domain dialogue management. In *The International Workshop on Dialogue, Explanation and Argumentation in Human-Agent Interaction (DEXAHAI)*.

Duran, Nathan, 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.

Eickhoff, Carsten. 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.

Enfield, Nick J and Jack Sidnell. 2017. On the concept of action in the study of interaction. *Discourse Studies*, 19(5):515–535.

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

Feng, Donghui, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. Learning to detect conversation focus of threaded discussions. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 208–215, New York City, USA. Association for Computational Linguistics.

Ferracane, Elisa, 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.

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

Fortuna, Blaz, Eduarda Mendes Rodrigues, and Natasa Milic-Frayling. 2007. Improving the classification of newsgroup messages through social network analysis. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 877–880, New York, NY, USA. Association for Computing Machinery.

Fuscone, Simone, 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.

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

Ghosh, Souvick 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.

Giglietto, Fabio, Nicola Righetti, Luca Rossi, and Giada Marino. 2020. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society*, 23(6):867–891.

Giles, David, 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.

Glickman, Oren 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.

Godfrey, John J., 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.

Goffman, Erving. 1974. *Frame Analysis: An Essay on the Organization of Experience.* Harvard University Press.

Guo, Yuting and Abeed Sarker. 2023. SocBERT: A pretrained model for social media text. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 45–52, Dubrovnik, Croatia. Association for Computational Linguistics.

Haverinen, Katri, 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.

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

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

Jeng, Wei, Spencer DesAutels, Daqing He, and Lei Li. 2017. Information exchange on an academic social networking site: A multidiscipline comparison on researchgate Q&A. *Journal of the Association of Information Science and Technology*, 68(3):638–652.

Jiang, Nan-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.

Joty, Shafiq 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.

Joty, Shafiq and Tasnim Mohiuddin. 2018. Modeling speech acts in asynchronous conversations: A neural-CRF approach. *Computational Linguistics*, 44(4):859–894.

Jünger, Jakob 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).*

Jurafsky, Dan. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical report.

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

Kim, Su Nam, Li Wang, and Timothy Baldwin. 2010. Tagging and linking web forum posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 192–202.

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

Laurenti, Enzo, Nils Bourgon, Farah Benamara, Mari Alda, Véronique Moriceau, and Courgeon Camille. 2022. Speech acts and communicative intentions for urgency detection. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 289–298, Seattle, Washington. Association for Computational Linguistics.

Levinson, Stephen C. 2013. Action formation and ascription. In *The Handbook of Conversation Analysis*, volume 1, pages 103–130. Wiley Online Library.

Lindell, Michael K. 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.

Linell, Per and Ivana Marková. 1993. Acts in discourse: From monological speech acts to dialogical inter-acts. *Journal for the theory of social behaviour*, 23(2):173–195.

Luukkonen, Risto, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726. Association for Computational Linguistics.

Mann, William and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

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

Meredith, Joanne. 2019. Conversation analysis and online interaction. *Research on Language and Social Interaction*, 52(3):241–256.

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

Moldovan, Cristian, 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.

Nie, Yixin, 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.

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

Paakki, Henna, 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.

Paakki, Henna, Heidi Vepsäläinen, Antti Salovaara, and Bushra Zafar. 2024. Detecting covert disruptive behavior in online interaction by analyzing conversational features and norm violations. *ACM Transactions on Computer-Human Interaction*, 31(2):1–43.

Pamungkas, Endang Wahyu, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Passonneau, Rebecca J., 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.

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

Pedregosa, Fabian, 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.

Peterson, Joshua C., 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.

Plank, Barbara. 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.

Pomerantz, Anita. 1984. *Pursuing a response.* Cambridge University Press.

Poth, Clifton, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore. Association for Computational Linguistics.

Qadir, Ashequl and Ellen Riloff. 2011. Classifying sentences as speech acts in message board posts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 748–758, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Rezapour, Rezvaneh, 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.

Rossi, Giovanni. 2018. Composite social actions: The case of factual declaratives in everyday interaction. *Research on Language and Social Interaction*, 51(4):379–397.

