What is the Real Intention behind this Question? Dataset Collection and Intention Classification

Maryam Sadat Mirzaei, Kourosh Meshgi & Satoshi Sekine
RIKEN Center for Advanced Intelligence Project (AIP)
Tokyo, Japan
{maryam.mirzaei, kourosh.meshgi, satoshi.sekine}@riken.jp

Abstract

Asking and answering questions are inseparable parts of human social life. The primary purposes of asking questions are to gain knowledge or request help which has been the subject of question-answering studies. However, questions can also reflect negative intentions and include implicit offenses, such as highlighting one's lack of knowledge or bolstering an alleged superior knowledge, which can lead to conflict in conversations; yet has been scarcely researched. This paper is the first study to introduce a dataset (Question Intention Dataset) that includes questions with positive/neutral and negative intentions and the underlying intention categories within each group. We further conduct a meta-analysis to highlight tacit and apparent intents. We also propose a classification method using Transformers augmented by TF-IDF-based features and report the results of several models for classifying the main intention categories. We aim to highlight the importance of taking intentions into account, especially implicit and negative ones, to gain insight into conflict-evoking questions and better understand human-human communication on the web for NLP applications.

1 Introduction

The essence of conversation is to communicate intentions; however, the uptake of what has been communicated entails more than merely decoding the words in the message (Galinsky et al., 2005). Many layers underlie a communicative message, and as we interact, we try to decode the surface meaning as well as the tacit aspects (Sperber and Wilson, 2015). We further use established codes to interpret the meanings. For instance, a question is known to be a means of asking for information or requesting help. However, as humans, we also apply meta-knowledge to override the established rules; thus, we may interpret a question as deceitful, then consider it as a means to criticize. Our interpretation is a byproduct of multiple factors, including utterance and context, as well as our beliefs, desire, presuppositions, mental states (Wellman, 1992) and cultural background that leads to several possible interpretations of a single message (Creswell, 1996). In this view, Table 1 shows a question with neutral or negative intentions when different perspectives are taken into account.

Question	*Have you been invited to edit this article?*
Positive/ neutral intention	Clarification/Confirmation: Inquiry to disambiguate whether an invitation was involved.
Negative intention	Putdown/Embarrass: An insulting remark to inflict harm as no one has permitted the person to conduct any edits.

Table 1: A question with multiple perceived intentions.

Studies that focus on the intentions of questions in conversations can be categorized as: those focusing on question-answering (Soares and Parreiras, 2020), those related to search engines for accurate information retrieval (Kwiatkowski et al., 2019), and studies that analyze the linguistics aspect of questions (Freed and Ehrlich, 2010). In these studies, questions are generally considered as having the true intention of eliciting a response (Dimitrakis et al., 2020). Yet, the research lacks analyses of the questions' communicative intentions and their interpretations from various perspectives. This is especially true for negative cases that require the application of semantic/inferential knowledge to grasp the underlying intentions. The presence of such questions in conversation is indisputable, making it crucial for systems to learn such knowledge.

In a related direction, studies aim to build systems able to detect attacks in conversations (Coleman et al., 2014), especially explicit attacks and offensive language such as hate speech and political, racial, or religious hatred that manifest in social media (Chetty and Alathur, 2018; Solovev

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13606–13622

and Pröllochs, 2022). However, not all instances of insults are explicit (Jurgens et al., 2019; Poletto et al., 2021) and occasionally, we require ampliative reasoning to interpret the underlying intentions. Whether the speaker has implicit deceitful intentions or the listener ascribes negative intentions to what the speaker says, it may raise a conflict in the conversations, highlighting the importance of analyzing such instances. When questions are used to attack a person, group, or someone's work, the negative intentions are at times implicated rather than explicitly expressed. As a result, seemingly harmless questions can contain concealed attacks or be interpreted as having negative intentions, thereby potentially causing conflicts. Getting to know intention categories helps to understand why a question is negatively perceived. Additionally, it sheds light on people's perspectives and mental states as well as their thresholds in perceiving the message.

In this context, our study aims to analyze the positive and negative perceived intentions behind questions from the reader's perspective. We collected a dataset of questions (*Question Intention Dataset*) on Wikipedia discussion pages to investigate the underlying intention categories. Discussions are a form of communication through sharing or contrasting ideas leading to (dis)agreement. It provides a rich resource of interactions with different intents and goals (Jowett, 2015). On Wikipedia Talk pages, for example, the general goal is to improve a Wiki page, and there are many questions and answers to fulfill this goal. However, it may, at times, be influenced by a personal agenda, such as showing off knowledge by asking questions. Wikipedia discussions can serve as a sample of real-world interaction and a plausible resource for our study. **Within the scope of this dataset**, we probe the following questions:

(**RQ1**) Can different intentions be pursued by asking questions? (**RQ2**) What are the most used intentions when questions have positive/neutral vs. negative purposes? (**RQ3**) Can a question's intention have different interpretations? (**RQ4**) Can we classify the intention categories behind questions?

Our contributions include: *(i)* introducing negative and tacit intention categories for questions and designing a rubric to annotate them *(ii)* gathering perceived intentions from readers' point-of-view, *(iii)* conducting a meta-analysis, and *(iv)* building TF-IDF-based dictionary on intentions and add it to a transformer to benchmark intention classification. Our dataset is available at `https://github.com/marymirzaei/Question-Intention-Dataset`.

2 Background Research

Studies that focus on intentions can be divided into those considering intentions from a linguistics viewpoint and those that consider the psycholinguistic view. The former is in respect to language itself, as in NLP studies on detecting intentions in dialogue systems (Wen et al., 2017), analyzing goals and purposes such as intent to purchase something or to travel (Wang et al., 2015) and those focusing on open-domain question answering (Rajpurkar et al., 2016). The latter focuses on the speaker's meaning, belief, desire, and mental states, hence involving a wider scope.

