Paakki, Henna; Vepsäläinen, Heidi; Salovaara, Antti; Zafar, Bushra

Detecting covert disruptive behavior in online interaction by analyzing conversational features and norm violations

Published in:
ACM Transactions on Computer-Human Interaction

DOI:
10.1145/3635143

Published: 29/01/2024

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Detecting Covert Disruptive Behavior in Online Interaction by Analyzing Conversational Features and Norm Violations

HENNA PAAKKI, Aalto University, Finland
HEIDI VEPSÄLÄINEN, University of Helsinki, Finland
ANTTI SALOVAARA, Aalto University, Finland
BUSHRA ZAFAR, University of Helsinki, Finland

Disruptive behavior is a prevalent threat to constructive online engagement. Covert behaviors, such as trolling, are especially challenging to detect automatically, because they utilize deceptive strategies to manipulate conversation. We illustrate a novel approach to their detection: analyzing conversational structures instead of focusing only on messages in isolation. Building on conversation analysis, we demonstrate that (1) conversational actions and their norms provide concepts for a deeper understanding of covert disruption, and that (2) machine learning, natural language processing and structural analysis of conversation can complement message-level features to create models that surpass earlier approaches to trolling detection. Our models, developed for detecting overt (aggression) as well as covert (trolling) behaviors using prior studies’ message-level features and new conversational action features, achieved high accuracies (0.90 and 0.92, respectively). The findings offer a theoretically grounded approach to computationally analyzing social media interaction and novel methods for effectively detecting covert disruptive conversations online.

CCS Concepts: • Human-centered computing → Social media; Collaborative and social computing theory, concepts and paradigms;

Additional Key Words and Phrases: Disruptive behavior, detection, online aggression, online trolling, machine learning, natural language processing

ACM Reference format:

1 INTRODUCTION
Social media was first seen as an opportunity for new modes of open civic discussion and interactive participation [31, 32]. However, it has become evident that open online media have properties

This research has been funded by the Academy of Finland, projects “Reconstructing Crisis Narratives for Trustworthy Communication and Cooperative Agency” (project number 339931), and “Automated trolling and fake news generation in future social media: Computational and empirical investigations of the threat and its implications” (project number 320694).

Authors’ addresses: H. Paakki, Konemiehentie 2, 02150 Espoo, PL 15400, 00076 AALTO, Finland; e-mail: henna.paakki@aalto.fi; H. Vepsäläinen and B. Zafar, P.O. Box 68, Pietari Kalmin katu 5, 00014 University of Helsinki, Finland; e-mails: heidi.vepsalainen@helsinki.fi, bushra.zafar@helsinki.fi; A. Salovaara, Otaniementie 14, 02150 Espoo Finland; e-mail: antti.salovaara@aalto.fi.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2024 Copyright held by the owner/author(s).
1073-0516/2024/01-ART20
https://doi.org/10.1145/3635143
that often overshadow the anticipated positive aspects by permitting a range of disruptive behaviors that deteriorate constructive discussion in the forms of manipulation, provocation, and circulation of disinformation [73]. Two cases of harmful behavioral phenomena that are often found in online forum interaction are aggression and trolling. Both are common and evidently problematic online: recent societal events have shown that they can be used to harass groups of people as well as to perform large-scale ideological manipulation, both of which pose significant threats to egalitarian and democratic ideals [1, 8]. Detection and prevention of such harmful behaviors are essential for protecting these ideals and maintaining civility online [12].

This article focuses on the identifiable conversational properties of disruptive behavior in online interaction where messages relate to one another in a conversation, e.g., through turn-taking. Disruptive online behavior can encompass phenomena such as, but not limited only to, aggression, trolling, bullying, spamming and hate speech, when these materialize as behaviors within conversational interaction that disrupt, harass, or break the flow of conversation. By attending to interactive characteristics of disruptive behavior, our focus excludes cases that cannot be analyzed as part of an interaction, e.g., aggressive language or swearing in non-interactive contexts such as blog texts or isolated messages.

Disruptive behaviors can be hard or impossible to recognize and detect without examining the dynamics between messages. Thus, analysis of messages in isolation, outside of their conversational contexts, may not be sufficient for addressing covert, deceptive forms of disruptive online behavior. Although some message-level aggression detection systems can be very accurate (e.g., [93]), they are also vulnerable to user deception and manipulation [49]. Thus, to detect and prevent disruptive behavior online, we need detection methods that can identify both overt and covert interaction strategies. This is needed for models to be less vulnerable to deception and more generalizable to various contexts. As aggression and trolling are common phenomena in conversational interaction, analysis of their wider conversational contexts can help in understanding and detecting them. Such an analysis can build on the existing understanding of language use in conversation [40, 94]. Investigation of patterns in conversational dynamics can reveal covert manipulation such as violations of conversational norms of turn-taking that may be left unnoticed if investigating only individual messages (see, e.g., [53, 82]).

In this study, we analyze aggression and trolling, two forms of disruptive behavior that cause harm in social media conversations [73]: the former being mostly an overt form of behavior [81] and the latter often being more covertly acted out [36, 54]. We are primarily focusing on covert behavior – trolling – while our analysis of aggression serves to highlight the differences between the two behaviors. By doing this, we emphasize the need for structural analysis of conversational dynamics, especially for covert disruptive behaviors such as trolling.

We suggest that identifying patterns in the use of conversational actions and related norm violations provides more conversational context, generalizability, and better accuracy in the computational detection of covert disruptive interaction. By conversational actions, we refer to the functions of a message, i.e., what a turn does in a conversation in relation to previous turns: e.g., asking someone a question. We investigate the use of common conversational actions and how they are used to violate conversational norms in disruptive as opposed to non-disruptive interaction. Besides analyzing action taking in general, we attend to action-related conversational norm violations in which the expectations set by actions are broken. An example of such a violation would be to avoid answering a question, replying with an unexpected action such as an accusation instead. Norms related to actions are pervasive across various contexts [39]; thus, people can be expected to usually follow them in quite a similar fashion online as well [77, 114]. This suggests that analyzing the dynamics of common conversational actions and their norm violations may offer a generalizable approach for detecting disruptive behavior.
The reason for focusing on aggression and trolling specifically is that while we consider them to differ in their portrayed levels of overtness (i.e., immediate perceivability), they have often been analyzed together [14, 15, 59, 69]. Building on earlier research on aggression enables us to conduct our analysis on an existing verified dataset [126]. Our findings on trolling detection, in turn, provide a basis for future research. In particular, computational trolling detection has not been studied widely from a conversational perspective as our study does. Instead, computational approaches have remained mostly limited to message-level analyses. Moreover, the demarcation between aggression and trolling has not always been clear-cut. Aggression is often seen as hostile behavior involving the use of profanity, derogatory language, or aggressive sentiment [81]. These have also been considered as definitive features of trolling (e.g., [15, 97]), although the literature has also maintained that they are not equivalent phenomena [54]. In this article, by aggression we refer to overt means of offense within a conversation, actions that result in an *ad hominem* attack (per [126, p. 1353]). Trolling, in turn, has been defined in several ways, emphasizing different characteristics (see [26]), such as ideological [124] or political manipulation (or “Twitter trolling”; [27]), classic trolling (i.e., baiting newbies on discussion forums; [36, 48]), or deceptive harassment of minority groups [56]. In our article, by trolling we refer to *conversational trolling*, which manifests as strategically deceptive interaction online, aiming to confuse, provoke and manipulate others into participating in pointless and even harmful discussions (per [53, 54]).

We seek to show that *the ways in which people use common conversational actions and how they adhere to or violate action-related norms can be used to complement message-level features and thus to better detect covert disruptive interaction (trolling)*. To show the benefits of a conversational approach, we build on linguistics and *conversation analysis (CA)* to develop a framework for analyzing online interaction as turn-taking, in which each turn may perform one or several conversational actions.

Besides generally attending to the types of conversational actions used, we analyze one phenomenon in particular. We observe that humans have a tendency to prefer *symmetry* in conversation: some action pairs in turn-taking (e.g., question–answer) are normatively bound. Thereby, certain types of responses are expected after response-mobilizing actions [106]. Unexpected responses are, in turn, referred to as *asymmetric* actions in this article. We expect disruptive behaviors to manifest conversational *norm violations* where responses are asymmetric with respect to earlier actions (see [82]).

However, a non-trivial problem is that all turns in conversation need to be tagged with actions in order to analyze the conversational structure. We overcome the lack of suitable CA-based automated action tagging systems by utilizing a state-of-the-art zero-shot *Natural Language Inference (NLI)* model [121] to implement our action tagging scheme. Action-tagged messages are then transformed into various conversational features for computational detection. Drawing from our theoretical foundations, we address three *research questions (RQs)*:

- **RQ1 (Principle of symmetry):** Do aggression and trolling involve more asymmetry than constructive discussion?
- **RQ2 (Principle of conversational distinctiveness):** Are important conversational features such as actions and norm violations different in online aggression as compared with trolling?
- **RQ3 (Principle of context):** Do features related to conversational context – i.e., conversational actions, action pairs, and norm violations instead of isolated messages – allow more accurate detection of disruptive interaction as compared with earlier approaches?

We expect conversations involving aggression and trolling to have different identifiable characteristics in conversational features, e.g., the use of actions and conversational norm violations. By investigating conversations in the way presented above, our findings suggest that especially
trolling detection can be significantly improved if conversational action features are included as predictive variables. RQ2 functions to provide context for RQ1 and RQ3, and to illustrate why structural analysis of conversation dynamics (i.e., analysis of the interrelations of turns in conversation, drawing from linguistic theories) is essential, showing that in comparison with aggression, trolling involves more indirectness and turn-taking-based conversational strategies such as asymmetries. Our approach offers a new theoretically grounded approach for understanding social media interaction and new means for detecting covert disruptive online behaviors.

2 THEORETICAL BACKGROUND

CA theory and the concepts of turn-taking, actions, and action pairs provide the foundations for our approach. CA is a microanalytical method for investigating the practices, conversational actions, and sequential structure through which interaction is built and maintained [99]. Though most of its fundamental findings derive from observations on how spoken conversation is built in real time, it has by now been well established in digital CA that, even in online conversations, participants orient themselves to meaningful and normatively organized conversational actions in interaction with others [77, 114]. In the following subsections, we will discuss the theoretical foundations of our three RQs that stem from this starting position. The RQs will be empirically investigated in Section 5.

2.1 Conversational Actions and Action Pairs in Disruptive Conversations

The foundations for analyzing the conversational characteristics of online discussion are rooted in how people utilize common actions to achieve mutual understanding and how human interaction strives toward symmetry. Humans usually coordinate interaction by utilizing conversational actions and responses that match their expectations [106] in order to maintain a common ground with their interlocutors [25]. These characteristics of human conversation have been well established in CA research. Also in online conversations, participants orient themselves to meaningful and normatively organized conversational actions in interaction with others [77]. Even disputes are built symmetrically, utilizing contributions to the conversation that are expected in the situation [35].

The CA concept of action is essential in our approach. In CA, actions are what a turn does in a conversation in relation to previous turns. They share similarities with speech acts (per [5]) and dialogue acts that have been studied in natural language processing and online communication research [3, 23, 38, 75] as well as in Computer-Mediated Communication (CMC) research [43, 111]. However, unlike speech act analysis, CA research maintains that the sense of an action is negotiated through the next-turn proof procedure: what a turn does in conversation is not based solely on its appearance but rather on how it is responded to [92]. Thus, actions emphasize the dynamics between turns in conversation and can be used to analyze the conversational flow and reasons for conversational breakdowns [82].

Some actions appear together more often than others. These so-called adjacency pairs, henceforth referred to as action pairs, are common in ordinary well-structured conversations. By identifying recurring patterns in action pairs, computational models can analyze how people adhere to conversational norms in online interaction. However, this kind of computationally oriented (digital) CA research has been uncommon – and critiqued in the CA community due to contradicting views about the necessary depth of interpretation and aims for generalizability [58, 95, 103]. Thus, only a few coding schemes amenable for computational analyses of actions and action pairs (e.g. [37, 105]) have been published. Here, we take an applied stance to digital CA, utilizing the concept of action. Conducting computational analysis, our approach differs from qualitative
(digital) CA in that we cannot subscribe to all the methodological requirements within traditional CA research, which is typically based on in-depth qualitative analysis of a limited number of cases.

Our RQ1 builds on the expectation that asymmetries in action pairs would be more common in disruptive interaction. We suggest that people aiming to create disruption do not strive toward the principle of symmetry in interaction similarly to people aiming at a constructive conversation. As actions taken have an effect on subsequent actions, we expect action-related symmetry violations (e.g., asymmetric responses) to be more common in disruptive conversation in both aggression as compared with non-aggression and in trolling as compared with non-trolling. For example, asymmetries emerge when one ignores an action that expects an answer (e.g., leaving a question unanswered), provides a response that is a mismatch to the previous action (e.g., a “thank you” to an accusation), or challenges others or their messages [82]. These have been found to be characteristic of trolling: for example, a troll may respond unexpectedly to an earlier action, pretend not to understand something, or ignore what was previously said [82].

In ordinary cooperative conversation, asymmetries are seen to signal a problem, such as a misunderstanding, a gap in presumed mutual knowledge, or reluctance [24, 86]. However, they can also be used to provoke others [82]. In particular, the norms of common action types (questions, accusations, requests, statements) may be violated by performing any of the above-listed asymmetric interaction strategies of ignoring, mismatching, and challenging [82]. Based on this, we expect that a range of conversational trolling behaviors (per [54]) have shared and automatically identifiable characteristics in action-related conversational norm violations. Besides trolling, it can be expected that violations of the norms of conversational actions are present also in aggression.

We therefore ask whether the principle of symmetry between conversational actions separates disruptive (aggression, trolling) and more constructive online interaction (non-aggression, non-trolling) from each other (RQ1). Based on the discussion above, we expect this to be the case.

### 2.2 Differences between Aggression and Trolling on Social Media

In the previous subsection’s formulation of RQ1, we did not attempt to separate aggression and trolling from each other. Instead, we considered both of them as examples of disruptive behavior in online interaction. In many earlier computational studies, these two phenomena have not been clearly separated. In the extant research, aggression has often been considered a defining element of trolling [14, 15, 59], although not all trolling behaviors involve aggression (see, e.g., [54]; also, Table 1). However, identifying norm violations in conversation may allow us to distinguish between different types of covert versus overt disruptive behavior, for instance, trolling and aggression.