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

Salonen, Margareta, Margarethe Olbertz-Siitonen, Turo Uskali, and Salla-Maaria Laaksonen. 2022. Conversational gatekeeping—social interactional practices of post-publication gatekeeping on newspapers' facebook pages. *Journalism Practice*, pages 1–25.

Savolainen, Reijo. 2020. Dialogue processes in online information seeking and sharing: a study of an asynchronous discussion group. *Information Research*, 25(3).

Savy, Renata. 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).

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

Schober, Patrick, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Stivers, Tanya. 2013. *Sequence organization*. Wiley-Blackwell: Blackwell Publishing Ltd.

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

Stivers, Tanya and Federico Rossano. 2010. Mobilizing Response. *Research on Language & Social Interaction*, 43(1):3–31.

Stolcke, Andreas, 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.

Stommel, Wyke 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.

Sudhahar, Saatviga, 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).

Taniguchi, Motoki, 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.

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

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

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Tunstall, Lewis, 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.

Turowetz, Jason J. and Douglas W. Maynard. 2010. *Morality in the social interactional and discursive world of everyday life*. Springer, New York.

Uma, Alexandra N., 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.

Virtanen, Antti, 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*.

Virtanen, Mikko T. and Liisa Kääntä. 2018. At the intersection of text and conversation analysis: analysing asynchronous online written interaction. *AFinLA-e : soveltavan kielitieteen tutkimuksia 2018*, (11):137–155.

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

Wang, G. Alan, Harry Jiannan Wang, Jiexun Li, Alan S. Abrahams, and Weiguo Fan. 2014. An analytical framework for understanding knowledge-sharing processes in online Q&A communities. *ACM Transactions on Management Information Systems*, 5(4).

Weber, René, 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.

Wendler, Chris, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Willard, Brandon T. and Rémi Louf. 2023. Efficient guided generation for LLMs. *arXiv preprint arXiv:2307.09702*.

Wu, Hanqian, Lu Cheng, Jiahui Jin, and Feng Yuan. 2019. Dialog acts classification with semantic and structural information. In *2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS)*, pages 438–442. IEEE.

Xiao, Yimin, 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.

Xu, Guanghao, Hyunjung Lee, Myoung-Wan Koo, and Jungyun Seo. 2017. Convolutional neural network using a threshold predictor for multi-label speech act classification. In *2017 IEEE international conference on big data and smart computing (BigComp)*, pages 126–130. IEEE.

Yang, Diyi. 2021. 6 questions for socially aware language technologies. *Northern European Journal of Language Technology*, 7(1).

Zakharov, Stepan, 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.

Zhang, Amy X., 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.

Zhang, Justine, 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.

Zhang, Renxian, Dehong Gao, and Wenjie Li. 2011. What are tweeters doing: Recognizing speech acts in twitter. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Zhou, Xinyi and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A Micro-F1s for A2 model

Act.	**Neg.**	**Pos.**	**Act.**	**Neg.**	**Pos.**
quest.	0.98	0.94	chall.	0.88	0.47
req.	0.94	0.61	accept.	0.96	0.63
state.	0.77	0.96	denial	0.86	0.49
accus.	0.89	0.67	apprec.	0.99	0.51

Table 15: 10-fold Micro-F1s for Annotator2 model. Neg.=Negative and Pos.=Positive class.

Table 15 presents Micro-F1 scores for the SetFit-FinBERT1-A2 model, one of our best performing models. This enables an examination of how the specific model performs on the positive and the negative class, respectively, for each action. Based on the scores, performances for the positive class ('yes the action is present in the comment') tend to be lower. Especially for some of the face-threatening actions (*denial*, *challenge*), as well as *appreciations*, the performance for the positive class is much lower than for the negative. This further supports our observation that people tend to express face-threatening actions in an ambiguous manner, perhaps to avoid direct face-threats. It is interesting, though, that *appreciations* also show lower performance for the positive class. This might have been due to appreciations being quite rare in the dataset.