Intentions within NLP area: Research on intention analysis mainly draws on NLP and deep neural networks to detect the goal of the message and fulfill a task such as realizing a smooth conversation in chatbots (Adamopoulou and Moussiades, 2020; Ouyang et al., 2022), retrieving information in search engines (Zhang et al., 2019), detecting a user's personal need or classifying feedback for marketing purposes and developing recommender systems (Hamroun and Gouider, 2020; Wang et al., 2020; Hao et al., 2022). These studies mainly focus on affirmative or neutral intentions with the aim of associating users' intentions with pre-defined categories. They handle emerging intents via knowledge transfer from existing intents and group the utterances with similar intents (topics) to find the best response or strategy (Xia et al., 2018). Hence they rarely deal with finding implicitly negative intentions, even though it happens in real-world conversations. Research is often concerned with explicit attacks and hate speech, aiming to detect toxic behavior (Sharif and Hoque, 2022), such as hatred toward religious groups (Albadi et al., 2018), racism (Park and Fung, 2017), sexism (Waseem and Hovy, 2016), cyberbullying (Rosa et al., 2019), abusive (Waseem et al., 2017), and offensive language (Davidson et al., 2017; Zampieri et al., 2019).

While explicit attacks have high priority (Gelber and McNamara, 2016; Pérez-Escolar and Noguera-Vivo, 2022), implicit instances of offensive language use are also important since, in many cases, offensive behavior is not explicitly demonstrated (Poletto et al., 2021; Caselli et al., 2020). Thus, more recently studies have explored the use

of implicitly abusive language (Wiegand et al., 2021a,b), latent and indirect hatred on social media (ElSherief et al., 2021), abusive remarks on identity groups (Wiegand et al., 2022), stereotypes (Schmeisser-Nieto et al., 2022), disguised and implicit attacks (Mirzaei et al., 2022) and implicit hate speech detection on machine-generated dataset (Hartvigsen et al., 2022).

Interpretation and perceived meaning in conversation: Intention and perceived meaning have been investigated from different perspectives (Haugh and Jaszczolt, 2012) including associating intentions with the speaker meaning (Grice, 1989), intention as the characteristics of the message (Hall et al., 2001), or as perceived by the addressee.

Other studies consider the notion of perspective-taking and intention perceived as a product of joint communication between the speaker and listener (Clark and Krych, 2004), the speech acts (Searle et al., 1983), and cognitive processes involved in interpreting the meaning and action (Bara, 2010). Meaning is not always perceived by the listener as intended by the speaker (Clark and Krych, 2004; Rosa et al., 2019). Considering both speaker's and addressee's perspectives is optimal for accurate interpretation (Mirzaei et al., 2018), yet it is not always feasible. Thus, studies collect annotations from the readers but provide clear guidelines for higher annotation agreement (Poletto et al., 2017), yet ascertain that a certain level of disagreement should be allowed in annotation (Pavlick and Kwiatkowski, 2019).

Intentions behind questions: Before answering a question, the meaning and intention of it need to be decoded. This is a necessary step for many NLP applications that deal with questions (Zhang et al., 2019; Adiwardana et al., 2020; Soares and Parreiras, 2020). Most research on detecting question intention centers on finding the mapping between the user's query and the knowledge base to provide a user-satisfying response (Bhutani et al., 2019). The candidate answer is selected based on sentences ranked by the model score of its suitability (Yang et al., 2015; Hao et al., 2022). Thus in these studies, a question is considered a query, and its intention is associated with the user's purpose within a specific or open domain (Lazaridou et al., 2022). Other studies investigated the form and function of questions (Freed, 1994; Koshik, 2003; Tsui, 2013), inferring appropriate questions for a given personal narrative such as advice-seeking (Fu et al., 2019), and the questions' semantic and pragmatic properties, such as rhetorical questions (Caponigro and Sprouse, 2007; Bhattasali et al., 2015; Oraby et al., 2017; Kharaman et al., 2019).

In this research, we investigate the types of negative and implicit intentions behind questions that can be used as a means of attacking another person, as well as the positive/neutral questions that serve the primary purpose of asking to receive an answer.

3 Dataset collection

Our dataset is built on the Conversation Gone Awry dataset (Zhang et al., 2018), which encompasses the conversations on Wikipedia Talk Pages (Chang et al., 2020). A combination of machine learning and crowdsourced filtering was used to gather these conversations that begin with civil comments and either remain civil or end with a personal attack (4188 conversations, >30k comments).

Wikipedia's talk page discussions are similar to public forums where contributors convene to deliberate on issues related to editing a page, including quality evaluation (Zhang et al., 2018). Wikipedia comments are known to contain a small number of antisocial behaviors– around one percent (Wulczyn et al., 2017), but it includes many cases of negative attitudes (Schluger et al., 2022), hence a good resource for our analysis. Such cases may interfere with the original goal of improving articles and are disruptive to those seeking to contribute to improving the article in peace by collecting, sharing, and disseminating knowledge. From this dataset, we extracted 2,084 questions and annotated their underlying intentions.

3.1 Crowd-sourced Annotation

We used Amazon Mechanical Turk (MTurk) to collect our annotations. To find reliable annotators, we adopted a hierarchical strategy: *i)* using worker profiles, we limited our workers to those who completed over 700 tasks on MTruk with over 99% acceptance rate, *ii)* conducting pre-qualification test and filtering those who earn low scores (<80), and *iii)* pilot testing to check the quality of workers' annotation. We also looked for instances of random labeling by intentionally including marker questions (red herrings) and checked for serial selection of the same options to exclude such annotators.

	Categories	Definitions	Examples
Neutral / Positive	Seek/share information, knowledge	general or specific inquiry, attempt to obtain info or knowledge, sharing info or news	How much does it cost? What year did he publish?* Did I tell you about my party next month?
	Seek/offer help, opinion, solution, invitation	intentions to seek help for problems, offer help, solution (no bragging), ask someone's viewpoint on issues, invite someone to do something	Could you help with this article?* May I carry the box for you? What do you think is best to do?* Will you join us for meetings?
	Clarification/ confirmation	find direction in what is confusing, eliminate ambiguity in lack of info, ask for more details, examples clarify by info, summary	Is it ok to proceed with plan B? What do you mean by external?* Could you share more details?* So far 1 page done, am I right?
	True guidance/ create awareness	highlight problems to be addressed in a friendly manner, encourage to find solutions, strive to improve and guide without trace of ego	Do you think we can organize it in a table? How about linking it to help the readers?* Would it be a good idea to add examples?*
Negative	Judgemental/ over-critical	display an overly critical viewpoint, unfair judge, blame, accuse, unfair question of credibility, discriminate, fault-finding harshly	Why can't you come up with a simple solution for this? Will your idea be useful at all? Isn't somebody else better at this?* You made wrong edits, correct?*
	Put Down/ embarrass	inflict pain, undermine, diminish importance, put somebody to shame, belittle, make someone feel/look stupid	Is it so hard for you to understand? Why do you think your opinion matters?* Anything you can do right? What skill you've except talking too much?*
	Manipulate	any instance of high-level manipulation or abuse with disguised intention, play a trick make somebody feel emotionally charged/guilty, ask a question to show off/ one up	You do not have any friend, do you?* You know how much you made him suffer? It needs skills, you wanna rely on me again?*
	Show hostility	attack someone with profane words, threaten/ dictatorship/ authoritarian	What are you? piece of sh**?* Get lost, or you'll regret, understood?*

Table 2: Defined categories for annotation of intentions. Examples with * are from (or paraphrased from) the dataset.