Aggressive online behavior has been characterized by profanity, derogatory language, spam, hate speech against individuals or groups, and aggressive sentiment (see, e.g., [81]). As we are analyzing text-based online conversations, we are dealing with verbal aggression [52], which can be characterized through the use of aggressive or offensive, vulgar, opinionated, or rude language [52, p. 7]. Aggression has been identified by looking at sentiment [116], blacklist terms, or offensive words [87, 101]. Also, the use of particular rhetorical strategies or (lack of) politeness can help distinguish peaceful interactions from ones that lead to aggression [126]. Although many studies have focused on identifying online aggression, its definition, for example, in contrast to hate speech, is not yet very clearly delineated (see [81]). Based on these considerations, we specify our earlier-presented definition further by considering online aggression as directly visible (overt) abusive language directed at another individual or a group: an ad hominem attack classifiable as “rude, insulting, or disrespectful” (similar to Zhang et al. [126, p. 1353]).

Trolling, then again, has unique characteristics among antisocial online behaviors. Theoretical and qualitative examinations of trolling stress its interactive characteristics [53, 56]. Trolling is
Table 1. Nonaggressive Trolling on Reddit/cats

<table>
<thead>
<tr>
<th>Position</th>
<th>Speaker</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original message*</td>
<td>Troll</td>
<td>Does my cat have autism?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>So I have 2 cats. One loves to be cuddled the other one hates to be hugged and doesn’t like to be touched for more than a few seconds. He has about 20 toys but only ever plays with his birdy toy that makes noises. He is really sensitive to low noises and will hide under the bed if he hears one. And he doesn’t look you in the eyes, EVER. I have a vet appointment scheduled for next week for routine vaccines. Should I ask the vet if my cat has autism, or is that a dumb question. Could his rabies vaccines be giving him autism?</td>
</tr>
<tr>
<td>Level 1</td>
<td>A</td>
<td>I can’t tell if you’re really that stupid or if you’re a troll. Either way, dumbest thing I’ve read all day</td>
</tr>
<tr>
<td>Level 1</td>
<td>B</td>
<td>…you’re joking about the vaccine causing autism right?</td>
</tr>
<tr>
<td>Level 2</td>
<td>Troll</td>
<td>He was a little kitten when he first got the vaccines. I don’t know if that changes anything…</td>
</tr>
</tbody>
</table>

*Original message refers to the message that initiates a conversation. Levels 1–2 refer to the positions of the messages in the extract from a threaded forum conversation.

T stands for Troll, A and B for other participants in the interaction.

identifiable by common trolling styles [54], for example, posting about taboos or sensitive subjects, giving ingenuine advice that may put someone in danger, excessively criticizing others, e.g., on spelling mistakes, or luring others into tangential discussions. These considerations allow for a further specification of trolling as deception and strategically uncooperative behavior in conversational interaction, which aims at manipulating others into participating in tangential or even harmful discussion [54, 56, 82, 113]. Such a demeanor in interaction provides a unifying frame for trolling behaviors across different contexts. This frame helps address the context-bound nature of trolling [30]. Computational identification of trolling has previously been highly problematic due to the aforementioned context dependency of trolling behaviors, but also because trolling is by nature deceptive [36, 56] and because online research is often unable to recover what the original unexpressed intentions or thoughts of participants were. Thus, in this article, automated detection of trolling is based on recurring identifiable patterns of manipulation within interaction.

Aggression and trolling also differ in sentiment. Though negative sentiment is central to aggression and can be involved in trolling, it is not an overarching defining feature of all trolling [54, 83]. Even so, computational trolling identification has been largely based on sentiment analysis (e.g., [14, 15, 83, 97, 98]), operationalizing trolling as aggression. This is problematic, because aggression and trolling can differ in terms of directness: one being a direct form of abuse and the other being characterized by more covert behavior, such as deception and elusiveness. Sentiment-analysis-type approaches fail to capture cases of nonaggressive trolling such as one presented in Table 1.

Trolling and aggressive behavior are therefore not equivalent to each other, though they are not mutually exclusive: although trolling might be acted out aggressively, it may also take a range of different forms of deceptive or manipulative behavior that may seem civil on the surface. We expect conversational norm violations, however, to emerge differently in aggression and trolling, to some extent differentiating trolling from aggression. We expect that aggression as overt offensive action is likely to involve more exaggerated directness in norm violations based on Nobata et al. [81] and politeness theory [11] [45, p. 37]. These would include, for instance, challenging conversational partners’ actions or exhibiting compliance-avoiding behavior, but less often other more indirect asymmetries such a mismatched responses. As for trolling, we expect that though
there are different trolling styles [54], many of them will share interactional characteristics: systematic, strategically uncooperative behaviors related to conversational actions [82]. Thus, we ask whether the principle of conversational distinctiveness holds: whether trolling involves different important features as compared with aggression, e.g., more action-related indirect conversational norm violations such as asymmetric actions (RQ2), and expect this to be the case.

2.3 Conversational Context in Disruptive Interaction

In computational detection of aggression and trolling, approaches building on conversational features have been relatively rare, whereas message-level detection models have been dominant. We will review the extant studies in more depth here. We start with message-level approaches and complement them by presenting user-level studies. We conclude by reviewing the few existing approaches that focus on conversation-oriented features, such as conversational actions.

Message-level approaches to aggression (e.g., [55, 90, 123]) and trolling detection (e.g., [14, 83]) have been most common. In message-level aggression detection, sentiment features have often been used to distinguish aggression from nonaggression (e.g. [4, 70, 123]). Other features used have included blacklist terms and offensive words [87, 101], uni-, bi- and trigrams, and part-of-speech tags [93] as well as punctuation and connectives (see [81]). In trolling detection, mostly including studies on aggressive [112] and political messaging [61] on X (formerly known as Twitter), message-level features have included word use, sentiment, and posting times [78, 83]. Outside of the X context, trolling has often been identified strictly in terms of negative sentiment within individual messages (e.g., [14, 15]), which illustrates the unclear definition of trolling in its computational detection. Furthermore, sometimes the description of data and their gathering methods have also been incomplete (e.g. in [14, 15]). This has left it uncertain as to what the models have actually detected. Mojica de la Vega and Ng [79] included reactions to trolling as features in their model: thus, this is an example of a study that did not use solely sentiments in classification. However, instead of classifying trolling directly, the authors trained different classifiers for identifying intention, its disclosure and interpretation, and response strategy. Although they analyzed reaction types, the classifiers were focused on message-level features.

User-level features have been used occasionally in identifying aggression and trolling beyond the message level. Chatzakou et al. [18], for example, detected aggression on Twitter using information regarding the content of different users’ messages, networks, and attributes such as negativity and popularity. They found that aggressors are relatively popular within their networks but tend to include more negativity in their messages than normal users. Bag-of-word features, 2-to-5-grams (word combinations) and sentiment have also been used to detect aggressive user behavior [20]. Wulczyn et al. [120] analyzed user activities based on posting toxicity but only used automated detection to tag individual messages as toxic or nontoxic. As for trolling detection, user-level features have included overall sentiment found in all messages posted by each user [97] and part-of-speech tags, sentiment, metadata, posting time, negative word use, and punctuation [78].

At the conversation level, finally, in relation to aggression detection, graph-based modeling has been employed to analyze interaction patterns and to forecast antisocial actions in online discussion [125]. In addition, features related to politeness and conversation prompts have shown potential in predicting aggression [126]. Here, prompts referred to rhetorical functions of a message, i.e., how the message orients to other users (e.g., fact-checking, editing), whereas politeness strategies captured properties such as hedging, personal pronoun use, and negative politeness (see [126]). Recent studies have also proved that examining conversation structure can be highly useful in analyzing, detecting, and predicting toxicity in online conversations [17, 57, 126, 128]. As for trolling, to our knowledge, only two studies have used conversational features to detect trolling. The first mostly used message-level features to identify troll questions within Yahoo Answers data but also...
reported that using answers that specific questions got as features led to promising results [50]. The second used network dynamics and information about how text edits relate to the original text [71]. However, the features used were dependent on platform-specific functions that are unavailable on many other social media sites. Conversational features related to sentiment, topic, and similarity between message–response pairs have also been used successfully to detect disputes on Wikipedia [116]. However, although disputes share some similarities with aggression and trolling, similar features have not been applied to detect these two phenomena.

To sum, most research on aggression and trolling identification has focused on message-level features. Studies have shown, however, that (linguistic) conversational features can yield promising results in studying online aggression (e.g., [50, 126]). Thus, studying the interactions between users (e.g., argumentation: [51]; discursive strategies: [122]; acts in dialogue: [10]), beyond just question–answer relations (as in [50]) could offer valuable insights in aggression and trolling detection. Based on the discussion above, we suggest that conversational features found in aggression and trolling can significantly improve the detection accuracy of these antisocial behaviors, as the wider conversational context is more telling than individual messages that violate social norms. We suspect that investigating the wider context of disruptive behavior would allow more accurate and generalizable automated detection of such interaction, aggression on the one hand and especially for trolling on the other (RQ3).

3 DATA

We studied disruptive interaction using two datasets: one that we used to distinguish aggressive conversations from nonaggressive ones and another that enabled us to compare trolling-like and nontrolling conversations. As aggression and trolling are different phenomena, we needed clearly and robustly labeled data. This was important in order to test whether our conversation-oriented approach would work for both aggression and trolling detection and whether the two behaviors would entail differing conversational features.

Many datasets are available on aggressive or hostile behavior and hate speech (e.g., [47]). However, most focus on aggression at the message level and thus provide isolated messages instead of full conversations (e.g., [47, 69]). Moreover, most are X data, which are not necessarily very conversational (see, e.g., datasets listed in [127]). When it comes to more conversational interaction, the Wikipedia Conversations Gone Awry dataset [126] fits our criteria of conversationality. Here, aggressive conversations are defined as starting out civil but ending up in a personal attack. The dataset has been annotated by crowdworkers who have filtered and then labeled conversations ending up in “rude, insulting, or disrespectful” behavior as aggressive. The dataset has a robust annotation scheme for personal attacks. Thus, we selected it as our aggression data. The dataset contains 30,021 messages from 4,188 Wikipedia conversations (see Table 2), each conversation including at least three messages. The same authors also provide a similar Reddit dataset, which we excluded because it considers deleted messages as indicators of personal attacks. At this stage, we did not want to use data from which central messages were missing as it would not allow sufficient analysis of entire conversations, especially because our idea was to analyze entire conversations and the interrelations between actions. The missing messages could have been extremely important for our analysis. Another potential dataset could have been the ComMA dataset of Kumar et al. [68], but as the data is multilingual, at this stage we chose to work with the English-language Wikipedia dataset, as the only suitable conversational trolling dataset is in English only. While multilingual analysis is important, we feel that limiting the language of the data to English only is an important first step in the detection of covert disruptive online behavior in conversational interaction.
Table 2. Main Characteristics of the Datasets Used in the Study

<table>
<thead>
<tr>
<th>Corpus (type, language)</th>
<th>#conversations</th>
<th>#messages</th>
<th>#authors</th>
<th>Average msg length (chars.)</th>
<th>#messages per conv.</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggression (whole corpus, English)</td>
<td>4,188</td>
<td>30,021</td>
<td>8,069</td>
<td>372.5 (aggression)</td>
<td>373.7 (non-aggression)</td>
<td>3–20</td>
</tr>
<tr>
<td>Trolling* (whole corpus, English)</td>
<td>4,911</td>
<td>54,071</td>
<td>18,829</td>
<td>199.5 (trolling)</td>
<td>205.6 (non-trolling)</td>
<td>3–286</td>
</tr>
</tbody>
</table>

*The trolling dataset is highly imbalanced, including 257 cases of trolling and 4,654 non-trolling conversations. We used the SMOTE algorithm [19] to create a balanced dataset. The procedure involved combined undersampling and oversampling of the data: randomly undersampling the majority group and then oversampling the minority group to produce a balanced dataset.

Finding a suitable dataset for analyzing trolling was more challenging. Trolling data collection has previously been based on (1) revealed X’s Russian troll accounts1 (2) aggressive behavior [59], or (3) suspected or mentioned trolls [79]. However, these do not exhibit enough conversationality or are in other ways unsuitable in the following ways. First, as was the case with aggression datasets, few datasets on trolling include sufficient meta-level information for analyzing conversational behavior (e.g., to which message another one is responding) to be able to analyze action pairs. Second, wanting to study conversational trolling behaviors exhibiting a range of different strategies (as in [54]), we could not use data consisting of only aggression-type trolling. Similarly, manipulative X accounts are but one form of trolling that may not be conversational either. Moreover, (possibly biased) lists of confirmed troll accounts reveal only a limited number of such accounts. Lastly, although data can be gathered by searching for troll mentions on online forums (e.g., [79]) and such data could also portray enough conversationality, it could fail to include deceptive forms of trolling unnoticed by others. Conversations sampled this way may also include cases labeled as trolling in which trolling accusations are actually used as a rhetorical strategy to counter an unwanted argument in a conversation – cases that do not necessarily include any trolling (see [82]). We therefore concluded that user-based flagging or reporting of trolling behavior as an identification method would be liable to be biased (e.g., [50]).

Due to these issues, we decided to use and extend a dataset collected by us in two earlier studies [82, 113].2 These datasets followed our definition of trolling as conversational behavior. We expanded the original dataset following the original data collection procedure and annotation scheme used in Paakki et al. [82]. The data sampling was based on Hardaker’s categorization of commonly used trolling styles [54] and operationalization of their identification on the guidelines defined by Paakki et al. [82]; the original paper provides further details on collection and annotation criteria.

In the operationalization of trolling identification for our manual data collection to expand the data, we assumed the definition of conversational trolling described in Paakki et al. [82] as well as the annotation guidelines for trolling identification used in the aforementioned qualitative study. Per this definition, a troll is a participant in a discussion who feigns sincerity, but whose real goals are to disrupt, digress, provoke, or otherwise prolong futile discussion by utilizing strategies described in Hardaker’s typology: aggressing, shocking, endangering others, antipathizing, (hypo)criticizing, or digressing [54]. Trolling behaviors are delimited to ‘successful’ trolling, cases in which trolls manage to elicit responses.

1Twitter’s list of 2,752 Russian trolls – Vox: https://www.vox.com/2017/11/2/16598312/russia-twitter-trump-twitter-deactivated-handle-list
2The anonymized trolling dataset is available upon request. See https://github.com/Aalto-CRAI-CIS/Detecting-covert-disruption for instructions on how to apply for data access.
We expanded the dataset using the collection and annotation criteria described in Paakki et al. [82] (see the article for detailed guidelines): identifying systematic or repetitious (see [128]) deceptive behavior utilizing one or more of the six trolling strategies per Hardaker [54]. We collected novel examples from the same sources the original data stemmed from: comment sections of The Washington Post, The Guardian, The Telegraph, and Reddit newsgroups, pertaining to similar topics as in Paakki et al. [82], namely, political discussions around Brexit and climate change, and interest-based or recreational topics around fitness and well-being, relationships (similar to [113]) and pets.3 We first read comment sections on each topic until we had identified approximately 200 conversations containing trolling-like behavior (including the 68 cases present in the original data [82,113]). A conversation here refers to a branch within a comment section thread. We set the minimum number of comments in a conversation at 3 messages to match the minimum conversation length in the aggression dataset [126]. We attempted to capture a wide range of trolling styles across different topics and platforms, similar to the earlier qualitative study on conversational trolling [82].