Model	**Threshold**	**Action**							
		question	request	statement	accusation	challenge	acceptance	denial	appreciation
FinBERT1-MS	≥ 1	**0.87**	0.68	**0.73**	0.63	**0.61**	0.64	0.55	0.60
	≥ 2	0.86	**0.69**	0.72	0.61	**0.61**	0.64	0.54	0.59
	≥ 3	**0.87**	0.68	**0.73**	**0.76**	0.60	0.62	0.54	0.63
	≥ 4	0.86	0.64	0.69	0.61	0.59	0.60	**0.56**	**0.65**
	≥ 5	0.85	0.60	0.66	0.54	0.56	**0.66**	0.53	0.64
	≥ 6	0.81	0.54	0.62	0.55	0.58	**0.66**	0.64	0.50
SetFit-FinBERT1-MS	≥ 1	**0.94**	0.79	**0.77**	0.71	0.59	0.85	0.56	0.62
	≥ 2	0.93	**0.82**	0.76	**0.73**	0.63	0.84	0.65	0.58
	≥ 3	**0.94**	0.79	0.75	0.71	0.60	**0.86**	**0.68**	0.65
	≥ 4	**0.94**	0.73	**0.77**	0.69	**0.64**	0.81	0.64	**0.78**
	≥ 5	0.90	0.73	0.72	0.54	0.50	0.66	0.59	0.75
	≥ 6	0.79	0.75	0.64	0.55	0.50	0.69	0.60	0.50

Table 16: 10-fold cross-validated macro-F1 scores with different thresholds.

B Threshold tests

To gain a better understanding of how threshold setting might affect action detection, we compared how well models performed when setting different threshold θ as a classification boundary. Here, we used the single annotations data, in particular the FinBERT1-MS and SetFit-FinBERT1-MS models. In this paper, we predefined our $\theta = 3$ for our experiments related to RQ1 and RQ2, as this allows us to focus on actions present at moderate to strong signal strength. However, here we wish to provide more insight into how thresholding might affect classification. Classifier training and testing follows the same procedures as described in section 5. See details specifically related to SetFit and the multilabel single annotation (MS) model configuration. It should be noted that with different θ, the dataset imbalances will notably change (see Appendix C). Class weighting (FinBERT1-MS) and sentence pairing used for SetFit help to account for imbalances when measuring model performance. However, due to differences in data distributions here we note that the test presented in Table 16 is only suggestive. More comprehensive analysis would be needed for more reliable results.

Based on the test (Table 16), it seems that for most actions, especially those that invite responses (question, request, statement, accusation, challenge), there might be no great difference between scores when using θ between 1–3, or even 1–4 for some. It seems that a $\theta \geq$ 4–6 (5–6 for some), might result in lower performance for many actions. It seems also that for responsive actions (and appreciations), a higher threshold might increase performance. This could perhaps be due to responsive actions being expressed in in a shorter or even more ambiguous manner, e.g. in a subordinate clause or shortly at the beginning of the comment, more emphasis being given to other actions in the comment.

C Label distributions

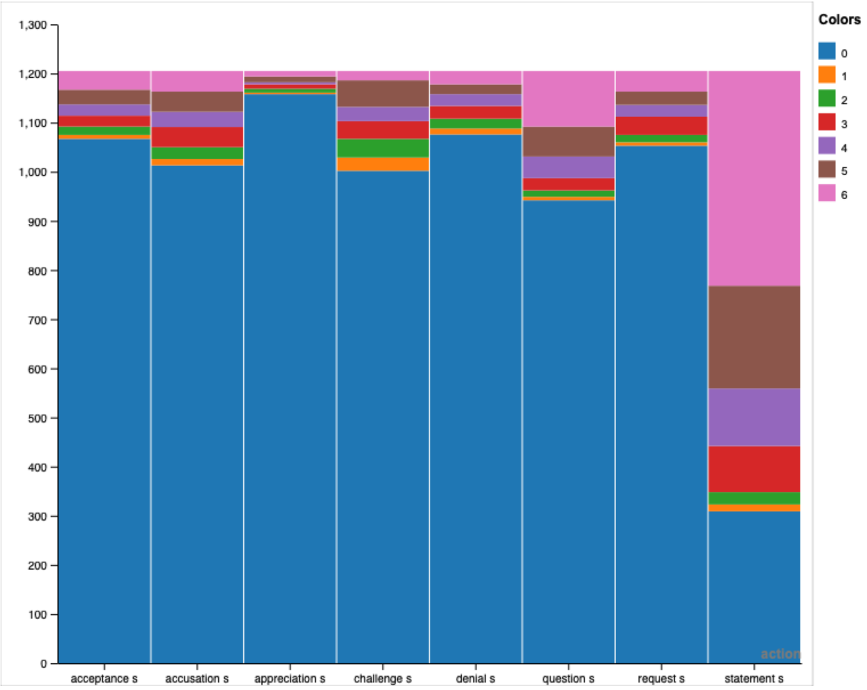

Figure 2: Distribution of labels in s=single annotations. Colors indicate label scores.