3.2 Annotation protocols

We laid out annotation guidelines, defined annotation categories, and provided examples for each category. We depicted some of our positive/neutral intention categories based on studies of questions (Freed, 1994; Freed and Ehrlich, 2010; Tsui, 2013). We went through the process of selecting and refining negative intention categories by analyzing data, defining categories based on the discovered patterns, followed by annotating questions (by two researchers independently), discussions, revising categories and guidelines, then pilot testing with workers, updating and re-annotating. We did this iterative cycle several times to select the final categories in Table 2 and used it as a guideline.

3.3 Annotation Procedure

We designed our annotation scheme and created a friendly interface for the MTurk website. We replaced URLs or personal information with a reference keyword and marked the target question in red. One whole conversation was presented to the annotators (to include the context) with one question in red color at a time. We also included a disclaimer for the offensive content. Our multiple-step process involved the annotations of *i)* intention polarity and *ii)* the intention category. First, we collected the data on polarity (neutral/ positive vs. negative) intentions. Each question was labeled by 7 annotators, after which the majority votes were calculated in order to identify the low-agreement cases (<5 out of 7) to be annotated by four more annotators. In a few cases, low agreement persisted even after re-annotation. We observed that insufficient background information on a particular topic, the involvement of a third party in the conversation, the need for clarifying the question in subsequent comments, self-reflective inquiries (e.g., *"Am I the stupid one here?"*), and the inherent challenge of discerning positive versus negative intentions were among the most frequent factors contributing to the low agreement among annotators.

Once the intention polarity was decided, we ran the second step of our annotation, i.e., choosing the intention category of the questions. The annotators (7 workers) could choose up to two categories of intention per question, but they also had to specify which one had the highest priority. Here, we only focus on the category selected as a priority. The pay rate was between 0.35-0.45 based on the length of the conversation, adding polarity and intention tasks together per question. Our most active annotators were primarily native English speakers, with two individuals who had English as their second language. They came from diverse backgrounds, including American, Italian, British, Brazilian, and

Intention Polarity/ Categories	
Positive/neutral intentions	**24.96**%
Seek/share (info, knowledge)	4.51%
Seek/offer (opinion, help)	13.73%
Clarification/ confirmation	4.61%
True guidance	2.11%
Negative intentions	**75.04** %
Judgemental/ over-critical	40.83 %
Put Down/ embarrass	26.73%
Manipulate/ Abuse/ show off	5.13%
Show hostility/ dictatorship	2.35%
Uncertain/Low Agreements	17.35%

Table 3: Annotation statistics: intention polarity (positive/neutral vs. negative) divided into subcategories.

French, and held undergraduate or graduate degrees. Additionally, our annotators spanned a wide age range, from 21 to 67 years old.

4 Meta-analysis of the data

Table 3 contains the statistics of our dataset. As the data shows, a majority of our questions are labeled as conveying a negative intention (∼75%), leaving only one-third as having positive or neutral purposes. This finding is important in showing how frequently questions can be perceived negatively. Data also suggests that questions are not always used to gain information but can frequently pursue different intentions (RQ1). The table shows that the most used intention category among negative questions is overcriticism (40.83%) while asking for opinion (13.73%) is the most used category for positive/neutral questions w.r.t our dataset (RQ2).

4.1 Questions with positive/neutral intentions

The primary drive behind positive/neutral questions was to gain/offer insights and information or to verify and confirm certain aspects.

Seek/share information, knowledge: The first category regards seeking, providing information, asking for ***news*** and inquiring about ***general/specific/personal information*** which is considered the main reason for asking ***sincere*** questions. Questions in this group were usually addressed in general rather than to specific editors, such as "*There are many different dates for this; does anyone know the real ones?*". The small number of questions in this class is explainable as discussions are held among editors who are knowledgeable on the topic.

Seek/offer help, opinion, solution: Questions intended to ***ask other's opinions*** such as "*What do you think?*" or those aiming to seek/offer help such as in "*Could you please vote in that talk page?*" shape the main category of positive intentions. Our analysis corroborated other studies (Goody, 1980) and intuitively revealed that in most cases, the questions requesting help/solutions are ***politely formulated***, as in: "*Would it be possible to have the lyrics on Wikisource and then link to them?*".

Clarification/Confirmation: These questions are aimed to receive ***reassurance*** as in, "*Are u sure they were moved?*". They are also used to ***confirm agreement or disagreement*** such as "*Any objections to removing it?*" or to ***disambiguate*** for example, "*What part of her article do you particularly want sourced?*". These are used when seeking information about the immediate conversational context in an attempt to eliminate ambiguity and confirm what was understood is indeed correct (Freed, 1994). These questions are typically formulated clearly w.r.t the vocabulary and grammar and are followed by relevant confirmation answers.

True guidance/Create awareness: The last category of intention is to provide true guidance, ***feedback*** and ***positive/constructive criticism*** as in "*Would it not be easier to have a table of the countries [...]?*". Such questions are mostly recognized as ***suggestions*** rather than negative criticism. They have a friendly tone reflected in the vocabulary used such as "*how about*", "*what if*", and "*shall we*". The small number of cases in this category (2.11%) shows that criticism is more often perceived negatively (40.83% in the Judgemental category). However, it can also be explained by several reasons, such as the nature of Wikipedia discussions, where each editor is responsible for providing accurate information and avoiding the inclusion of edits that do not follow regulations or are not based on strong evidence. Thus, cooperative guidance is not frequently observed. Editors sometimes enforce their opinion and criticize others, attempting to show off or establish/maintain face. Similar to the real-world, criticism on this platform is mostly perceived negatively. It forms an attack on the editor's work or personality rather than a friendly suggestion.