We collected the data by scraping whole comment sections that we had first read manually, by using screen scraping (Selenium webdriver scripts) and the Reddit application programming interface (API), and transforming the data into the computer-readable JSON format. This resulted in 4,911 conversations (54,071 messages), including both trolling-like behavior and non-trolling conversations. We made sure that the data included similar message-level and user-level details (e.g., user ID, posting time, reply-to-ID) to the Awry dataset that we chose for studying aggression. We used user IDs and reply-to-IDs to identify response patterns.4

We then carried out all the manual annotation work required for this article. The original dataset was validated as part of the qualitative study [82] by conducting an inter-annotator test with three annotators for discovering trolling in conversation threads extracted and separated from entire comment sections under either news articles or items starting Reddit threads [82]. One annotator selected a random set of conversations separated from whole comment sections, including a set of non-trolling and trolling threads, altogether 100 threads. The three annotators then proceeded to annotate these threads as either trolling or non-trolling. Fleiss kappa [42] measure for the inter-annotator agreement achieved an overall agreement of 86.5%. The free-marginal Fleiss kappa was 0.74, which signifies substantial interrater agreement (range: 0.40–0.75; [42]) [82]. To further validate the dataset, we conducted an inter-annotator test for identifying trolling within whole comment sections, as we expected this to be a harder task. For this task, one annotator created a set of documents with whole comment sections (N = 100), which either contained or did not contain trolling. With two annotators, we reached inter-annotator reliability scores of 87.10% for overall agreement and 0.74 (substantial agreement) for Fleiss kappa [42]. For both annotation procedures, we performed iterative annotation, meaning that we first annotated 3–4 practice batches of data, negotiating our annotations in between batches. The disagreements

3Reddit conversations were collected from the following subreddits: worldnews, ukpolitics, cats, fitness, weightlifting, vegan, Binauralbeats, Advice, Meditation, relationship_advice, askwomenadvice, TwoXChromosomes. The Washington Post data was collected from the politics section (for the Brexit topic), Climate & Environment and Climate Solutions sections (climate change), Well + Being (fitness), Lifestyle (pets). The Telegraph data: News section’s Environment and Global Health Security (climate), News – Politics (Brexit), Life (fitness), News and Life sections (pets). The Guardian: News section (climate, Brexit), Lifestyle section (pets, fitness, relationships). All included data is in the English language only; non-English content was excluded.

4Most sources provided both user IDs and reply-to-IDs (ID of parent). However, for The Telegraph, this needed to be extracted with scripts during data collection based on location of message in a thread, user IDs, and message content or username mentions in parent and response messages. Often, responses were located far away from the parent in a thread; thus, we could not assume that the next message in a thread would always be a response to the previous message.
often stemmed from overlooking some detail in the guidelines, missing some important feature in the conversation, or different understandings of the guidelines or the conversations. In the case of different understandings, we analyzed the difficult case, the guidelines, and our interpretations in detail, and, if needed, updated the guidelines so that difficult cases could thereafter be resolved correctly and the classes could be differentiated. We continued iterative annotation until we achieved a good enough understanding of the guidelines and the classes, meaning that a simple agreement percent was high enough (over 80%). After this, we annotated the final set of approximately 100 conversations to calculate the inter-annotator agreement.

Both datasets’ main characteristics are summarized in Table 2. Similar to the aggression data, we included in our final trolling dataset only conversations that involved at least two responses to the message initiating a new conversation. The discussions in the aggression data involve a range of topics. Also, the trolling data include a variety of trolling strategies (as defined in [54]), as well as conversations from several different online forums – centered on politics and society as well as on leisure and recreational activities. Finally, we used data augmentation methods (SMOTE [19]) to address the high imbalance between trolling and non-trolling cases: there were 257 trolling conversations and 4,654 non-trolling conversations. This produced an augmented training set with a balanced set of trolling and non-trolling conversations. We used this dataset for classification tasks only, following recommendations by Chawla et al. [19].

4 DATA PROCESSING AND METHODS

The processing and analysis of data involved three steps, as depicted in Figure 1. The following subsections describe these three steps in more detail, including:

Step 1 – Action tagging: The tagging of each message with the probability of how likely it is to belong to each CA-based action type in our tagging scheme (see Table 3).

Step 2 – Feature extraction: Extraction of required features from the tagged datasets, including action-based features and features that allow comparisons to earlier research.

Step 3 – Data analysis: Analyzing the data in order to answer our three RQs.

These steps will be described in more detail below.
Table 3. CA-Based Action Tagging Scheme and Descriptions of the Dependencies between Action Pairs

<table>
<thead>
<tr>
<th>First action</th>
<th>Normative*</th>
<th>Symmetric responding actions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>question</td>
<td>Y</td>
<td>statement</td>
<td>Questions require an informative answer or indication of inability or willingness to provide the information [29, pp. 217–248].</td>
</tr>
<tr>
<td>request for action</td>
<td>Y</td>
<td>denial, acceptance</td>
<td>The action requested [28] expects an acceptance or rejection [28, pp. 259–266] (rejection operationalized here as part of denial category).</td>
</tr>
<tr>
<td>proposal</td>
<td>Y</td>
<td>denial, acceptance</td>
<td>These are considered similar to requests, with the difference being that the beneficiary of the action is the respondent (suggestions in [28]).</td>
</tr>
<tr>
<td>accusation</td>
<td>Y</td>
<td>denial, acceptance, apology</td>
<td>Accusations convey a complaint, ascribing the responsibility for an unsatisfactory event to a person or a group [110]. The accused is expected to deny or contest the accusation [85], acknowledge it and explain themselves [34] or perhaps apologize.</td>
</tr>
<tr>
<td>challenge</td>
<td>Y</td>
<td>denial, acceptance, statement, apology</td>
<td>Challenges convey a negative stance towards another party, denying the epistemic basis of the interlocutor’s priorly expressed claims, actions, or feelings, through various forms (e.g., assertive/interrogative; [67]).</td>
</tr>
<tr>
<td>statement</td>
<td>N</td>
<td>– (no normatively expected responding actions)</td>
<td>Statements assert a fact or a claim but do not set a high expectation for a response [106]. The statement class includes informings and assertions, such as contributions of information, statements of opinion, answers to questions (that provide information), and announcements, which e.g., report a general rule or information.</td>
</tr>
<tr>
<td>appreciation</td>
<td>N</td>
<td>– (no normatively expected responding actions)</td>
<td>Appreciations include various forms of positive evaluative reactions, evaluations and thanking [22]. They can invite a response but do not strongly require one.</td>
</tr>
</tbody>
</table>

*Expects a response.

4.1 Step 1 – Action Tagging

For tagging messages with actions, we first developed a CA-based action tagging scheme based on earlier research [25, 82, 104]. We then used a transformer-based NLI model to computationally tag all messages found in our data with actions, following our tagging scheme.

4.1.1 CA-Based Action Tagging Scheme. In Section 2.1, we suggested that conversational actions can reveal possible norm violations within a conversation as well as interaction dynamics between participants. Based on CA research [25, 39, 104] and earlier research on actions commonly performed in trolling [82], we chose to focus on 7 pair-initiating actions (see Table 3; first column) and 3 responsive actions (Table 3; additional actions in the third column: denial, acceptance and apology). All of them are either first actions (actions that initiate an action) or responding actions. These form action pairs (as seen in Table 3) that we expect to reveal how participants follow or violate social rules. On the left are first actions that commonly expect certain types of replies whereas
the column “Symmetric responding actions” indicates which actions are acceptable responses to them. Some actions do not expect (a certain type of) a response; this is indicated in the second column (“Normative”). Normativity refers to whether the action expects a response (Y) or not (N).

While the action pairs in Table 3 adhere to social norms, we consider (based on [82]) norm violations (i.e., asymmetric responses that go against the expectation presented in the table) to be informative in trolling as well as in aggression.

4.1.2 NLI-based Action Tagging. To tag all messages in our data with actions, we used NLP and a pre-trained language model [118]. This was possible because several suitable large language models have recently been published for public use. A suitable model for our purposes had to fulfill the following four requirements: it had to (1) be able to identify relevant action classes; (2) offer adequate accuracy (models trained on some existing datasets do not perform well on some of the required classes; see [65]); (3) manifest generalizability, because we have learned in our own tests that while models trained on chat conversations (in which turns are typically short) can be quite accurate [107], they are not fully applicable to long forum messages; and, finally, (4) be able to classify messages as stand-alone texts, because using previous messages’ tags as features would possibly confound the analysis of asymmetries.

From the available pre-trained language models, as a solution to these requirements, we decided to use a zero-shot NLI approach [121] (using MNLI corpus: [118]) with BART sequence classification as a text encoder [74]. In NLI tasks, a language model is used for inferring whether a text includes or expresses an idea of interest (e.g., a topic or certain sentiment; [121]). Formally, this can be described as a hypothesis test carried out on a given premise text [118, 121]. When employing a zero-shot approach, the tagging task does not require any training or labeled data from the researcher [121]. This is possible because the NLI model has already been pre-trained with labeled large corpora that have been annotated and trained for textual entailment tasks (see [118]). This leads to a system that can represent both the hypothesis text as well as the to-be-labelled input text [121]. The model encodes both parts as tokenized sentence embeddings. Based on these, it can output a probability for textual entailment, i.e., the probability that the text entails what is suggested by the hypothesis. For example, given the hypothesis “This text is an apology.” and the premise text “Are there things I might be doing normally that I don’t _know_ are bad?”, the model gives a probability of 1.28%, i.e., the text is very unlikely to express an apology. In our tagging task, we used the NLI model to provide probabilities for 10 action hypotheses, e.g., “This text is a question”. We performed this procedure for each action presented in Table 3.

NLI allowed us to tag messages with actions, which is necessary for us to identify symmetric and asymmetric action pairs and other relevant conversational features. Zero-shot NLI enabled us to tag conversations even with tags for which we found no existing labeled training data, e.g., accusations and challenges. This tagging approach has already been used successfully in a number of related classification tasks, for example, to assign topic, emotion, and situation labels to texts (see [121]). We utilized the open-source Python code5 by Yin et al. [121] as well as code openly published in the transformers library [119]6 in our own scripts that processed all conversations in our data. For tokenization, we used BartTokenizer [74].

The downside was that NLI is computationally demanding: although we used a high-performance computing cluster, tagging a single message took 2–3 seconds to perform. Another limitation was that the BART encoder did not accept long messages (its limit was approximately 1,000 tokens). We classified the messages that exceeded the limit (230 in aggression set, 685 in

---

5https://github.com/yinwenpeng/BenchmarkingZeroShot
6https://github.com/huggingface/transformers
trolling) as ‘statements’ with the probability of 75.0, because statements commonly had the highest probability as main actions for other long, yet classifiable, messages in our data. In addition, longer messages tended to contain background information leading to a statement.

When evaluating the zero-shot NLI approach, Yin et al. [121] reported a label-weighted $F_1$ of 53.7 on tests they conducted on Yahoo Answers when using the large and more recent BART model pre-trained on MNLI. We also tested the suitability of NLI action classification for our task using a small hand-labeled set of action-tagged social media messages (50 messages per class). The macro-$F_1$ score for action tagging was 0.69 and accuracy was 0.71. Appendix A provides evaluation scores (recall, precision, $F_1$) for all actions separately.

4.2 Step 2 – Feature Extraction

From our textual data, we extracted features that have been found common and informative in the type of conversations we are analyzing: action-based features that we see having potential in generalizability in studying especially covert disruptive behavior (based on [82]) and complementary features such as negative sentiment that have been considered central in earlier research (agreement: [115]; sentiment: [4]; politeness and prompt types: [126]). Features similar to those used by Zhang et al. [126] also allow comparisons between features used in earlier research and our own results with action-based features. As we conducted conversation-level detection, the unit of analysis was a whole conversation: a discussion branch within a thread, initiated by a root-level comment to the topic (e.g., news article, Reddit original message). Next, we will describe our features in more detail, starting with actions, and then proceed to other features.

4.2.1 Action Features. We based the action features on NLI tagging and the dynamics between action pairs. These included frequencies of main actions used in a conversation, their NLI probabilities, and the number of (a)symmetries found between pairs.

Main actions and action probabilities. It is known that a forum message may perform several actions in one message and can simultaneously both complete an action pair initiated by another actor and start a new one [108]. For instance, a question may also involve a challenge, and a response that concludes an action pair may also initiate a new pair by having a provocative follow-up claim. Our pre-analysis indicated this was also common in many messages in our data. To account for this, we postulated that among those 10 different action types that are listed in Table 3 (7 different first actions and 3 additional actions that can act as responses), every message can have two (“top-2”) main actions, that we will consider in our analysis of asymmetries. Several considerations led to this decision. First, with a median length of 26 words, many messages in our data were relatively short and thus unlikely to have a higher number of actions than 2 uniformly. Second, the average NLI probability for the second most likely action type remained above 0.9 (a cut-off point that we feel is suitable as the third most likely action’s average probability was approximately 0.85). Moreover, the most prevalent actions will more likely be normatively binding than less central, subsidiary actions. Third, speech act and conversation analysis research alike suggest that a turn can have two actions, especially if separate parts in the turn can be identified [46, 72]. This is often relevant in online conversations in which participants tend to include a responsive action to an earlier turn as well as an action initiating a novel contribution in the same message, to accomplish more at one go [114]. Fourth, our yet unpublished analyses of comparable social media datasets also suggest that two is the most common action count in social media conversations on the type of forums included in our data.

Finally, since our classification task was to detect aggression and trolling within entire conversations instead of individual messages, we had to transform the previously mentioned features to a level where the unit of analysis is one conversation. For each conversation, we calculated the
number of times each action type appeared within the top-2 actions and divided the count with the total number of main actions in the conversation. Also, for each action type, we calculated its mean NLI probability across all messages in the conversation. Using this feature, we also allowed other (subsidiary) actions besides the top-2 main actions to contribute to the classification. For simplicity, however, these were not included in the analysis of asymmetries in action pairs.

(A)symmetries and replies. To extract symmetry and asymmetry features in action pairs, we investigated whether acceptable responses were given to pair-initiating actions as defined in Table 3. Our script (see Appendix F7) identified all main actions that are considered as normatively binding (i.e., initiating an action pair that expects completion from the recipient side), and investigated the responses they received. We counted all symmetric pairs (i.e., normative first actions that received an expected response), and asymmetric pairs: (a) mismatching pairs in which the normative first action received an unexpected response and (b) missing responses in which the normative first action received no response at all.8 These were also counted separately for each action type. In counting mismatched responses, we took a conservative approach, i.e., for a response to count as mismatched, both of its main actions needed to be mismatches with the actions found in its parent message. For statements and appreciations, asymmetric responses were not counted, as they do not strongly expect a response. We also extracted as features the number of any types of responses to normative actions (symmetric or asymmetric) and the time passed between a normative action and its reply. These counts were transformed into conversational action-based features across entire conversations in our conversation-level detection.