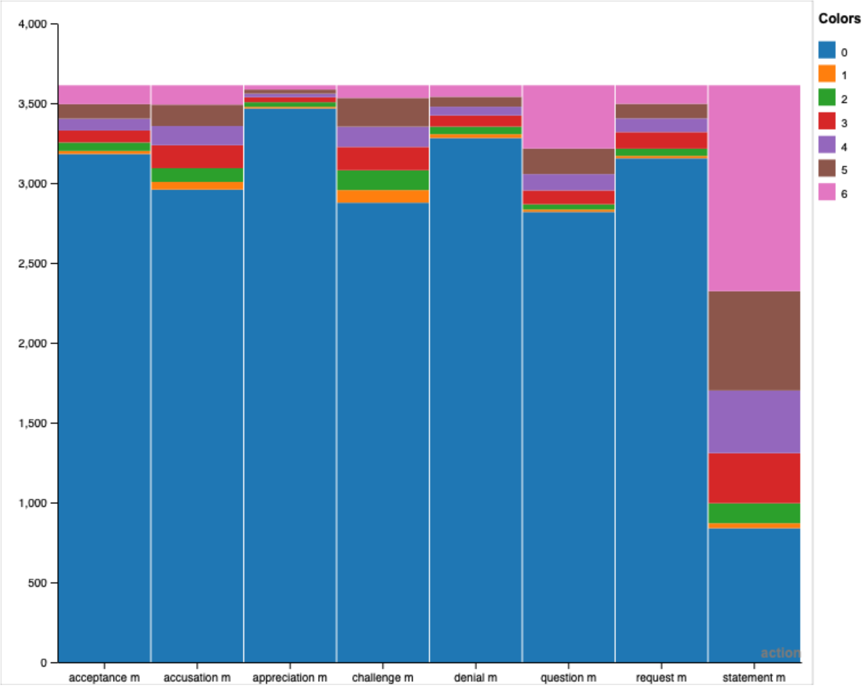

Figure 3: Distribution of labels in m=multiple annotations. Colors indicate label scores.

Dataset	Action	Label score							Total
		0	1	2	3	4	5	6	(positive label)
single	question	942	7	13	25	44	61	113	263
	request	1053	7	15	37	24	27	42	152
	statement	309	14	25	94	117	209	437	896
	accusation	1013	13	24	41	31	41	42	192
	challenge	1002	27	38	36	29	54	19	203
	acceptance	1067	8	17	22	23	30	38	138
	denial	1076	12	20	26	24	20	27	129
	appreciation	1158	3	8	9	4	12	11	47
		0	1	2	3	4	5	6	(positive label)
multiple	question	2819	15	33	87	102	160	396	793
	request	3155	15	45	104	86	91	116	457
	statement	839	31	125	314	392	622	1289	2773
	accusation	2959	48	86	144	120	134	121	653
	challenge	2877	79	124	146	127	180	79	735
	acceptance	3182	19	53	75	74	91	118	430
	denial	3282	24	47	70	55	62	72	330
	appreciation	3467	10	29	31	23	25	27	145

Table 17: Distributions of action labels by score in single annotations and multiple annotations. The label distributions are quite similar for both single annotations and multiple annotations. Overall, scores 3-6 are more commonly used than 1-2, although there are some differences between actions. For *challenges*, scores 1-2 seem a bit more common than for other actions. These are small differences, though. *Statements* have a positive score ($\geq$ 1) much more often than other actions, highlighting the assertive nature of the comments in our data. *Questions*, *challenges* and *accusations* also appear to be more commonly labeled with a positive score.