4.2 Questions with negative intentions

Judgemental/overcritical: In the negative group, most of the questions belong to the Judgemental and Over-critical class. Such questions convey judgment, and their underlying tone and attitude often express scorn or accusations, leaving criticized people to feel attacked or blamed.

The main characteristics of this category are ***criticisms and accusations***. The question, "*None of these links is commercial, and none of these links is inappropriate. Did you click on them before you acted inappropriately?*" is one such example. It denotes that the person is *i)* criticizing the addressee for improper action [*"before you acted inappropriately*"] and *ii)* accusing him/her of not checking the content before editing [*"did you click on them*"]. This question does not genuinely seek whether he/she has checked the link, thus holds a negative intent.

A distinctive feature of this group is to ***condemn and blame***. For instance, "*Why remove a less-ambiguous sourced statement and replace it with your personal interpretation?*" intends to blame the person, not asking why the source is replaced.

This class also includes questions that ***discredit someone*** and/or impose a threat through criticism, as in "*I didn't see you writing anything to support your revert on the discussion page. Did you, or did you simply use the undo button*?".

Criticism can hold a ***complaint***: "*Why don't instead of keeping on doing these blind reverts which are getting nowhere you'll look for some serious sourced info?*". It can be politely formulated but perceived negatively: "*Can you please explain why you would delete what is probably the most reliable and pertinent source of information this article could have? [...] I will give you the opportunity to explain before I decide what my next step will be.*".

Some cases do not even follow the grammatical form of a question such as "*So you disregarded all the above established consensus and discussion?*". The declarative form makes the questions similar to "*Clarification/confirmation*" questions. However, the question's perceived interpretation is criticism that is implicated, not asserted (Creswell, 1996).

Putdown/embarrass: The next category includes 26.73% of the questions, which is about an effort to put down or embarrass. Questions in this category show some degree of offensiveness through being insulting or belittling, causing ***humiliation***. These intentions are not necessarily expressed with explicit hostility, similar to sarcastic, rhetorical, and unpalatable questions (Bagga et al., 2021). The main characteristic of this group is an indirect ***insult***, and the question is rather ***rhetorical*** than a real one, such as "*who cares about your idea?*" or "*How can I make it any simpler?, This is beyond stupid.*". The context of the last example clearly shows that the speaker is not genuinely seeking an opinion but indirectly making the addressee ***feel/look ignorant***. The same assumption holds for "*Do you really think, that the word "failure" is neutral?*", by which the speaker is ***embarrassing*** the other party. A similar example is "*Are you some sort of super-editor here or something?*", in which the intent is to ***belittle and diminish the importance*** of the other person, another manifestation of putting down. "*Can't you read your own words?*" is another example of implicitly attacking another person by putting him/her down and ***ridiculing***. However, these questions should be distinguished from sincere ones that may seem similar such as "*Maybe someone who knows more about the game could merge it?*". The true intention here is to ask for help from a knowledgeable person, not diminish current editors' expertise. Context is the key to deciphering intention accurately.

Communicative acts causing offenses include simple criticism, insult, accusation, and mockery (Poggi and D'Errico, 2018), which conform to our data. Such actions will make the addressee feel offended since these are implied as unjust criticism, overly judgemental, and insulting reproach. On the other end of the spectrum is hostile behavior or a personal attack, which forms our next category.

Show hostility: This class includes questions that show a high level of hostility and any form of ***clear insult***, ***profanity***, and ***attack*** on the conversational partner, such as "*Is that personal enough for you, you irritating, infuriating little man?*". This class often reaches a high annotation agreement as hatred/offense is explicit (Wojatzki et al., 2018).

Manipulation: This category is perhaps the most abstract among all negative intentions as it involves a certain level of pragmatics and includes a ***hidden agenda*** expressed in an unscrupulous way. In such cases, it is often not the communicative words that are offensive but the implied intention. The category entails cases where someone ***plays the victim***, ***gaslights***, ***denies wrongdoing***, ***takes control over*** and ***abuses*** another. An example like "*You don't have many friends do you?*" presents a case of exercising harmful influence by the speaker. "*I apologize for getting his name wrong (one letter off, and you have to correct me?)*" is another case that shows the speaker is inducing ***guilt*** and ***disapproval*** of what the other person did (Baumeister, 1998). Similarly, "*I try to help out and you call it condescending?*" is another example of a speaker

playing the victim role and ***guilt-tripping***. Finally, ***show off/ one up*** is another case that represents manipulation and includes the questions that the speaker intentionally asks to reinforce his/her alleged superior knowledge, work, and skills. "*Have you noticed that there hasn't been any significant CONTENT or cited material contributed apart from my work*?". These questions are more about ***preening*** and ***grandstanding*** and are used to make others agree with the questioner's mindset/viewpoint. The entire effort is to be seen and influence opinions, not to ask questions out of curiosity and sincerity.

4.3 Uncertain intention categories

Table 3 also presents low agreement/uncertain annotations (17.35%). We have found that the "*Manipulation*" category has the least average annotation agreement (∼61%). This class involves the most indirect, deceptive tactics to **conceal an intention**; it may even seem benign or friendly, making it hard to spot (Billig and Marinho, 2014).

The sensitivity and tolerance ***threshold*** of the reader/listener plays a role in choosing categories. For instance, a question like "*[...] How old are you and where do you come up with this garbage?? Get some sunshine and a breath of fresh air*" was considered an act of insult and ridicule through sarcasm by some annotators and a case of explicit hostility by others causing low-agreement annotations. Similarly, true criticism can be perceived negatively by sensitive people, and a judgmental question can be interpreted as an act of insult and humiliation. Uncertainty can emerge from varying ***perspectives*** that lead to associating different but possible intentions with a question (See Table 1). These indicate that both perspective and threshold for tolerating offense play roles in perceiving questions and different interpretations (RQ3).

5 Classification of intention categories

To address RQ4, we classify the intention categories based on their polarity. For positive/neutral categories, we integrated the "*True guidance*" category with "*Clarification/confirmation*" cases as these two categories were most often selected together (>80% overlap) when annotators could choose up to two categories. These cases were found to be complementary w.r.t our dataset. We also excluded the "*Show hostility*" class from our classification to only focus on implicit cases.

Pre-processing: Our pre-processing included replacing usernames, email addresses, URLs, hashtags, and special symbols with assigned tokens and handling misspellings with TextBlob (Loria, 2018).