4.2.2 Complementary Features: Allowing Comparisons to Earlier Research. We included the following features in the analysis as frequencies within conversation, averages, or probabilities (see Table 4 for information on which method was used).

Sentiment. A blacklist of offensive words [76] was used to extract frequencies of offensive word use per message (see details in [76]). For this task, some preprocessing of the data was needed (see Appendix D). To allow analyses of sentiment features’ importance in detecting aggression as compared with trolling and to compare our results with earlier research, we first extracted message-level toxicity measures. In this task, we utilized Perspective API9 (similar to [126]), which uses machine learning (ML) to automatically rate the toxicity of a message. It defines toxicity as rude, disrespectful, or unreasonable commenting that may lead to someone leaving the discussion. Perspective API does not allow an empty comment as input and only allows 3,000 characters per message, discarding the rest. However, as there were only 125 messages in total with length greater than 3,000 characters, we did not further process the discarded parts. In addition, we included three complementary features to characterize affect in reactions to others’ messages (negative reaction, positive reaction, or neutral). We wished to investigate whether affect type in reactions to others’ contributions in conversation would play an important role in detection tasks, since sentiment has been found to be important in detecting aggressive messages (e.g., [4]). These features were extracted using NLI. NLI was applied here similarly to what was done with actions: a sentiment-related hypothesis was given (e.g., “This text is a positive reaction”), and the related NLI-provided probabilities were gathered for positive and negative reaction hypotheses for each message. A

7We will publish our code on GitHub as open source code: https://github.com/Aalto-CRAI-CIS/Detecting-covert-disruption
8If the user for whom the response-expecting message was intended did not take part in the conversation at all after the message (i.e., if there were no messages from the user to anyone after the message was posted), this was not counted as a missing response. Only when the targeted user continued to participate in the conversation but chose not to respond to the message in question did this count as a missing response.
9https://www.perspectiveapi.com/
### Table 4. Summary of Features Extracted from the Datasets

<table>
<thead>
<tr>
<th>Feature group (N)</th>
<th>Calculation principle</th>
<th>Feature group (N)</th>
<th>Calculation principle</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentiments</strong></td>
<td></td>
<td><strong>Action-based features</strong></td>
<td></td>
</tr>
<tr>
<td>Blacklist words (1 feature)</td>
<td>Number of offensive words (based on [76]) found within the conversation, divided by the total number of messages.</td>
<td>Main actions (10 features)</td>
<td>The number of times a main action occurs within the conversation, divided by the total number of actions.</td>
</tr>
<tr>
<td>Toxicity (2)</td>
<td>Total conversation toxicity and mean toxicity (toxicity per number of messages) of the conversation.</td>
<td>Action probabilities (10)</td>
<td>Mean NLI action probability across messages within a conversation.</td>
</tr>
<tr>
<td>Affect in reactions (3)</td>
<td>Number of positive, negative and neutral reactions within the conversation, divided by the number of messages.</td>
<td>Asymmetries Asymmetries per normative actions (15)</td>
<td>For each normative action ($N = 5$; see Table 3, column 2): the number of mismatched responses, missing responses, total asymmetric pairs (sum of the previous two), divided by number of actions in the conversation.</td>
</tr>
<tr>
<td>Probability of affect (4)</td>
<td>Mean NLI probability of positive and negative reactions within the conversation.</td>
<td>Probability of agree (4)</td>
<td>Mean NLI probability of positive and negative reactions within the conversation.</td>
</tr>
<tr>
<td><strong>POLPR</strong></td>
<td></td>
<td><strong>Total symmetries</strong> (1)</td>
<td>Total number of symmetric action pairs.</td>
</tr>
<tr>
<td>Politeness (21)</td>
<td>Number of times each politeness strategy is used in the conversation.</td>
<td>Total asymmetries (3)</td>
<td>Total number of missing responses, mismatched responses, total asymmetric pairs (sum of the previous two), divided by number of actions in the conversation.</td>
</tr>
<tr>
<td>Prompts (6)</td>
<td>Number of times each Prompt type is used in the conversation.</td>
<td>Total symmetries (5)</td>
<td>The number of symmetric responses to each normative action ($N = 5$; see Table 3, column 2), divided by number of actions.</td>
</tr>
<tr>
<td><strong>Agreement</strong></td>
<td></td>
<td><strong>Symmetries</strong></td>
<td></td>
</tr>
<tr>
<td>Agreement level (3)</td>
<td>Number of agreeing, disagreeing and neutral messages within the conversation, divided by number of messages.</td>
<td>Symmetries per normative actions (5)</td>
<td>The number of symmetric responses to each normative action ($N = 5$; see Table 3, column 2), divided by number of actions.</td>
</tr>
<tr>
<td>Probability of agreement (2)</td>
<td>Mean NLI probability of agreement and disagreement in reactions within the conversation.</td>
<td>Total symmetries (1)</td>
<td>Total number of symmetric action pairs.</td>
</tr>
<tr>
<td><strong>Troll mentions</strong></td>
<td></td>
<td><strong>Action pairs</strong></td>
<td></td>
</tr>
<tr>
<td>Troll mentions (2)</td>
<td>Number of trolling mentions and number of different users mentioning trolling in the conversation, divided by number of messages.</td>
<td>Replies to actions (5)</td>
<td>Number of any types of responses to normative actions (asymmetric or symmetric), divided by number of actions.</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td><strong>Reply time (1)</strong></td>
<td>Average time in seconds passed between a normative action and its response.</td>
</tr>
<tr>
<td>Message length (1)</td>
<td>Mean message length within the conversation (in characters including spaces).</td>
<td>Semantic textual similarity (3)</td>
<td>Mean cosine similarity between symmetric message pairs, mismatching asymmetric pairs, and other non-normative pairs (see Table 3, column 2), messages represented as sentence embeddings (using sentence-BERT: [88])a.</td>
</tr>
<tr>
<td>Messages (1)</td>
<td>Number of messages in the conversation.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Users (1)</td>
<td>Number of users in the conversation.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ahttps://github.com/UKPLab/sentence-transformers

A message was classified as negative or positive if one of these probabilities was significantly higher than the other. This means that a message was classified as neutral when NLI’s probabilities for positive reaction and negative reaction differed by 0.2 or less, thus suggesting that either label was close to being likely. This NLI-based tagging task was evaluated by a small hand-labelled dataset.
Detecting Covert Disruptive Behavior in Online Interaction

Politeness and prompt types (POLPR\(^{10}\)). We also extracted politeness strategies and prompt types used in the conversation, similar to previous research [126]. For their extraction, for both datasets (aggression and trolling), we employed Convokit libraries [16], similar to Zhang et al. [126].

Trolling mentions. Trolling-related mentions (e.g., "Don’t troll") were counted per a dictionary of trolling-related words found within the conversation (based on [53]; see Appendix E), as they have been considered important for identifying trolling [53, 78, 79]. Text preprocessing was also needed to extract these features (see Appendix D).

Agreement. We also added three features related to the level of agreement in a message (agreeing, disagreeing, or neutral). This was to allow comparing our results to earlier research that has deemed the level of agreement as important in studying online disputes [115]. The comparison interested us due to findings indicating that the measure of disagreement–agreement within a conversation does not necessarily correspond with the constructiveness of a conversation or vice versa (see [122]). These features were extracted using NLI, exactly like in the case of NLI-based tags for sentiment in reactions to others’ messages: the hypothesis being related to agreement in this case (e.g., “This text is an agreement.”). The message was tagged as agreeing or disagreeing if one of these probabilities was significantly higher than the other; a message was classified as neutral when NLI’s probabilities for agreement and disagreement differed by 0.2 or less, thus suggesting that either label was close to being likely. This NLI-based tagging task was also evaluated by a small hand-labelled dataset (50 examples per category), performance reaching 0.93 accuracy and 0.93 macro-F1 (see Appendix A).

4.3 Step 3 – Analysis of Data

Next, we conducted analyses on the extracted datasets to answer our three RQs. We first removed 905 conversations from our trolling data, mostly from the non-trolling Reddit subset, as these included unsuitable content (non-English language texts or deleted messages) or were considered outliers, being much longer than most conversations in our data – longer than 100 messages. This resulted in a final analysis dataset including 257 trolling conversations and 3,749 non-trolling conversations.

To test group differences (stochastic equality) between aggression versus nonaggression and between trolling versus non-trolling (for RQ1), we used the generalized non-parametric Mann–Whitney–Wilcoxon test, also known as the Brunner–Munzel (B–M) test. We discarded the use of a non-parametric \(U\) test because Shapiro’s test of normality was \(p < 0.001\) for each variable measured, the variables had non-equal variances, and were also skewed. The B–M test has been reported working robustly also in such situations [13, 66, 80]. For this test, we used the brunner-munzel.test() function provided in the lawstat package [60] for \(R\). It comes with an in-built calculation of probabilistic effect size, which is deemed suitable for assessing effect size when using the B–M test [66].

Because the test is rarely used and as the trolling dataset was highly imbalanced, for comparability we conducted also permutation testing [117]. We used the \(R\) function permTS() found in the perm package [41]. The permutation test does not make assumptions about the distribution of the underlying data [41, 117], and it can infer whether the responses of the dependent variable

\(^{10}\)We refer to politeness strategy and prompt type features here with POLPR. They correspond to the same features that were used on the same aggression dataset in the original study in which the dataset was published [126]. The original study used these features to predict aggression; we use these features on the same dataset as a point of comparison, referring to them from now on as POLPR features.
among groups are equal. We conducted permutation tests using 1,000 random permutations with $\alpha = 0.05$ (type I error probability).

To identify the most important predictors out of candidate features listed in Table 4 for aggression and trolling (for RQ2), we conducted feature-forward stepwise logistic regression selection with the stepAIC algorithm [89], where features were incrementally added to an originally empty model ([44]), based on which feature is the best predictor at each point. The Akaike Information Criterion (AIC) was used to limit the size of the final model (number of predictors) so that during the process, the model chose the best predictors, stopping when reaching the lowest AIC, i.e., the best fitting model with least information loss [44]. Thus, it output a selection of features that provided the best predictor features for distinguishing between two classes. In this analysis, we used the R package MASS [89], and a logistic regression model, utilizing functions stepAIC() and glm(). StepAIC avoids multicollinearity by removing highly correlated predictor variables from the final model. The selected features were later used in our ML tasks.

RQ3 was dealt with by training classifiers and evaluating them. Using Python scripts, we trained two separate classifiers to distinguish between (1) aggressive conversations versus non-aggressive conversations and (2) conversations containing trolling versus non-trolling conversations. We used the feature sets extracted in step 2 to train the two classifiers. Using sci-kit learn randomized split (train test split) [84] twice, we separated the dataset into train, validation, and test sets, with respective set sizes 60%, 20%, and 20%. The ML models used were Logistic Regression, Random Forest, XGBoost, and K-Nearest Neighbors (KNN) (from sci-kit learn: [84]). These models were selected as they have also been used in previous related research (e.g., [78, 116, 126]). In addition, their operational logic is simple enough for a study like ours that is interested in analyzing features, not for finding the optimally performing model. The models’ hyperparameters were optimized using the sci-kit learn random search algorithm [84] and Bayesian search [100]. We normalized the action-based features per all actions (per number of messages for other features, see Table 4) so that all action features represent their proportion of the measured action per number of all actions in the conversation. We also experimented with scaling, but this was inferior to the aforementioned normalization strategy. We also used the Nemenyi test (sci-kit-posthocs implementation for Python [109]), which has been recommended for comparing the performances of different ML models on the same task [33], to compare the Macro-F1 performances across the folds between the model using POLPR+Tox features versus the model using POLPR+Tox+Actions. The aim was to statistically compare the gains from adding conversational action-based features into trolling detection, which is the main focus of this article.

For the feature-forward selection set in Table 7, we used the most important features from analysis done related to RQ2. We used the features drawn from forward selection to train models using the statistically optimal features for both datasets. For trolling data, we used upsampling and downsampling similarly to Chawla et al. [19] for data augmentation, to account for the imbalance between trolling and non-trolling data. In data augmentation, the majority class (non-trolling) was undersampled first (down to 1,047 conversations for the training set). Then, the minority class was oversampled to increase its size to fit the minority group using the SMOTE algorithm [19]. The downsampling and upsampling sizes were based on performance results reached in earlier research [19]. Oversampling was applied to the training set. This has been shown to be an effective method to even out imbalanced datasets [19]. This, together with SMOTE resampling, resulted in a balanced set of 2,094 conversations in total in the training set.

5 RESULTS
The following three subsections will present the findings related to the article’s three RQs: how common asymmetric and symmetric actions are in online aggression and trolling, how aggression
and trolling conversations differ in most informative predictive features, and whether our conversational features relating to actions and norm violations can improve the detection of these behaviors. Following the best practice of using separate datasets for training, validation, and testing, the classification results in these subsections are obtained using the test dataset.

5.1 RQ1: Do Aggression and Trolling Involve More Asymmetry than Constructive Interaction?

Our first RQ asked whether disruptive online conversations contain more asymmetry in terms of conversational action pairs when compared with well-intending discussion, i.e., whether aggression contains more asymmetry than non-aggression, and trolling more than non-trolling. For instance, to analyze whether symmetric responses to accusations are different in aggression as compared with non-aggression, we first computed the percentage of symmetric responses in each conversation by dividing the number of symmetric responses with the total number of actions in the given conversation. We then compared the distributions of these percentages between conversations with and without aggression. We used both permutation tests and the Generalized Mann–Whitney–Wilcoxon (i.e., B–M) test [66] to test for the equality of the group distributions with the null hypothesis that the distributions would be equal. Table 5 shows both the significance levels and effect sizes for these tests. Statistically significant results indicate cases in which the percentages of an action pair were different in disruptive and non-disruptive conversations. Effect sizes below 0.5 point out the action pairs in which asymmetries (or symmetries) were more frequent in disruptive conversations. Effect sizes above 0.5 indicate higher frequencies in non-disruptive conversations.

Table 5 shows differences in the percentages of (a)symmetric action pairs in both datasets. For aggression, the B–M and permutation tests were closely in line with each other. Unexpectedly, the overall percentage of symmetric action pairs was slightly lower (28.2) in conversations without aggression than in conversations with aggression (29.0). This may be due to accusations and requests getting symmetric responses more often in aggressive conversations. Also, overall, accusations seem to be more common in disruptive conversations (see Appendix B).