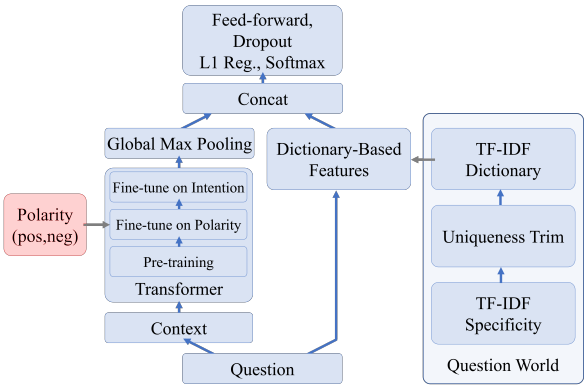

Figure 1: Proposed method with TF-IDF-based dictionary, fine-tuned on polarity to classify intentions.

Context: We added the sentences preceding and succeeding the question, to provide the context to the classifier. If the question started/ended the comment, we used the remaining adjacent sentences. Note that this procedure is done within the comment containing the question. While we recognize that including subsequent comments can enhance accuracy and may even be necessary to understand the question fully, we have restricted the context to only the comment containing the question in order to generate predictions for each individual comment as it is posted.

Classification Method: We classify the intention categories using binary and multi-class classification methods. For binary classification, we target each category of intention individually and fine-tune a Transformer model as a proof-of-concept.

For the main task of multi-class classification, we propose an architecture to fine-tune transformers augmented by a TF-IDF-based dictionary, depicted in Figure 1. The use of dictionary plus transformer has led to improvement in previous studies on relatively similar tasks (Caselli et al., 2021). In the left branch, the question, together with its context, is given to a transformer, which is pre-trained and fine-tuned on the polarity labels of our dataset (positive/neutral vs. negative intentions). The transformer is then trained on our intention categories, and its outputs are fed into global max pooling.

On the right side, we applied TF-IDF to our proposed intention categories to find the most specific words within each class. We trim these vocabularies so that each word appears in only one category (uniqueness trimming) and discard words with

Intention Categories	Binary classification		
	P	*R*	*F1*
Seek/share Information	0.60	0.47	0.53
Seek/offer Opinion	0.71	0.77	0.74
Clarification/Confirmation	0.50	0.61	0.55
Judgemental/over critical	0.76	0.61	0.68
Putdown/embarrass	0.52	0.57	0.54
Manipulation	0.56	0.32	0.41

Table 4: Model performance on the binary classification of intention categories in questions.

higher frequency in spoken/written texts to build our tailored dictionary (See Appendix B). The TF-IDF dictionary was populated with highly specific words from the training portion of the dataset. We did this not only to avoid the risk of target validation leakage but also to enhance the transferability of the model to unseen conversations. Words in each question are lemmatized and matched to dictionary vocabulary to make the feature list. We concatenate the max pooling output with dictionary-based features to be processed by three FC layers and output the label of the question's intention.

Competitive Models: We chose SVM, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019) to build our baseline models for intention category classification and compared their results. These models were selected to yield competitive results in NLP tasks based on previous studies (Nobata et al., 2016; Malmasi and Zampieri, 2017; Tanase et al., 2020). We also used SVM as a strong competitor for being fast and working well with fewer data. Implementation details are provided in Appendix C. We stratified the annotated data and randomly split it into training, val, and test sets (70:10:20). Results are an average of 3 runs.

6 Results and analysis

We conducted two experiments: binary and multi-class classification and reported the results based on precision (P), recall (R), and F1 score.

Binary Classification: In this experiment we labeled the target category as positive while the rest are considered negative. We conducted several experiments with different Transformer models and observed that RoBERTa has the best performance with binary classification, thus the results in Table 4 are based on the RoBERTa model. This table indicates that the best performance belongs to the "*Seek/offer opinion*" class, while the "*Manipulation*" class has the least performance.

Multi-class Classification: The benchmarking results are listed in Table 5. The table shows that all BERT-based models outperform SVM, with the RoBERTa model yielding the best results. The self-attention and the multi-head attention mechanism in Transformers encode each input w.r.t all other inputs, enabling the use of context and considering the relationship between words which is beyond matching sole words. Moreover, the pre-training and transfer learning in the BERT-based models allow for significant performance even with few examples compared with traditional SVM.

The table also shows that the results using the proposed method (RoBERTa+dictionary) overpass the RoBERTa-only model. Using the distinguished words found by TF-IDF analysis assists the model in better classification of intention categories. This improvement is particularly dominant in positive categories, likely because these categories are more explicit and less disguised, and oftentimes politeness and requests are explicitly expressed through specific vocabularies (e.g., "*help*"). On the contrary, the negative groups are inherently more implicit and challenging. However, TF-IDF also proved helpful in unveiling certain negative intentions within the Manipulation category, and it enhanced performance in the Judgemental category by mitigating bias towards this particular category when compared to the BERT-only model. For instance, the words "*allegation*" and "*liar*" were associated with *Manipulation* category, and words like "*ridiculous*", "*meaningless*" and "*nonsense*" were found in *Putdown* category whereas the words such as "*suggestion*", "*help*", and "*thoughts*" were among the vocabulary representing *Seek/offer help or opinion* category.

The data also reveals that the pre-training model on polarity and using the dictionary helps with detecting the intention categories, with the results of this method mostly overpassing baseline and binary classifiers.

As the results show, all classifiers had difficulty in classifying the "*Manipulation*" category, with SVM and BERT facing the most difficulty. One explanation lies in the difficult nature of this category, which also led to low- agreement scores among human annotators ($\sim$61%). The tacitness in the "*Manipulation*" category is the highest among all. Moreover, the boundaries between "*Judgemental*" and "*Putdown*" questions are not always clear as

Methods	Positive/neutral Intentions									Negative Intentions								
	Seek Info			Seek Opinion			Clarify			Judgement			Putdown			Manipulation		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
RoBERTa+TFIDF	0.69	0.58	0.63	0.77	0.77	0.77	0.57	0.63	0.60	0.67	0.71	0.69	0.64	0.59	0.61	0.42	0.38	0.40
RoBERTa	0.69	0.52	0.59	0.74	0.77	0.75	0.52	0.57	0.54	0.72	0.65	0.68	0.57	0.67	0.61	0.38	0.31	0.34
BERT	0.82	0.43	0.56	0.64	0.79	0.70	0.50	0.44	0.47	0.72	0.70	0.71	0.59	0.64	0.61	0.38	0.31	0.34
XLNet	0.75	0.43	0.55	0.66	0.68	0.67	0.41	0.52	0.46	0.69	0.59	0.64	0.51	0.64	0.57	0.43	0.35	0.38
SVM	0.32	0.38	0.35	0.67	0.52	0.59	0.36	0.48	0.41	0.63	0.46	0.53	0.52	0.40	0.45	0.14	0.50	0.21

Table 5: Performance of different models on detecting positive/neutral vs. negative intention categories in questions.

they are tied to people's threshold for tolerating offense, as well as the cultural background or word choices that led to the emergence of uncertain annotations, which in return affected the model.

Error analysis: We found cases where annotators reached a strong consensus, but the model failed to capture the intention. These cases included the complicated structure of the questions such as "*I would only ask that you be more careful with your reverts in the future. Experienced contributors, who make good edits are not usually treated like vandals, okay?*", which starts politely but ends with a warning and forms a long combination of statements and a short form of a question. More complex cases were the fallacy of answering a question with a question where the actual intention is not to ask for information. Another case is the questions that require a significant amount of context to determine the label, in which prior comments played a role in understanding the intention, which has to be addressed in future research. Other cases regarded when a clear indication of a negative intention was missing from the question, as in, "*what is the problem with that?*", which in the context, the speaker is intentionally ignoring the alleged problems and chooses to play dumb or act innocent. Moreover, problems arose when the intensity of the negative intention was not apparent in the vocabulary used to construct the question as in "*Did you take up my suggestion to consult the dictionary?*". Others include using pragmatics that implicates intentions, such as addressing someone with a question "*how old are you?*" and degrading him/her. It requires a higher-level knowledge to interpret the actual intention (Haugh, 2008; Leech, 2016), easy for humans, but hard for the model.

7 Conclusion

This paper proposes the new problem of investigating how humans use questions as a means to attack others and disguise their intentions rather than asking sincere questions to get information. The goal is to incorporate such knowledge into the NLP area. We used the Wikipedia discussions where the editors actively collaborate with the goal of improving Wiki pages. We gathered and annotated questions from discussions to distinguish positive/neutral and negative intentions, plus the intention types. It is only after considering such information that we can learn why a question is perceived negatively. We did a meta-analysis to explore each class's characteristics and the role of thresholds and perspectives in interpreting questions. We also built a TF-IDF dictionary-based transformer and benchmarked several classifiers on intention detection.

Questions are frequently used in conversations, and finding their true intentions is a non-negligible task for AI to understand human communications. The type of intention pursued and how it is perceived by people of different cultures/backgrounds need illumination through the inclusion of diverse perspectives. This future task will enrich research on human reasoning, thereby largely impacting the NLP area on understating human interactions.

Limitations

The intention classification task is not trivial even for humans, especially when the intention is implicit or disguised. The sample size of our study is small, which makes classification more challenging. Currently, we are extending the dataset to include more samples in each category. We aimed to use this data as a proof of concept to shed light on using questions as a means to attack someone or disguise intention. Future directions involve enlarging the dataset and including a variety of social interactions from different sources such as social media (e.g., Twitter), forums (e.g., Reddit), and spoken conversations to investigate other emerging categories based on context, topics, and events.

Moreover, the dataset is imbalanced. Wikipedia

editors should follow strict rules and avoid explicit hostility otherwise get blocked. The nature of Wikipedia discussions is special in the sense that editors need to save face, which refers to the positive social value a person effectively claims (Goffman, 1967) and a professional profile in mainstream interpersonal activities. Implicit and explicit offenses can impact one's face and are closely related to the position and the social fabric of the community, which can lead to righteous indignation by the addressee to save face. On the other hand, since negative questions may disrupt a certain level of interpersonal relations, a speaker will try to minimize this disruption by being polite or implicitly conveying it. Even though this provides us with more implicit samples, which is in line with the focus of this research, the results of this study may not be generalizable to other datasets where the level of offense is higher, and the overall threshold for tolerating offense may be different.

We acknowledge that there may be additional categories that did not emerge in our data. Furthermore, it is important to consider dividing intention categories into more fine-grained criteria. For example, a close analysis of the criticism category reveals a wide spectrum of intensity and threshold of tolerance that plays a role in the perception of criticism. On one end of the spectrum, we have positive and constructive criticism that is more of a guidance and a suggestion, whereas, on the other end, we have an extreme case of criticism accompanied by abusive and hateful language that is more like a personal attack. The following two questions represent both ends of the spectrum, while both can be regarded as criticism: in one, the speaker pursues the goal of improvement by providing a constructive comment "*Can you give a reliable reference for that?*", and in another, the speaker directly attacks the other person "*Why you are being so unhelpful and arrogant?*". This shows that different intention categories inherit the criticism nature to some extent while each involves other characteristics as well.

This highlights the importance of defining more fine-grained categories to distinguish the cases along the spectrum. It should also be noted that even though criticizing questions are associated with the speaker's action and intention, categorizing criticism-implicating questions is explained from the addressee's perspective rather than the speaker's viewpoint, i.e., the addressee should hold the belief that the speaker intended to raise a criticism by asking a question. These beliefs result in a pattern of inferences, leading to correct or incorrect interpretations of the question (Creswell, 1996).

This calls for attention to the difference between perceived intention and the speaker's intended intention. This is another limitation of this study which is the case for many of the NLP studies where the annotations are done by a third person out of context. Having a contrastive analysis between the speaker's intention and the addressee's interpretation can shed light on the similarities and differences, yet not always feasible. Moreover, when dealing with text-based interaction, many aspects of communication, such as the speaker's prosody and tone, are lacking from the textual context; as a result, this gap is filled by the addressee. This is another reason that may lead to inaccurate interpretations of the message.

On a relevant topic, annotators' background, culture, the threshold for tolerating offense, and many more factors can affect their annotation of perceived intention, causing problems in reaching a consensus, but at the same time, different viewpoints need to be included to avoid model bias.

Finally, even though we provided the whole conversation context for annotators to choose the question's intention category, sometimes it is hard to understand the background discussion of the target question. Editors often deliberate on a topic with a follow-up discussion. However, the annotators do not have access to such context (previous discussions, editor's profile) and may not be able to have a clear picture of the questions being asked hence inaccurate interpretations.