Also, the findings regarding asymmetries involved some unexpected results in our aggression dataset. Among those differences that proved statistically significant, in aggressive conversations, accusations and requests were more frequent both in mismatched and missing responses (i.e., their effect size was below 0.5). In other response types, however, asymmetries were more frequent in non-aggressive conversations. Finally, overall, the percentages of asymmetric responses seemed higher for non-aggression both in mismatched responses (12.4 vs. 12.0) and in missing responses (31.7 vs. 31.4). These differences were not, however, statistically significant. Overall, effect sizes for the aggression task were quite close to 0.5.

However, for trolling, the results proved more interesting and were closer to our expectations. Asymmetries were more common in trolling overall. Among asymmetric mismatched responses, based on the B–M tests and effect sizes, asymmetries in all action pairs except for proposals were more frequent in conversations that contained trolling. Asymmetric missing responses yielded the same result. Also, the three comparisons of totals in asymmetries provided similar statistically significant results. The more conservative permutation tests corroborate the results for all totals and accusations as well as for missing responses to questions and proposals.

The effect sizes, on the other hand, were not strong in many cases. While the effect sizes were between 0.379 and 0.435 for responses to accusations, mismatched or missing responses to questions, and comparisons of totals, the other values were close to 0.5 even in cases in which a significant difference in frequencies was observed. Also, in light of the permutation tests, requests and
Table 5. Empirical Probabilities of Symmetric and Asymmetric Action Pairs in All Datasets

<table>
<thead>
<tr>
<th>Action pair</th>
<th>Non-effect</th>
<th>Effect</th>
<th>Non-trolling</th>
<th>Effect</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aggression</td>
<td>$p^a$</td>
<td>size$^b$</td>
<td>Perm.$^c$</td>
<td>size$^b$</td>
</tr>
<tr>
<td>Symmetric pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Questions</td>
<td>3.5</td>
<td>4.5</td>
<td>0.502</td>
<td>4.7</td>
<td>***</td>
</tr>
<tr>
<td>Accusations</td>
<td>0.4</td>
<td>0.7</td>
<td>**</td>
<td>0.490</td>
<td>**</td>
</tr>
<tr>
<td>Requests</td>
<td>1.8</td>
<td>2.2</td>
<td>**</td>
<td>0.482</td>
<td>*</td>
</tr>
<tr>
<td>Challenges</td>
<td>10.4</td>
<td>10.8</td>
<td>0.490</td>
<td>6.1</td>
<td>6.5</td>
</tr>
<tr>
<td>Proposals</td>
<td>3.1</td>
<td>2.7</td>
<td></td>
<td>0.513</td>
<td>*</td>
</tr>
<tr>
<td>All together</td>
<td>28.2</td>
<td>29.0</td>
<td>*</td>
<td>0.480</td>
<td>*</td>
</tr>
<tr>
<td>Asymmetric mismatched responses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Questions</td>
<td>4.5</td>
<td>4.3</td>
<td>0.506</td>
<td>3.5</td>
<td>4.6</td>
</tr>
<tr>
<td>Accusations</td>
<td>1.3</td>
<td>2.0</td>
<td>***</td>
<td>0.480</td>
<td>***</td>
</tr>
<tr>
<td>Requests</td>
<td>5.2</td>
<td>5.7</td>
<td>**</td>
<td>0.480</td>
<td>3.1 3.2</td>
</tr>
<tr>
<td>Challenges</td>
<td>17.6</td>
<td>17.3</td>
<td>0.505</td>
<td>7.0</td>
<td>7.5</td>
</tr>
<tr>
<td>Proposals</td>
<td>10.7</td>
<td>9.1</td>
<td>*</td>
<td>0.520</td>
<td>**</td>
</tr>
<tr>
<td>All together</td>
<td>12.4</td>
<td>12.0</td>
<td>0.508</td>
<td>7.2</td>
<td>8.1</td>
</tr>
<tr>
<td>Asymmetric missing responses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Questions</td>
<td>5.1</td>
<td>5.0</td>
<td>0.505</td>
<td>4.3</td>
<td>5.5</td>
</tr>
<tr>
<td>Accusations</td>
<td>0.7</td>
<td>1.7</td>
<td>***</td>
<td>0.437</td>
<td>***</td>
</tr>
<tr>
<td>Requests</td>
<td>4.0</td>
<td>4.6</td>
<td>***</td>
<td>0.466</td>
<td>2.4 2.2</td>
</tr>
<tr>
<td>Challenges</td>
<td>17.0</td>
<td>16.3</td>
<td>*</td>
<td>0.520</td>
<td>7.7 8.9</td>
</tr>
<tr>
<td>Proposals</td>
<td>4.9</td>
<td>3.8</td>
<td>***</td>
<td>0.538</td>
<td>2.5 1.7</td>
</tr>
<tr>
<td>All together</td>
<td>31.7</td>
<td>31.4</td>
<td>0.506</td>
<td>19.0</td>
<td>21.7</td>
</tr>
<tr>
<td>All asymmetric action pairs</td>
<td>44.1</td>
<td>44.1</td>
<td>0.509</td>
<td>28.2</td>
<td>29.9</td>
</tr>
</tbody>
</table>

Calculated by counting the number of (a) symmetric pairs per number of actions in individual conversations and then taking the average across all conversations to obtain dataset level measures.

- $p^a$: Significance result of the Generalized Mann–Whitney–Wilcoxon test (i.e., Brunner–Munzel test); * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
- $p^*$: Computed as stochastic superiority estimate $\hat{p}^*$ from the Brunner–Munzel test using R package lawstat. Compared with Cohen’s $d$, which assumes normality, $\hat{p}^*$ is a non-parametric estimate that is linked to stochastic equality where the null hypothesis is $P(X > Y) = P(X < Y)$. Its corresponding test statistic is $\hat{p}^* = (P(X < Y) + 0.5 P(X = Y)) / 0.5$. The effect size is measured by the amount by which the estimate $\hat{p}^*$ deviates from 0.5: If $X$ and $Y$ are stochastically equal, $\hat{p}^* = 0.5$. If $P(X < Y) > P(X > Y)$, then $\hat{p}^* > 0.5$. Similarly, if $P(X < Y) < P(X > Y)$, then $\hat{p}^* < 0.5$. A detailed presentation of this estimate is available in [66].
- Perm.$^c$: Permutation test.

Challenges did not have statistically significant differences in asymmetric responses in trolling versus non-trolling and, in the case of the permutation tests for mismatched responses, only accusations got a statistically significant result in the comparison. As for requests and challenges, the results may be partly due to the fact that the action tagger had lower performance in identifying these actions (see Appendix A). This might have affected the identification of (a)symmetries in these cases as well as the statistical tests. In fact, when we investigated the action tagging results for requests more closely, the tagger often mistakenly interpreted questions or even challenges as requests or proposals. For these latter classes, norms related to action pairs and symmetric responses would be different (especially so for questions; see Table 3), which likely also affected the results.

Surprisingly, although this was not part of our hypotheses, many symmetric responses were more frequent in trolling as compared with non-trolling. One possible reason can be that these actions (questions, accusations, and challenges, to be specific) were also more frequent in trolling (see Appendix B). It might also be due to the fact that we are considering whole conversations instead of individual user behaviors. While norm violations in action pairs might be more common for troll users (or aggressive users), the same might not be the case for non-troll users engaging
in the conversation. The opposite might even be true: non-troll users might even start responding to other actions in an overtly symmetric manner to hold the norm violator accountable, as can be seen in some examples in studies by Hardaker [53] and Paakki et al. [82].

The results seem to support the importance of paying attention to the principle of symmetry between conversational actions (RQ1; see Section 2.1) in trolling but not for aggression. Trolling conversations are more asymmetric in terms of action pairs as compared with non-trolling, although this does not apply similarly to all actions we examined. The results also emphasize the need to recognize the differences in the frequencies of specific action types and how they are used in conversation.

5.2 RQ2: Are Important Conversational Features Such As Actions and Norm Violations Different in Online Aggression As Compared to Trolling?

Our second RQ asked whether there are significant differences in important features, both related to conversational norm violations (asymmetries in action pairs) and in conversational features in general (such as frequencies of different individual actions). With this question, we were interested in finding evidence for our claim that aggression and trolling should be treated as different phenomena, which would also be visible in their features.

The features extracted from the action-tagged data allowed comparisons between most frequent action types and (a)symmetries in aggression on the one hand and in trolling on the other hand. We were also able to compare the importance of other features, including sentiments, different politeness strategies, prompt types, and agreement between the datasets. We started by conducting a stepwise feature-forward selection (stepAIC) using logistic regression to find the most important features out of all our conversation features (see Table 4) that set apart aggression versus non-aggression, and trolling versus non-trolling conversations. We set out to prove that the most important features would be different for trolling in contrast to aggression.

Table 6 presents the results. Out of all the 47 comparison and 53 action-based features in total (see Table 4), the table lists the features that proved optimal in stepwise feature-forward selection for distinguishing aggression from non-aggression and trolling from non-trolling. The algorithm selected 32 features for aggression and 29 for trolling.

The results suggest that the two types of disruptive behavior are best characterized by different feature sets. In light of comparison features, sentiment and POLPR features are more predictive in aggression than in trolling. Sentiment features are especially important in aggression detection (1st, 2nd, and 9th places). This is in line with earlier research [4, 116]. For trolling detection, in contrast, sentiment features are less essential, with blacklist words reaching only the 18th position. POLPR features, in turn, appear important in both detection tasks, with 13/32 (41%) of the features in aggression detection and 9/29 (31%) features in trolling belonging to that category.

However, most important to RQ2, the table shows that many features related to actions and (a)symmetries carry significant weight in both feature sets, although differently. Action features (i.e., specific individual action frequencies as well as their replies) are important for the aggression task: denials, appreciations, and apologies can be found at the 8th, 10th, and 20th places, and there are 8 action features other than asymmetries in total on the list. With regard to specific actions, the numbers of types of replies to specific actions, i.e., accusations, questions, and challenges, were informative. These actions are seen in the literature to have face-threatening qualities [11, 45]; thus, in ordinary well-intending discussion, people tend to avoid them. Several such actions were more frequent in aggression as compared with non-aggression (see Appendix B). Symmetries and asymmetries, on the other hand, appear lower in the list. Only 5/32 (16%) of the top features as listed in Table 6 for aggression are related to symmetries or asymmetries altogether.

For trolling, the findings are different: action-based features carry more weight than in aggression detection. Symmetries and asymmetries, as well as other action-based features, are central
Table 6. Most Important Features across Both Datasets using Stepwise Feature Forward Selection

<table>
<thead>
<tr>
<th>Feature Category</th>
<th>Weight</th>
<th>Feature Category</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Toxicity Sentiments</td>
<td>0.449</td>
<td>Users mentioning trolling</td>
<td>0.189</td>
</tr>
<tr>
<td>2. Toxicity mean Sentiments</td>
<td>1.347</td>
<td>Probability of accusation Individual actions</td>
<td>0.043</td>
</tr>
<tr>
<td>3. Users in conversation Other</td>
<td>−0.046</td>
<td>Troll mention count Troll mentions</td>
<td>−0.058</td>
</tr>
<tr>
<td>4. Has hedge POLPR</td>
<td>−0.019</td>
<td>Prompt: Procedures POLPR</td>
<td>−0.013</td>
</tr>
<tr>
<td>5. 2nd person start POLPR</td>
<td>−0.036</td>
<td>Please POLPR</td>
<td>0.020</td>
</tr>
<tr>
<td>6. Mean cosine similarity between matching action pair parts</td>
<td>−0.223</td>
<td>Apologies Individual actions</td>
<td>0.025</td>
</tr>
<tr>
<td>7. 1st person POLPR</td>
<td>−0.018</td>
<td>Mismatched replies Asymmetries</td>
<td>0.037</td>
</tr>
<tr>
<td>8. Probability of denial Individual actions</td>
<td>0.145</td>
<td>Mean cosine similarity between matching pair parts</td>
<td>−0.041</td>
</tr>
<tr>
<td>9. Blacklist words Sentiments</td>
<td>−0.002</td>
<td>Symmetric action pairs Symmetries</td>
<td>0.053</td>
</tr>
<tr>
<td>10. Appreciations Individual actions</td>
<td>−0.043</td>
<td>Accusations Individual actions</td>
<td>0.137</td>
</tr>
<tr>
<td>11. Mismatched responses to accusations Asymmetries</td>
<td>0.045</td>
<td>Indirect greeting POLPR</td>
<td>−0.033</td>
</tr>
<tr>
<td>12. 2nd person POLPR</td>
<td>0.013</td>
<td>Questions Individual actions</td>
<td>0.079</td>
</tr>
<tr>
<td>13. Prompt: Casual POLPR</td>
<td>0.018</td>
<td>1st person plural POLPR</td>
<td>−0.003</td>
</tr>
<tr>
<td>14. Symmetric replies to proposals Symmetries</td>
<td>0.014</td>
<td>2nd person POLPR</td>
<td>0.007</td>
</tr>
<tr>
<td>15. 1st person start POLPR</td>
<td>−0.012</td>
<td>Factuality POLPR</td>
<td>−0.005</td>
</tr>
<tr>
<td>16. Factuality POLPR</td>
<td>0.013</td>
<td>Users in conversation Other</td>
<td>0.003</td>
</tr>
<tr>
<td>17. Replies to challenges Action pairs</td>
<td>−0.015</td>
<td>Has hedge POLPR</td>
<td>−0.005</td>
</tr>
<tr>
<td>18. Indirect POLPR</td>
<td>−0.074</td>
<td>Blacklist words Sentiments</td>
<td>0.001</td>
</tr>
<tr>
<td>19. Has positive POLPR</td>
<td>−0.006</td>
<td>Statements Individual actions</td>
<td>0.060</td>
</tr>
<tr>
<td>20. Apologies Individual actions</td>
<td>−0.051</td>
<td>Prompt: Casual POLPR</td>
<td>0.007</td>
</tr>
<tr>
<td>21. Replies to accusations Action pairs</td>
<td>−0.040</td>
<td>Message count Other</td>
<td>0.003</td>
</tr>
<tr>
<td>22. Missing replies to proposals Asymmetries</td>
<td>−0.017</td>
<td>Probability of negative reaction Sentiments</td>
<td>−0.030</td>
</tr>
<tr>
<td>23. Announcements Individual actions</td>
<td>−0.028</td>
<td>Subjunctive POLPR</td>
<td>−0.017</td>
</tr>
<tr>
<td>24. Probability of statement Individual actions</td>
<td>−0.099</td>
<td>Missing responses to accusations Asymmetries</td>
<td>−0.004</td>
</tr>
<tr>
<td>25. Missing replies to requests Asymmetries</td>
<td>0.012</td>
<td>Neutral valence Sentiments</td>
<td>−0.004</td>
</tr>
<tr>
<td>26. Prompt: Editing POLPR</td>
<td>−0.010</td>
<td>Symmetric responses to proposals Symmetries</td>
<td>0.106</td>
</tr>
<tr>
<td>27. Apologizing POLPR</td>
<td>0.024</td>
<td>Mismatched responses to proposals Asymmetries</td>
<td>−0.041</td>
</tr>
<tr>
<td>28. Disagreement Agreement</td>
<td>−0.009</td>
<td>Missing responses to challenges Asymmetries</td>
<td>−0.046</td>
</tr>
<tr>
<td>29. Has negative POLPR</td>
<td>0.007</td>
<td>Mismatched responses to accusations Asymmetries</td>
<td>0.006</td>
</tr>
<tr>
<td>30. Replies to questions Action pairs</td>
<td>0.007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31. Symmetric replies to accusations Symmetries</td>
<td>−0.015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32. Direct question POLPR</td>
<td>0.009</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the order in which they were selected starting with the first feature at the top. Category = Category in Table 4; Weight = the weight coefficient of each feature in the final model. NRO = Number of. Action-based features have been emphasized.
in trolling detection: 7/29 (24%) of the top features are related to symmetries or asymmetries in action pairs, and individual action features and other action pair–related features appear 6 times, resulting in 13/29 (45%) action-based features in total. Other action features besides (a)symmetries do not consist of merely directly face-threatening actions or replies to them; rather, they consist of textual symmetry of matching pairs, number of questions asked, apologies, and statements made. Overall, trolling seems to include more accusations as compared with non-trolling (see Appendix B). Besides norm violations in action pairs (e.g., mismatched replies), the use of accusations, types of responses to them, and asymmetries related to accusations are informative. Some politeness features also carry weight in separating trolling from non-trolling conversations, as they do for distinguishing aggression from non-aggression, e.g., hedging and directness.