Ethics Statement

While the goal of this study is for social good, an intention classifier, if deployed, could also lead to potential negative impacts. For example, a biased intention classifier that picks up spurious features of certain language patterns might be more frequently used by a subgroup of people hence negatively impacting certain users. Our aim is to use this in a collaborative way for willing users to provide hints on the possibility of their questions being perceived with a different intention. In other words, the model can indicate if questions may be perceived by another person as conflict-invoking; hence the user considers rephrasing their questions if they prefer to do so. Our goal is not to restrict free

expressions or take any actions against users, but the opposite, which is to promote friendly discussion and raise awareness of multiple interpretations (only if the users are interested). Yet, this technology, like others, may be misused or might be used in a way that systematically or erroneously silences certain social groups (Gorwa et al., 2020). One solution might be having a threshold that can be moderated by the users since different people have different levels of tolerance to offense, and this also holds for different cultures. Such aspects could be accommodated by collecting viewpoints from different personalities, cultural backgrounds, genders, or generations in order to make a more comprehensive system and avoid model biases. Finally, our model does not provide any indication of where the negative intention lies within the question, which may confuse the users. This calls for extending the system to boost explainability, and transparency, also mentioned in (Chang et al., 2022). In this case, collecting user feedback and annotator reasoning may help identify the problems, and conducting error analysis and training a hybrid model (rule-based guidance on top of machine learning) may improve the performance.

References

Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

Sunyam Bagga, Andrew Piper, and Derek Ruths. 2021. "are you kidding me?": Detecting unpalatable questions on reddit. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2083–2099.

Bruno G Bara. 2010. *Cognitive pragmatics: The mental processes of communication*. MIT press.

Roy F Baumeister. 1998. Inducing guilt. In *Guilt and children*, pages 127–138. Elsevier.

Shohini Bhattasali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic identification of rhetorical questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 743–749.

Nikita Bhutani, Xinyi Zheng, and HV Jagadish. 2019. Learning to answer complex questions over knowledge bases with query composition. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 739–748.

Michael Billig and Cristina Marinho. 2014. Manipulating information and manipulating people: Examples from the 2004 portuguese parliamentary celebration of the april revolution. *Critical Discourse Studies*, 11(2):158–174.

Ivano Caponigro and Jon Sprouse. 2007. Rhetorical questions as questions. In *Proceedings of Sinn und Bedeutung*, volume 11, pages 121–133.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th language resources and evaluation conference*, pages 6193–6202.

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021. DALC: the Dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 54–66, Online. Association for Computational Linguistics.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60.

Jonathan P Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. 2022. Thread with caution: Proactively helping users assess and deescalate tension in their online discussions. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–37.

Naganna Chetty and Sreejith Alathur. 2018. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118.

Herbert H Clark and Meredyth A Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of memory and language*, 50(1):62–81.

Peter T Coleman, Morton Deutsch, and Eric C Marcus. 2014. *The handbook of conflict resolution: Theory and practice*. John Wiley & Sons.

Cassandre Creswell. 1996. Criticizing with a question.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

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

Eleftherios Dimitrakis, Konstantinos Sgontzos, and Yannis Tzitzikas. 2020. A survey on question answering systems over linked data and documents. *Journal of intelligent information systems*, 55(2):233–259.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Alice Freed and Susan Ehrlich. 2010. *Why do you ask?: The function of questions in institutional discourse*. Oxford University Press.

Alice F Freed. 1994. The form and function of questions in informal dyadic conversation. *Journal of pragmatics*, 21(6):621–644.

Liye Fu, Jonathan P Chang, and Cristian Danescu-Niculescu-Mizil. 2019. Asking the right question: Inferring advice-seeking intentions from personal narratives. *arXiv preprint arXiv:1904.01587*.

Adam D Galinsky, Gillian Ku, and Cynthia S Wang. 2005. Perspective-taking and self-other overlap: Fostering social bonds and facilitating social coordination. *Group processes & intergroup relations*, 8(2):109–124.

Katharine Gelber and Luke McNamara. 2016. Evidencing the harms of hate speech. *Social Identities*, 22(3):324–341.

Erving Goffman. 1967. On face-work. *Interaction ritual*, pages 5–45.

Esther N Goody. 1980. Questions and politeness: Strategies in social interaction. *Philosophy and Rhetoric*, 13(3).

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Stuart Hall et al. 2001. Encoding/decoding. *Media and cultural studies: Keyworks*, 2:163–173.

Mohamed Hamroun and Mohamed Salah Gouider. 2020. A survey on intention analysis: successful approaches and open challenges. *Journal of Intelligent Information Systems*, 55(3):423–443.

Tianyong Hao, Xinxin Li, Yulan He, Fu Lee Wang, and Yingying Qu. 2022. Recent progress in leveraging deep learning methods for question answering. *Neural Computing and Applications*, pages 1–19.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Michael Haugh. 2008. Intention in pragmatics.

Michael Haugh and Kasia M Jaszczolt. 2012. Speaker intentions and intentionality. *The Cambridge handbook of pragmatics*, 87:112.

Adam Jowett. 2015. A case for using online discussion forums in critical psychological research. *Qualitative Research in Psychology*, 12(3):287–297.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using nlp to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666.

Mariya Kharaman, Manluolan Xu, Carsten Eulitz, and Bettina Braun. 2019. The processing of prosodic cues to rhetorical question interpretation: Psycholinguistic and neurolinguistics evidence. In *Interspeech 2019*, pages 1218–1222.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

Geoffrey Leech. 2016. *Principles of pragmatics*. Routledge.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Steven Loria. 2018. textblob documentation. *Release 0.15*, 2:269.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472.

Maryam Sadat Mirzaei, Kourosh Meshgi, and Satoshi Sekine. 2022. Is this question real? dataset collection on perceived intentions and implicit attack detection. In *Proceedings of the ACM Web Conference 2022*, pages 2850–2859.

Maryam Sadat Mirzaei, Qiang Zhang, Stef van der Struijk, and Toyoaki Nishida. 2018. Language learning through conversation envisioning in virtual reality: a sociocultural approach. In *Future-Proof CALL: Language Learning as Exploration and Encounters-EUROCALL Conference*, pages 207–213.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical questions and sarcasm in social media dialog. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 310–319.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45.

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

Marta Pérez-Escolar and José Manuel Noguera-Vivo. 2022. *Hate speech and polarization in participatory society*. Taylor & Francis.