To sum, the results speak for the relevance of conversational features in trolling detection but to a lesser extent in aggression detection. Our results differ from earlier research findings in some respects. This demonstrates that comparing aggression and trolling has been fruitful. First, the level of disagreement is informative in aggression prediction, in contrast to the findings of Zakharov et al. [122]. Second, trolling mentioned by several users is significant, although it needs to be noted that this happens only in about 26% of all trolling conversations (see Appendix C). By nature, in many cases, troll mentions are indicators of trolling but might not necessarily constitute cases of trolling themselves. Third, while Zhang et al. [126] reported that both politeness and prompt types are important in aggression, Table 6 shows that politeness features are more prominent than prompt features for both datasets. 11 Although trolling has been seen as a phenomenon that cannot solely be studied based on politeness since there are many more subtle layers to the deceptive phenomenon (e.g., [53]), POLPR features also seem to be important here in addition to action-based features.

Finally, while aggression and trolling have often been analyzed together (see the Introduction and Section 2.2), the results demonstrate their differences at the feature level. Aggression and trolling differ by their sentiment (e.g., toxicity is a top feature for aggression but does not appear in the trolling list), although negative sentiment has been a central feature in previous studies, both in aggression and trolling detection [4, 14, 15, 59]. These results illustrate the differences between aggression and trolling, and emphasize that sentiment measures alone are not enough to detect especially covert disruptive behaviors (trolling) online.

Overall, the results demonstrate the differences between conversational strategies and directness (principle of conversational distinctiveness) in these disruptive behaviors (RQ2): aggression and trolling do, indeed, involve significantly different important features and norm violations. This can be seen in their use of specific individual actions, (a)symmetries, politeness, and sentiment. Aggressive interaction is characterized most importantly by sentiment-related features, pair similarity between matching action pairs, (lack of) reconciling actions, and the use of face-threatening actions. Some (a)symmetries are important in the task, however, less so as compared with trolling. Conversations involving trolling are defined largely by various asymmetries such as mismatched responses, the use of specific actions such as accusations (and related asymmetries), and politeness strategies such as indirect greetings and hedging. These can be seen as more indirect strategies to direct or manipulate conversation, ones that the perpetrator can shrug off as unintended or misinterpreted, and that cannot be concretely pointed out, such as offensive language. We conclude that trolling behavior is thus essentially responsive and can be recognized by elusive norm violations in responses to other users.

11Prompts are marked among POLPR features with “Prompt:” in front of a feature name. Other POLPR features are all politeness features.
5.3 RQ3: Do Features Related to Conversational Actions and Norm Violations Allow More Accurate Detection of Disruptive Interaction?

Our second RQ addressed ML-based disruptive behavior detection and whether conversational features related to actions and their norm violations, together with other relevant features, would allow accurate detection of online conversations involving aggression or trolling. Also here, following our interest in detecting covert behavior in particular, we were primarily focused on the models’ performance in trolling detection. We also report the results for aggression detection in order to support comparison with prior works and to evaluate whether the two behaviors should be treated as different phenomena.

To compare our action-based analysis with earlier research, we trained models with increasingly comprehensive feature sets. In aggression detection, we used the feature set of Zhang et al. [126] (politeness and prompt types; POLPR) in order to include features used in earlier research. Since we used their dataset, we wished to use similar features. However, we stress that the original study by Zhang et al. [126] sought to predict a toxic ending to a conversation, whereas we sought to distinguish aggressive conversations from non-aggressive conversations. Thus, the performances of the models should not be directly compared. After POLPR and toxicity features, we added our conversational action feature sets to investigate how action-based features and norm violation features would affect the performances of ML models in detecting disruptive behaviors in online conversations. For trolling data – for which suitable prior baseline data were not available – we used the same feature sets but also included trolling mentions in the message-level feature set since earlier research has reported on their importance [78] and because our findings for RQ2 also demonstrated the importance of this feature. We will first introduce the overall results of our classification tasks and then discuss further details related to both tasks in more depth. Table 7 presents the results of both classification tasks.

For aggression detection, we obtained accuracy levels in the range 0.64–0.69 for the POLPR model. Our best models used POLPR + Tox + Actions and the feature-forward selection, both of which included a combination of action (Actions), toxicity (Tox), politeness and prompt type (POLPR) features. These surpassed the POLPR model with an accuracy of 0.90. The other key measures (recall, precision, and $F_1$ score) of the best performing aggression detection model are presented in Table 8. In the buildup towards the best model in aggression detection, toxicity features (Tox + Actions, or POLPR + Tox) offered the most prominent accuracy improvement in all the models. Concerning RQ2’s focus on the contribution of conversational actions, the model using only action-related features (Actions) (accuracy 0.68–0.75) was equal to or slightly better than the POLPR model.

The confusion matrix (Table 9) for aggression classification illustrates that the results for our best aggression detection model (using feature-forward selection) were well balanced for both aggression and non-aggression. The Receiver Operating Characteristics (ROC) Area Under the Curve (AUC) score for the model was 0.95, calculated by repeated stratified K-fold cross-validation using sci-kit learn. To summarize the most important results, aggression classification improved most when using toxicity together with POLPR or conversational action–based features as compared with only using POLPR features.

For trolling detection, we also expected that conversational action-based features would be important. Indeed, Table 7 shows that the best model including action-based features in its feature set can detect trolling with an accuracy as high as 0.92. The action features alone lead to the accuracy of 0.86 already. We conducted the Nemenyi test [33] to compare the Macro-$F_1$ performances across the folds between the model using POLPR + Tox features vs. the model using POLPR + Tox + Actions to confirm whether the addition of conversational action-based features would prove to
Table 7. Classification Accuracy in 5-fold Cross-Validation Tests in Aggression vs. Non-aggression and Trolling vs. Non-trolling Classification using Different Combinations of Feature Sets

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Model</th>
<th>FOLPR</th>
<th>POLPR + Tox</th>
<th>POLPR + Tox + Actions</th>
<th>POLPR + PR + Tox</th>
<th>POLPR + PR + Tox + Actions</th>
<th>POLPR + POLPR + Tox + Actions</th>
<th>POLPR + POLPR + Tox + Actions + All with forward selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggression vs. non-aggression</td>
<td>Logistic regression</td>
<td>0.69</td>
<td>0.87</td>
<td>0.68</td>
<td>0.87</td>
<td>0.71</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>Random forest</td>
<td>0.68</td>
<td>0.87</td>
<td>0.69</td>
<td>0.87</td>
<td>0.70</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.69</td>
<td>0.88</td>
<td>0.75</td>
<td>0.87</td>
<td>0.72</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.64</td>
<td>0.71</td>
<td>0.68</td>
<td>0.62</td>
<td>0.69</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>Trolling vs. non-trolling</td>
<td>Logistic regression</td>
<td>0.69</td>
<td>0.69</td>
<td>0.72</td>
<td>0.71</td>
<td>0.73</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Random forest</td>
<td>0.79</td>
<td>0.80</td>
<td>0.85</td>
<td>0.89</td>
<td>0.89</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.70</td>
<td>0.80</td>
<td>0.86</td>
<td>0.91</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.62</td>
<td>0.64</td>
<td>0.63</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 8. Classification Report for Models That Had the Best Accuracy in Table 7

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Precision</th>
<th>Recall</th>
<th>Micro-F1</th>
<th>Macro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trolling</td>
<td>0.86</td>
<td>0.93</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>Non-trolling</td>
<td>0.97</td>
<td>0.95</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>Aggression</td>
<td>0.88</td>
<td>0.93</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>Non-aggression</td>
<td>0.93</td>
<td>0.87</td>
<td>0.90</td>
<td></td>
</tr>
</tbody>
</table>

Table 9. Normalized Confusion Matrices

<table>
<thead>
<tr>
<th>True label</th>
<th>Non-aggression</th>
<th>Aggression</th>
<th>Non-trolling</th>
<th>Trolling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-aggression</td>
<td>0.87</td>
<td>0.13</td>
<td>0.97</td>
<td>0.03</td>
</tr>
<tr>
<td>Aggression</td>
<td>0.07</td>
<td>0.93</td>
<td>0.09</td>
<td>0.91</td>
</tr>
</tbody>
</table>

The results show a statistically significant difference between the performances of the models \((p < 0.05)\), showing that using action-based features together with features from earlier research can lead to significantly improved results in trolling detection. These results are better than the classification scores achieved by earlier trolling detectors (e.g., [50]: accuracy, 0.78–0.85, \(F1\), 0.79–0.85; [79]: accuracy, 0.58–0.65; [78]: accuracy, 0.81, \(F1\), 0.80). The results are not completely comparable, however, as the definitions of trolling and the units of analysis vary between the studies. However, social media data has been used in all of the mentioned studies, and we have also incorporated features deemed most relevant in previous work into our model for comparability.

A further ablation analysis in Table 10 on best model performances on conversations with different lengths showed that our models performed well on different conversation lengths, although very long conversations tended to be more difficult. The ROC AUC score for the best model (XGBoost) was 0.96. The confusion matrix in Table 9 reports that the results are quite well balanced.
Table 10. 5-fold Cross-Validated Classification Results with the Best Models Using Test Conversations with Different Lengths

<table>
<thead>
<tr>
<th>Dataset Quartile (msgs)</th>
<th>Precision</th>
<th>False Precision</th>
<th>True Recall</th>
<th>False Recall</th>
<th>True Accuracy</th>
<th>Macro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggression vs. non-aggression</td>
<td>Q1 (3–6)</td>
<td>0.94</td>
<td>0.90</td>
<td>0.90</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Q2 (6–7)</td>
<td>0.93</td>
<td>0.84</td>
<td>0.88</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>Q3 (7–9)</td>
<td>0.87</td>
<td>0.80</td>
<td>0.78</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Q4 (9–19)</td>
<td>0.90</td>
<td>0.78</td>
<td>0.68</td>
<td>0.94</td>
<td>0.83</td>
</tr>
<tr>
<td>Trolling vs. non-trolling</td>
<td>Q1 (3–4)</td>
<td>0.94</td>
<td>0.89</td>
<td>0.88</td>
<td>0.91</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Q2 (4–6)</td>
<td>0.93</td>
<td>0.89</td>
<td>0.90</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Q3 (6–11)</td>
<td>0.95</td>
<td>0.85</td>
<td>0.95</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>Q4 (11-79)</td>
<td>0.91</td>
<td>0.84</td>
<td>0.91</td>
<td>0.85</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 11. Classification Results of the Best Model on Conversations Including Trolling Mentions and Conversations without Trolling Mentions

<table>
<thead>
<tr>
<th>Group</th>
<th>Precision</th>
<th>Recall</th>
<th>Micro-F1</th>
<th>Accuracy</th>
<th>Macro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No mentions/non-trolling</td>
<td>0.92</td>
<td>0.97</td>
<td>0.94</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td>No mentions/trolling</td>
<td>0.94</td>
<td>0.87</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>Including mentions/non-trolling</td>
<td>0.60</td>
<td>1.00</td>
<td>0.75</td>
<td>0.98</td>
<td>0.87</td>
</tr>
<tr>
<td>Including mentions/trolling</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
</tbody>
</table>

for the trolling task. It is notable that, in contrast to aggression classification, the accuracy levels did not increase to the same extent when toxicity features were added to POLPR features.

As the forward selection of important features (see Table 6) revealed that the number of different users mentioning trolling or accusing someone of trolling is highly informative, we wished to further ensure that the classifier would also perform well on conversations in which trolling is not mentioned. As mentioned earlier, we considered it important for the classifier to be able to also detect covert trolling attempts that have not been identified by other users (see Section 2.3). This was also because detection based on trolling mentions would be biased, working only to detect trolling already noticed by other users (see Appendix C). Thus, we further tested how our best trolling classifier performed on conversations including trolling mentions or accusations as compared with conversations in which trolling was not mentioned at all. As can be seen in Table 11, the classifier could also accurately detect trolling in conversations in which trolling was not mentioned or no one was accused of trolling. However, the rare cases in which trolling is mentioned in non-trolling conversations seemed to be hardest for the model to identify correctly. Cases of non-trolling with trolling mentions being rare in our data (see Appendix C), we do not consider this a significant caveat.

The results provide evidence of the efficacy of using conversational features related to actions and norm violations together with message-level features to detect especially covert disruptive social media conversations (trolling): specific actions and norm violations in action pairs, complemented by features used in earlier research, allow better performance in trolling detection as compared with earlier research. Although the classification results for aggression are more modest in comparison with earlier detection models, for trolling they surpass earlier models. This supports the importance of tending to the principle of context in disruptive behavior detection (RQ3) but also emphasizes that conversational action features can add significant information to message-level features used in earlier research to more efficiently detect trolling. Action-based classification has advantages due to the following reasons: (a) instead of investigating only message-specific properties, action-based classification focuses on conversational features, i.e., conversational actions taken within
Detecting Covert Disruptive Behavior in Online Interaction

the conversation and the relationships between action pairs; and (b) the analysis of conversational symmetries and asymmetries between action pairs (i.e., dynamics between messages) attends to the particular interaction characteristics typical of especially covert disruptive behaviors (trolling). To sum, conversational action features and norm violations play a key role in our trolling classification task: they allow significantly improved classification results. Thus, instead of focusing solely on message-level features to develop trolling detection models, more focus should be given to conversational features related to conversational actions and norm violations to obtain better results.

6 DISCUSSION

The results of this article emphasize the relevance of linguistic and interaction-based conversational features related to conversational actions and norm violations in analyzing and detecting covert disruptive online conversations. They provide more context for understanding how these phenomena emerge not only through offensive language or negative sentiment but also through violations of conversational norms related to actions. The latter may be acted out, for example, by evading response-expecting actions and social accountability. The results show that analyzing conversational structures, such as action types and the dynamics between action pairs, can reveal differences in their use across different conversational behaviors (aggression, non-aggression, trolling, non-trolling). We verified the validity of our approach by using aggression data from earlier research as a comparison point. Compared with previous studies, conversational action features and norm violations allowed accurate detection of conversations containing aggression and trolling. We demonstrated that our model can especially detect covert behaviors accurately and surpass earlier models in performance; also, different styles of trolling [54] from different contexts were included in the dataset (see [82]). The CA-based computational approach to analyzing social media conversations proposed in this article is a novel method and robust in its action tagging scheme, which is rooted in a well-researched theoretical tradition.

Overall, the results stress the importance of three principles in the detection of covert disruptive conversations: the principle of symmetry between conversational actions, the principle of conversational distinctiveness in conversational strategies used, and the principle of context in interaction. Based on our three RQs and their results, we can further elaborate the need to account for these principles in detection models. To identify covert disruption in particular, it is important to pay attention to norm violations in actions in user interaction, indirect strategies of manipulation, and the wider conversational context of messages. This reveals more subtle or covert norm violations as compared with direct offense, e.g., evasion of accountability demands.

Our results in detecting conversations including covert disruption strategies provide a significant contribution for future detection and prevention of especially covert forms of disruption, such as trolling, on social media. The results emphasize that to have an efficient and sustainable solution to detecting covert attempts at disruption, it is necessary to move beyond the individual message, to look at conversational structure. The types of features we have described in this article could be used to computationally analyze and detect disruptive behavior online. Various norm violations, e.g., in action pairs, are amenable for user-based analysis as well: possible models could produce a probability of a user behaving in an unacceptable way, even during discussion, by using previous turns and inter-message dynamics in the conversation as features. Such features are also more generalizable than word use, for example: common conversational actions and their expectations are features that repeat in various contexts (see, e.g., [39]; although some exceptions can be found: [2]). Identification of a range of conversational features such as covert norm violations could allow detection models that would not be as vulnerable to user deception as message-level detection (see [49]). This, however, is a vein of research that still requires more work.
The results of the ablation study in which we compared classifier performance on conversations with different lengths (Table 10) showed that our model performs well and in quite a stable manner across different conversation lengths. However, very short (3–4 messages) and very long conversations (11–79 messages) tended to be a bit more difficult to classify. Thus, as expected, it seems that a 3-message-long conversation does not always provide enough information. However, with a few more messages, the identification of disruptive behavior seems to become a little easier. For potential approaches in which disruptive behavior could be identified during conversation, we suspect that very early detection, when there are only very few messages in the conversation, might be most challenging, especially when it comes to deceptive behaviors such as covert trolling. In addition, when operating on extremely long conversations, these should perhaps be further analyzed in smaller batches. It must be taken into account in future research that a very long conversation (e.g., including over 100 messages) might include a problematic thread but might not be problematic in its entirety.

6.1 Limitations

Our computational modeling and analysis of online conversations through the lens of conversational actions and asymmetries is still under development. From an applied digital CA perspective, more detailed analyses will be needed to be able to better account for common patterns found in online interaction. Firstly, we could get better insights through even more detailed analysis and distinction of actions and their functions. For instance, accusations that are targeted at a specific person in contrast to general accusations towards a group of people could be analyzed separately. Secondly, the analysis did not consider the existence of insert expansions that are commonly found as injections in between action pairs [96, pp. 97–114]. For example, a question as a response to another question might be acceptable if it contains a request for more detailed information with regard to the original question and the original question is returned to afterwards. In this article, we have supposed that they are less common than action pairs without insert expansions and, thus, in a large conversation dataset should not greatly affect the end result. Moreover, since we assumed a conservative approach in counting asymmetries, e.g., in the case of mismatched responses, future studies could attempt to identify asymmetries in more detail (e.g., between parts of messages).

As no action identification model trained on accusations and challenges was available, the zero-shot approach was our best option at this point. However, although we reached fruitful results in our conversation classification tasks, it should be noted that a zero-shot model is likely to be less accurate than a supervised model trained on annotated data [121]. Zero-shot NLI is also a black box. It lacks transparency; thus, the reasons by which it selects a particular class cannot be scrutinized even if we wished to. For this reason, supervised ML approaches might provide more insight into the analysis of conversational actions. NLI is also computationally demanding and, thus, very slow if we operate on very large datasets; faster models requiring less resources or computation should be developed for such cases. The NLI-based model does not reach a very high performance on our 10-way action classification task. However, we consider its performance to be adequate here for a first model that looks at conversational context as a basis for discerning conversations involving trolling and aggression from more peaceful conversations. Still, there is room for future research to implement more fine-grained high-performance models of action classification for asynchronous forum conversations, including classes such as accusations and challenges that are informative for trolling especially [7, 82]. However, developing such an action classification model was out of the scope of this article, as the effort to create an annotated dataset and to implement a robust model for this purpose would be worth a study of its own. In the future, a better action classification model could be trained incorporating digital CA-inspired action classes: accusations, challenges, and requests for action along with questions, proposals, statements, and their
expected responses. Future research should investigate whether this will yield even better results in social media conversation classification tasks.

Another limitation is the selection of datasets. One concern regarding the aggression data is the length of the discussions: the data by Zhang et al. [126] includes many conversations that are only 3–4 messages long. In the authors’ definition, a conversation that involves a personal attack always starts as neutral but ends up with an attack. It is debatable whether 3–4 messages would count as such a conversation. Therefore, in the future, it may be reasonable to analyze conversations in which the minimum number of turns is higher. Furthermore, datasets pose a limitation as there do not exist many datasets on trolling or aggression that would provide enough conversational information (sequential organization of messages, message and author IDs, and which messages they are responding to). Therefore, it was not possible to find conversational datasets on trolling and aggression that would have been similar enough in terms of discussion topics and source platforms to be perfectly comparable. We also could not find similar enough classification studies that could have been used as a comparison baseline – the most similar study has a slightly different focus [126]. Due to the lack of trolling (or aggression) datasets that would provide enough conversational information to conduct our analysis, we acknowledge that we had to choose datasets that are not exactly similar, as Wikipedia editor conversations may differ in interaction from the types of conversations included in the trolling dataset. Some of the potential datasets we did find either had a completely different definition of trolling [59], failed to include some of the very central messages in their data (as in the Reddit dataset of Zhang et al. [126]) or included multilingual content that we could not analyze at this point in our work [68].

It should also be stressed that, at this point, aggression and trolling data are not fully comparable. This is because the aggression data is solely obtained from conversations among Wikipedia editors that may possess qualitatively different characteristics as compared with the conversations in the trolling dataset. Thus, some of the differences that we observe in our results might be based on the different nature of the types of interaction found in the respective datasets, or, for example, the differences of audiences. In other words, aggressive behavior in interaction might be more or less toxic in Wikipedia editor conversations as compared with newspaper comment sections or Reddit conversations. Furthermore, we have no way to ensure that the aggression data does not contain some amount of trolling, though we ascertained with random checks that the data portrayed mainly aggression. Similarly, the trolling data has not been filtered for aggressive behavior that is not trolling. However, since the main focus of this study is on trolling detection and whether actions and norm violations enable improved performance therein, we feel that this is a minor problem. This also relates to limitations regarding RQ2, which functions in this article to provide more context for RQ1 and RQ3, and to illustrate the different features relevant to trolling and aggression. We acknowledge that the datasets are not equal and, therefore, direct comparisons of the features are not reliable. In order to more reliably study the exact differences between them, further research using similar data with careful annotation practices for both would be needed. This was out of the scope of this article. We feel that future work should investigate how novel datasets could be gathered to allow more versatile conversational analyses of trolling and aggression, and how novel datasets could in the future be used to further corroborate our results.

A third limitation in our study is the small size of the trolling dataset. This could not be avoided because, as of this writing, there exist no large datasets with robust collection criteria, encompassing all known and well-researched trolling styles [54] and including both audience-recognized trolling behavior (that has been filtered to account for audience bias) and behavior that has not been recognized or named by the audience as trolling. Incorporating the latter data is very important, as deceptiveness is at the very core of trolling [36] and, thus, presumably most successful
trolling attempts go unnoticed and unlabeled. The small size of our dataset might limit the generalizability of the models even with data augmentation. Nevertheless, the results are promising.

In future research, larger conversational trolling data should be gathered and the analyses ascertained by running them on the larger-scale dataset. Studies could also explore classification performance differences across different conversation topics, a task that was out of the scope of this article. Furthermore, conversational datasets with multilingual and low-resource language content would be an important next step for which to apply our detection approach and analyses. At the present moment, we have concentrated on English-only data as a relevant first step. However, trolling detection is a task that should be applicable to many language contexts as well as multilingual settings, which is why we consider this an important future direction for research.

6.2 Methodological Insights

Our results provide an interesting and novel angle to the study of computer-mediated communication. Although CMC is said to be more fragmented than ordinary face-to-face interaction (e.g., [56]) and thus not as diligently following the rules or customs of face-to-face conversation, the results in this article prove that well-intending online conversations also adhere to the norms of conversational action, e.g., the rules of action pairs (see also [77]). Norm-violating conversational actions, even in online discussions, are less common in the flow of well-intended conversation. Thus, the evidence here backs up the claim [77] that people orient to regular conversational social norms even when interacting with others on social media. We have shown that, at least to some extent, computational implementation of analyses of actions and action pairs is possible. This opens new possibilities for studying interaction in CMC – and for studying disruptive behavior online. Norm violation can sometimes be a way of using power in interaction, especially when it is systematic. Thus, some of the patterns found in this study could shed more light on how different actions are used to assert power online in trolling, aggressive behavior, political communication or other modes of online interaction attempting to influence other people. Action-based differences in conversational data could be used to differentiate between types of conversations based on their inner dynamics.

Our study reveals that online discussions can be analyzed computationally via conversational actions and their norm violations, which is a novel approach in computational CMC research. It also reveals that the sequential structure of conversation carries relevance in analyzing different online behaviors. In addition to bringing novel findings and methodologies to computational CMC, our study has important implications to applied CA. Firstly, we have shown that as an objective of study, online behaviors such as trolling and aggression can be analyzed and recognized by using CA-based coding schemes, which, to our knowledge, is a new area of application for them. Also, for covert behavior (trolling), we have shown that the approach can outperform earlier approaches. Future work may look at how to better implement more fine-grained CA-based action classification systems, which could result in better identification of trolling. Secondly, this study is an important indicator that applied computational CA, which thus far has been more of a hypothetical than a real field of study, is not only applicable but also integral when we want to detect covert disruptive online behavior (e.g., trolling).

We have successfully combined applied computational CA with other conversational and linguistic measures such as the toxicity of language use, and have shown that our approach may well be the most successful one when identifying covert disruptive online behaviors. Based on the importance of different conversational features, it seems that future research should also investigate how topical digressions play into how disruptive interactions unravel: pair similarities (cosine) between messages were included in the list of important features (Table 6), but the more exact
Detecting Covert Disruptive Behavior in Online Interaction

6.3 Automated Moderation and Social Media

Disruptive behavior online deteriorates online discussions and threatens democratic systems [8], especially in the form of trolling-like manipulative communication strategies (e.g., [1]). Since trolling strategies have been found to be used on the one hand to attack vulnerable groups and minorities (e.g., [56]) and on the other hand to manipulate public opinion and to spread disinformation and uncertainty [6, 102], automated detection of both aggressive and trolling behavior is sorely needed to reduce the impact of harmful and manipulative behaviors on individuals and society. We feel that norm violation–based identification such as what we have presented could be effectively used to find possible systematic manipulation or covert disruption such as trolling in online conversations. It could also be used to prevent the effects of content or sentiment amplification attempts [6], harassment of specific groups of people online (see, e.g., [91]), and possible automated provocation and trolling attempts on conversational forums. The three principles we have defined in this article outline the approach that automatic detection of covert disruptive online behavior could be based on. Models can benefit from identifying covert norm violations of symmetry between conversational actions (principle of symmetry), accounting for indirect strategies in interaction (principle of conversational distinctiveness), and analyzing the wider conversational context around turns (principle of context).

Tools for recognizing disruptive behavior are important for better social media: during significant societal events, there is a need for recognizing harmful manipulative behaviors that attempt to amplify certain messages or manipulate social media behaviors to boost the motivations of political players or other social media influencers (see, e.g., [6]). Recent conversations around social media responsibility and moderation (e.g., [12, 73]) might necessitate more governance and moderation on the side of social media platform providers when it comes to information spread and important social phenomena such as elections or health behaviors. This would further emphasize the need for detecting both covert and overt systematically disruptive behaviors online.

The reasons why automated moderation is important are twofold. First, the simplest manners of violating social norms, such as asymmetric responses, are also easy to automate. Thus, it is likely that these could be used to disrupt and manipulate online conversations around sensitive or political issues. As bots have already been shown to be used in the amplification of political agendas on X [6], it is plausible that bots capable of trolling and aggression on more conversational online forums could also be used in the future for similar purposes. This emphasizes the importance of automated moderation. Moreover, with recent advances in NLP, bots participating in conversations might not be discernible from human participants even with meta-level account information. We feel that, in this case, moderation should be based on systematic display of disruptive behavior, regardless of whether bot or human. Second, earlier research shows that many online conversations are easily volatile [62]. Thus, many platforms are also vulnerable to aggression and trolling if there already is a history of disruptive conversation practices in these discussion environments. Trolling, for instance, is often infectious [21]. This makes it doubly important to moderate.

To improve the identification of disruptive behaviors, future research will likely need to look into various possible information sources besides conversational analyses of interaction, e.g., combining conversation–internal information with conversation–external user information as well as message-level analyses. Although in this article we have stressed the importance of investigating the conversational context past only analyzing standalone messages, we think that message-level analyses are also needed and can be very helpful in some cases, for instance, when there is only one initial potentially provocative message in a thread. In such a case, of course, a conversational
approach would possibly not be needed. However, deceptive behaviors such as covert trolling cannot always be detected if only examining the first message in a thread. Thus, we think that even having just a few messages (and more than one from the troll) will provide more relevant information on potential consistent disruptive behavior than a single message.