Isabella Poggi and Francesca D'Errico. 2018. Feeling offended: a blow to our image and our social relationships. *Frontiers in Psychology*, 8:2221.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, volume 2006, pages 1–6. CEUR-WS.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Charlotte Schluger, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive moderation of online discussions: Existing practices and the potential for algorithmic support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–27.

Wolfgang Schmeisser-Nieto, Montserrat Nofre, and Mariona Taulé. 2022. Criteria for the annotation of implicit stereotypes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 753–762.

John R Searle, S Willis, et al. 1983. *Intentionality: An essay in the philosophy of mind*. Cambridge university press.

Omar Sharif and Mohammed Moshiul Hoque. 2022. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*, 490:462–481.

Marco Antonio Calijorne Soares and Fernando Silva Parreiras. 2020. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, 32(6):635–646.

Kirill Solovev and Nicolas Pröllochs. 2022. Hate speech in the political discourse on social media: disparities across parties, gender, and ethnicity. In *Proceedings of the ACM Web Conference 2022*, pages 3656–3661.

Dan Sperber and Deirdre Wilson. 2015. Beyond speaker's meaning. *Croatian Journal of Philosophy*, 15(2 (44)):117–149.

Mircea-Adrian Tanase, Dumitru-Clementin Cercel, and Costin Chiru. 2020. Upb at semeval-2020 task 12: Multilingual offensive language detection on social media by fine-tuning a variety of bert-based models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2222–2231.

Amy Tsui. 2013. A functional description of questions. In *Advances in spoken discourse analysis*, pages 89–110. Routledge.

Chenyang Wang, Weizhi Ma, Min Zhang, Chong Chen, Yiqun Liu, and Shaoping Ma. 2020. Toward dynamic user intention: Temporal evolutionary effects of item relations in sequential recommendation. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–33.

Jinpeng Wang, Gao Cong, Xin Wayne Zhao, and Xiaoming Li. 2015. Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Henry M Wellman. 1992. *The child's theory of mind.* The MIT Press.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In *International Conference on Machine Learning*, pages 3732–3741. PMLR.

Michael Wiegand, Elisabeth Eder, and Josef Ruppenhofer. 2022. Identifying implicitly abusive remarks about identity groups using a linguistically informed approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5600–5612.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021a. Implicitly abusive comparisons–a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021b. Implicitly abusive language–what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587.

Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, page 110–120.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and S Yu Philip. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL-HLT*, pages 1415–1420.

Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N Bennett, Nick Craswell, and Saurabh Tiwary. 2019. Generic intent representation in web search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361.

A Intention Polarity definitions for annotation task

We instructed the annotators to assign labels to the data by considering the question's context and determining the possible *real* or even *hidden* intention behind the question. They were asked to choose whether the question is perceived to have a *positive (including neutral)* or *negative* intention based on the following definition (Mirzaei et al., 2022).

- Questions are considered to have *positive or neutral intentions* if the purpose or plan of asking is perceived as innocuous, i.e., not harmful at all. These questions are considered sincere with good or neutral intention, such as showing innocent curiosity to elicit information, making a sincere request, or helping to clarify the situation, rather than aimed to hurt someone's feelings.

- Questions with *negative intentions*, do not belong to the above category as they imply negative motives and have an ill-natured inclination to stress fault or strongly criticize the other person (e.g., disqualifying, humiliating, and complaining). These questions are recognized with (obvious or disguised) spiteful purposes, thus raising objections, making the other party feel defensive, and are interpreted as being hurtful.

B TF-IDF based dictionary

To build out TF-IDF induced dictionary, we took the following steps:

- All words in the corpus are sorted ascendingly based on TF-IDF.
- The high-rank words were discarded if they appeared in more than one category.
- If a word is frequent or appears in glosses (e.g., proper names), it is discarded. We also used COCA/BNC corpus and discarded the words with ranks over 500.
- The top-ranked words remaining in the list of each category are included in the dictionary.

The results of the Transformer could be improved using a dictionary since the lexicon in the dictionary gathered by TF-IDF could emphasize the word/phrase in contrast to the attention mechanism in which the transformer set the weights based on the pre-training. We directly used the feature representation of RoBERTa as the word embedding feature of our task. At the same time, the TF-IDF ranked dictionary was fed to our model to improve predicting performance.

C Implementation Detail

For binary classification of each intention category, for the transformer classifiers, we used a dropout layer (with a rate of 0.5), followed by a fully connected layer and a Sigmoid output layer. For multiclass intention classification, for each classifier in the Transformer group, we used English pretraining, fine-tuned it on polarity, and trained the model on intention categories. We used two fully-connected layers with 128, 32, and ReLU activations with a Dropout of 0.5 and L2 regularization of 1e-03, followed by an FC with Softmax activation. We set the classes' weights with a grid search. For both experiments, the learning rate was set to 3e-5, and the batch size was 16. Other settings conform with HuggingFace implementation. The dictionary included 96 vocabularies after the uniqueness trimming procedure.

For the SVM classifier, the pattern of words and the frequency of their occurrence were measured by TF-IDF and the bigram and trigram. We optimized the hyperparameters, using a grid search to maximize the performance of this competitor. We adopted linear SVM to classify the intention categories. We conduct experiments using a P100 GPU.

D Question Selection

The Conversation Gone Awry dataset, which we used to extract our questions, includes 2094 conversations that start and remain civil and 2094 conversations that start civil but end with a personal attack. We extracted the questions equally from civil and uncivil conversations. We assumed the questions asked within civil conversations should be positive/neutral. However, around 21% of those questions had negative intentions. Within conversations that start civil but end with a personal attack, questions were picked from both civil comments and from the last comment that included attacks. Even though we assumed civil comments before an attack should be positive/neutral, around 30% of the questions in that group were also negative. The rest belonged to comments, including personal attacks (24%). One explanation may be that Wikipedia editors are experts and may not necessarily ask questions to get more information but to discuss and oftentimes criticize someone's edit.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section: Limitations

☑ A2. Did you discuss any potential risks of your work?
Section: Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract + Section: 1 Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3: Dataset collection

☑ B1. Did you cite the creators of artifacts you used?
Section 3: Dataset collection

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3: Dataset collection

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 3: Dataset collection

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section: Dataset collection

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section: 4 Meta-analysis of the data; Section 5

C ☑ Did you run computational experiments?

Section: 6 Results and analysis

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix C

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix C

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section: 6 Results and analysis

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 5 and 6

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3: Dataset collection

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3: Dataset collection

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3: Dataset collection

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3: Dataset collection