Beyond deciding whether a conversation includes disruptive behaviors or attempts to manipulate, the next steps would also involve deciding whether some specific users have acted as instigators of disruption or repeatedly violated community norms, and how to act on these issues. Here, various options are possible: either computational identification of problematic messages and user behaviors within and across conversations based on interactions, produced content, and meta information or identification by human moderators. In cases in which a conversation involves tens or hundreds of messages, a mixed approach might be a viable option.

Assuming that automated moderation of disruptive behaviors is implemented on online forums on a larger and more intricate scale than it currently is, what can be done and how we operate on instances identified as possible trolling remain questions for which in-depth ethical considerations should be initiated [12]. For instance, to discern between instigators of aggression or trolling justly, moderation could be based on the collaboration of human moderators and computational detection systems [63], i.e., a human could make the final decision. Moreover, extreme care is required in assigning an “aggressor” or “troll” label to single users, as this is a grave accusation and erroneous accusations can also have a negative impact on the person in question (e.g., unfair blocking: [64]). Safer and fair options need to be explored. A possibility would be to first inform users themselves of the observation that they have violated community norms (e.g., notifying them of messages they have left unanswered). Then, perhaps, if the behavior continues, there would be a need for human moderators in the loop. As algorithms and automated systems can lead to biased end results, careful use of computational models is needed to avoid unnecessarily placing blame on innocents or individuals who might unwittingly disregard social norms. Ethical design and employment of disruptive behavior detection is a vein of research we feel will be necessary and valuable if and when large-scale automated identification of these disruptive behaviors is to be used on online forums – to maintain civility and platform norms but not to increase censorship.

6.4 Conclusion

These results offer two important contributions to the analysis of covert disruptive behavior in online interaction. First, the conversational approach building on the concepts of conversational action and norm violation offers a new theoretically grounded approach for understanding disruptive online behavior. This offers a significant contribution to developing new computational models for analyzing online discussions and to (automated) data collection for online trolling research, e.g., to help capture a range of different styles of conversational trolling. Second, our computational model for detecting conversations involving trolling demonstrate practical improvements that surpass earlier detection accuracy levels. Our approach paves the way for possible computational models for automated moderation of harmful covert disruptive behaviors, especially trolling, which has so far been difficult to detect automatically. As the main interest in the computational identification of trolling is the prevention of disruption, more real-time methods are needed in contrast to leaked lists of paid trolls or troll mentions that often come too late or identify only a subsection of trolling attempts. Thus, the results of this article offer a significant contribution to developing methods to more accurately detect covert disruption. They also pave the way for possible future models utilizing analyses of conversational actions and norm violations that could help prevent covert disruptive behaviors.
APPENDICES

A EVALUATION MEASURES OF CA-BASED ACTION TAGGING WITH ZERO-SHOT NLI

Recall, precision, and $F_1$ score per tag individually, macro scores* separately for action tagging, and for agreement level, and affect, and specific scores for each tag.**

<table>
<thead>
<tr>
<th>Action Type</th>
<th>Recall*</th>
<th>Precision</th>
<th>$F_1$</th>
<th>Macro Accuracy</th>
<th>Macro $F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>First actions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>question</td>
<td>0.79</td>
<td>0.87</td>
<td>0.81</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>challenge</td>
<td>0.62</td>
<td>0.59</td>
<td>0.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>accusation</td>
<td>0.62</td>
<td>0.65</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>request for action</td>
<td>0.59</td>
<td>0.60</td>
<td>0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>proposal</td>
<td>0.78</td>
<td>0.69</td>
<td>0.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>statement***</td>
<td>0.65</td>
<td>0.70</td>
<td>0.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>appreciation</td>
<td>0.77</td>
<td>0.78</td>
<td>0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Responding actions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>denial</td>
<td>0.64</td>
<td>0.63</td>
<td>0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>apology</td>
<td>0.89</td>
<td>0.93</td>
<td>0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>acceptance</td>
<td>0.81</td>
<td>0.62</td>
<td>0.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complementary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>disagreement</td>
<td>0.91</td>
<td>0.88</td>
<td>0.89</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>agreement</td>
<td>0.65</td>
<td>0.63</td>
<td>0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>negative reaction</td>
<td>0.96</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>positive reaction</td>
<td>0.87</td>
<td>0.93</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Macro scores refer to the overall evaluation scores for a tagging task. For actions, for instance, this means a 10-way task of deciding between all actions included in our action tagging scheme (see Table 3).

**Note that the evaluation is based on our own tests using two small sets of manually action-annotated data: first set n = 30 per action, second set n = 20 per action. The messages annotated this way pertain to the conversations in our dataset and have been extracted randomly from larger data. Thus, the set of messages representing a class may be limited in addition to the size of the test set but also the type of messages representing that action. Thus, this is an approximate evaluation; a more accurate evaluation would need a larger set of actions that includes versatile cases of each class. Scores were counted based on our tagging scheme assumption of one message having two main actions, i.e., if the tested action was among the two most probable actions, the tagging was considered correct. The overall scores are sufficient, although not excellent. Based on our manual evaluations, it seems that the NLI model often had difficulty discerning accusations from challenges, requests from proposals, rejections from denials, acceptances from admissions, and answers to questions from statements. Thus, we combined the latter three cases for the final action tagger. However, more work is needed here.

***The statement class includes informing and assertions, such as statements of information, statements of opinion, announcements, and answers to questions.

B MAIN ACTION FREQUENCIES IN THE CONVERSATION DATASETS

Number of main actions per number of all actions in each individual conversation, mean across all conversations in each dataset. Statistical tests using Generalized Wilcoxon test, aka Brunner–Munzel test (BM), and permutation test (Perm.) a
aStatistical tests were conducted only for actions that are considered to have possible face-threatening implications in conversation: accusation, challenge, denial, request. Similar tests were used here as for testing (a)symmetries between action pairs: Brunner–Munzel tests were used, and permutation tests were used to corroborate the results and more strictly and reliably test the differences between the imbalanced trolling data groups.
bEffect size: for the Brunner–Munzel test, the related function in R packages lawstat and brunnermunzel calculates effect size while conducting the test. The interpretation of this measure is explained in detail by Karch [66]: in short, 0.50 means a similar effect for both groups, a value < 0.50 means the action is more frequent in trolling (or aggression), and a value > 0.50 means that the action is more frequent in non-trolling (or non-aggression).

*p < 0.05, **p < 0.01, ***p < 0.001.

C TROLL MENTIONS

In trolling detection, we also wished to account for the question of troll accusations and troll-related word mentions, which have been central in earlier approaches to sampling, studying, and classifying trolling (e.g., [53, 78]). We measured how many times trolling-related words were mentioned in trolling and non-trolling data (terms based on [53]) in order to get a better idea of how important these are in trolling classification and detection tasks. Judging by the results, as seen below, it is questionable whether troll mentions are a sufficient method for finding trolling in online conversations: most trolling conversations (190/257, 73.92%) in our data do not contain troll mentions. The portion of trolling conversations that do have trolling-related mentions have a considerably higher amount of such mentions as compared with non-trolling conversations that include troll mentions.

<table>
<thead>
<tr>
<th>Behavior</th>
<th># troll mentions</th>
<th># different users mentioning or accusing of trolling</th>
<th># conversations with trolling-related word mentions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trolling</td>
<td>168</td>
<td>135</td>
<td>67 / 257 (26.1%)</td>
</tr>
<tr>
<td>Non-trolling*</td>
<td>6–12</td>
<td>6–9</td>
<td>3 / 257 (1.2%)</td>
</tr>
</tbody>
</table>

*Counted by 5 random samples from the whole non-trolling conversations data, downsampled to the same size as the trolling set to account for the data imbalance.

Although many previous studies have sampled and analyzed trolling based on the mentions of “troll” (e.g., [53, 78]), in this article, using a different sampling method, we showed that conversations in which troll-related terms are mentioned are in the minority. Therefore, it is highly noteworthy to consider the bias that a “mentioned troll”–based data collection approach will inflict on the dataset. It is likely that some specific types of trolling will be especially underrepresented in this type of dataset, which will affect any computational classification models trained on such data. This type of model would also find only a small portion of all trolling-like behaviors.
D  PREPROCESSING

NLI required no preprocessing at our end other than tokenization for the embeddings: for this, we used the BartTokenizer [74], similar to [121].

However, for the extraction of blacklist words and troll mentions, some additional preprocessing was needed:

- Sentence tokenization with NLTK Sentence tokenizer [9]
- Tokenization with NLTK Tweettokenizer [9]
- Stemming with NLTK SnowballStemmer [9]12
- Excluding non-English language content with pypi langdetect.13

E  TROLLING-RELATED TERM DICTIONARY

Hardaker describes a term search for troll mentions in her highly influential paper [53, p. 225]. She first uses WordSmith to search for hits with the term “troll*” 14 as a catch-all search term. She then reviews all results manually, including relevant hits and excluding all irrelevant results. She concludes that, in her data, the following search terms are relevant: troll, trolls, trolling, trolled, derivations (trolly, trollish, trolldom), neologisms (trollometer, trollbait, trollerita) and typographic errors (trollign, troll).

We used a similar approach to find relevant search terms, using Python scripts instead of WordSmith. By examining our data first by using the search term troll*, we examined all results and their relevance manually. Finally, we used the following search terms that came up as relevant in our data: troll, trolling, troller, trolls, trollishly, trollish, trolled, russian-trolls, Russian trolls, troll’s, trolls’, paid-for-trolling, half-troll, full-troll, troll-like, trolling-like, tricktrolling, ricktrolling, rickrolling, facts.troll.

F  ALGORITHM FOR ACTION AND ASYMMETRY FEATURE EXTRACTION

In the pseudo-code below, $A$ refers to the list of actions in Table 3, $N$ is a list of normative, i.e., response-seeking actions, $R$ is a dictionary that uses an action key to list symmetric responses to all normative actions in $N$. $D$ is a dataset, $C$ a conversation in $D$, $M$ a dictionary including all fields related to a forum message that is part of a $C$ (e.g., posting time, user id, message text), and $V$ a feature dictionary of a conversation.

---

12https://www.nltk.org/_modules/nltk/stem/snowball.html
13https://pypi.org/project/langdetect/
14The asterisk (*) wildcard denotes ‘zero or more characters’. Thus, a search for troll* will retrieve troll, trolls, trolling, trolley, etc. [53].
Algorithm for Action and Asymmetry Feature Extraction

Select main actions($M$):

\begin{verbatim}
actionProbs = sort(NLIActionProbabilities ∈ M)
set two highest actionProbs as M[mainActions]
\end{verbatim}

Check if $M$ is mismatched response($C$, $M$):

\begin{verbatim}
if $M$ is a response to a specific message:
  targetM = search specific message ∈ C(C)
  targetActions = targetM[mainActions]
  for $a_i$ ∈ targetActions:
    if $a_i$ ∈ N:
      for $a$ ∈ M[mainActions]:
        if neither $a$ /nelement R[$a_i$]:
          if $M$ /nelement mismatchedMessages add it
          add V[mismatched to $a$] + 1
          else:
            if $M$ /nelement matchedMessages add it
            add V[matched to $a$] + 1
            add V[any response to $a$] + 1
          else if neither $a_i$ ∈ targetActions /nelement N:
            if $M$ /nelement nonNormativeMessages add it
\end{verbatim}

Check if response to $M$ is missing($C$, $M$):

\begin{verbatim}
currentMessage, targetedAuthor = M, M[responseToAuthor]
for $M_i$ ∈ C:
  if $M_i$[Author] == targetedAuthor:
    and if $M_i$ is a response to currentMessage:
      add V[nonmissingResponse] + 1
  else:
    if targetedAuthor ∈ C after currentMessage[postingTime]:
      add currentMessage to missingResponseMessages
    for $a$ ∈ currentMessage[mainActions]
      if $a$ ∈ N add V[missing response to $a$] + 1
\end{verbatim}

Get cosine similarities for message pairs ∈ C(C):

\begin{verbatim}
model = RobertaSentenceTransformer to count cosine similarities
V[mismatchedSimilarity] = mean(model(mismatchedMessages))
V[matchedSimilarity] = mean(model(matchedMessages))
V[nonNormativeSimilarity] = mean(model(nonNormativeMessages))
\end{verbatim}

FeatureVectorSet = empty set to add all $V$

for $C$ ∈ D:

\begin{verbatim}
V = feature dictionary with feature keys
mismatchedMessages, missingResponseMessages, matchedMessages,
nonNormativeMessages = empty lists × 4
for $M$ ∈ C:
  Select main actions($M$)
  for ∀$P(a)$ ∈ NLIActionProbabilities ∈ M:
    $V[probability(a)] = P(a)$
  Check if $M$ is mismatched response($C$, $M$)
  if at least one of mainActions ∈ N:
  Check if response to $M$ is missing($C$, $M$)
  for $a$ ∈ $A$:
    for $M$ ∈ C if $a$ ∈ M[mainActions]: add $V[a]$ + 1
    $V[mismatchedPairs] = len(mismatchedMessages)$
    $V[missingResponses] = len(missingResponseMessages)$
    $V[matchedPairs] = len(matchedMessages)$
    $V[nonNormativePairs] = len(nonNormativeMessages)$
  for $M ∈ C:
    $V[probability(a)] = (V[probability(a)] ÷ len(M ∈ C))$
    Get cosine similarities for message pairs($C$)
    add $V$ to FeatureVectorSet
\end{verbatim}

ACKNOWLEDGMENTS

The models and calculations presented in this paper utilized computer resources within the Aalto University School of Science “Science-IT” project. High-performance Computing resources were also provided by the Finnish CSC – IT Center for Science LTD., for some of the heavy computing required. We are extremely grateful for our colleagues Antti Ukkonen, Matti Nelimarkka, and Nitin Sawhney for feedback and review of our research, as well as support during the study process. We
Detecting Covert Disruptive Behavior in Online Interaction 20:37

would also like to thank our Editor and reviewers for all the insightful comments that helped make this paper much stronger.

REFERENCES


Detecting Covert Disruptive Behavior in Online Interaction

[53] Claire Hardaker. 2013. “Uh...not to be nitpicky,” but...the past tense of drag is dragged, not drug.: An overview of trolling strategies. *Journal of Language Aggression and Conflict* 1, 1 (2013), 58–86. https://doi.org/10.1075/jlac.1.1.04har
[58] Claire Hardaker. 2013. “Uh...not to be nitpicky,” but...the past tense of drag is dragged, not drug.: An overview of trolling strategies. *Journal of Language Aggression and Conflict* 1, 1 (2013), 58–86. https://doi.org/10.1075/jlac.1.1.04har


Detecting Covert Disruptive Behavior in Online Interaction


Received 21 December 2021; revised 20 August 2023; accepted 29 August 2023