“Tell me who you are and I tell you how you argue”: Predicting Stances and Arguments for Stakeholder Groups

Philipp Heinisch
Bielefeld University
pheinisch@techfak.uni-bielefeld.de

Lorik Dumani
Trier University
dumani@uni-trier.de

Philipp Cimiano
Bielefeld University
cimiano@techfak.uni-bielefeld.de

Ralf Schenkel
Trier University
schenkel@uni-trier.de

Abstract

Argument mining has focused so far mainly on the identification, extraction, and formalization of arguments. An important yet unaddressed task consists in the prediction of the argumentative behavior of stakeholders in a debate. Predicting the argumentative behavior in advance can support foreseeing issues in public policy making or help recognize potential disagreements early on and help to resolve them. In this paper, we consider the novel task of predicting the argumentative behavior of individual stakeholders. We present ArGenSt, a framework that relies on a recommender-based architecture to predict the stance and the argumentative main point on a specific controversial topic for a given stakeholder, which is described in terms of a profile including properties related to demographic attributes, religious and political orientation, socio-economic background, etc. We evaluate our approach on the well-known debate.org dataset in terms of accuracy for predicting stance as well as in terms of similarity of the generated arguments to the ground truth arguments using BERTScore. As part of a case study, we show how juries of members representing different stakeholder groups and perspectives can be assembled to simulate the public opinion on a given topic.

1 Introduction

Debates on societally controversial topics typically involve different camps or stakeholder groups that have opposing views, interests, and goals. Take the example of the debate on “abortion” (Ginsburg, 1998), which is typically divided into a more “liberal” pro-choice camp and a more “conservative” pro-life camp. The so-called stance (PRO/CON) as well as the argumentative behaviour of stakeholders can often be predicted by knowing their political inclination (left/right), membership to a certain religious group, socio-economic backgrounds, etc. Argumentative behavior is thus to some extent predictable given some knowledge about a group of stakeholders (Alshomary et al., 2022). This leads us to consider the new task of predicting stance and argumentative content given a certain controversial topic and a description of a particular stakeholder in terms of personal attributes. Towards this goal, we present ArGenSt, a framework for the prediction of stances and generation of arguments for stakeholder groups. Our framework is inspired by the approach of Gordon et al. (2022), who focus on hate speech detection and model the opinion towards a controversial topic via a jury containing a number of ambassadors for each group. Instead of predicting a general opinion for an entire group, we predict the stance and argumentative behavior of individual stakeholders, resulting in a distribution for groups assembled out of single stakeholders. Understanding the key positions of stakeholders well in advance would allow us to recognize issues, conflicts, or even general opposition to public policies early on, thus helping to foresee, de-escalate, or prevent a conflict.

More precisely, given a controversial topic t and a person p with a set of personal properties such as gender, income, or religion, we aim to predict p’s stance on t and generate an argument (henceforth called: major claim) justifying the stance of p. We present in particular a fine-tuned architecture combining neural recommender systems with large language models (LLMs) based on Gordon et al. (2022) as well as a prompt-based method utilizing pre-trained LLMs, GPT4 in particular.

We evaluate our approach on data from the debate portal debate.org (Durmus and Cardie, 2018, 2019) as described by Plenz et al. (2024). The dataset comprises of threaded discussions on controversial topics in addition to profiles of the different users of the portal including information about their political party, religion, education level, etc.

We evaluate the predicted stance and argument against gold standard data extracted from debate.org in terms of accuracy and F_1 measure (stance)

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1968–1982
June 16-21, 2024 ©2024 Association for Computational Linguistics

and BERTScore (major claim).

In this paper, we thus make the following contributions:

(1) We introduce ArGenSt, a framework for predicting the stance and generating a major claim for a single stakeholder given a topic.

(2) We present two instantiations of our framework, one relying on a fine-tuned architecture combining neural recommender systems with LLMs and another one relying on GPT4, an LLM that is prompted to predict a stance and major claim.

(3) We conduct a comprehensive automatic evaluation in addition to a manual study showing that, while the prompting approach outperforms the fine-tuned approach in predicting stance, the fine-tuned approach performs significantly better in generating major claims when measured with BERTScore. However, the manual study revealed that, while generating arguments following a simple surface-matching pattern, the fine-tuned approach often generates major claims that are very general or not aligned with the stance. In contrast, the prompting approach generates arguments that correspond to the stance in more cases.

(4) In a case study, we show how the predictions for single stakeholders can be meaningfully aggregated into juries to simulate the argumentative behavior of groups that capture the distribution of stakeholders as represented in the relevant population.

2 Related Work

Summarization of debates and opinion analysis are prominent tasks in the field of argument mining (Friedman et al., 2021). Chen et al. (2019) and Bar-Haim et al. (2020) introduced the field of key-point analysis as the task of identifying the most important aspects or arguments of a debate. While this task has received prominent attention in the form of shared tasks, e.g. by Friedman et al. (2021), these approaches are purely text-based and can only detect *which* core-points exist (including their frequency), but not *who*, that is, which stakeholder, stands in for which key-point.

With the aim of establishing a relation between argumentative content and the personal standpoints/dispositions of individuals, recent research in the field of argumentation mining has focused on modeling the labeling behavior of single individuals, instead of aggregating them into a majority vote (Plank, 2022; Romberg et al., 2022). Indeed, current approaches attempt to predict the label distribution (Pavlick and Kwiatkowski, 2019; Peterson et al., 2019) or even labels for single individuals (Gordon et al., 2022; Heinisch et al., 2023).

The benefit of modeling the argumentative behavior of single individuals has recently been demonstrated by considering recommender-style architectures that embed single individuals on the basis of the arguments they share (Heinisch et al., 2023). However, Heinisch et al. (2023) did not consider any personal information related to political orientation, religious attitudes etc. in the computation of the embeddings.

Beyond the model proposed by Heinisch et al. (2023), Gordon et al. (2022) also proposed a recommender-inspired model to develop a text classification system that relies on a component embedding the personal characteristics of annotators beyond considering their ID and the text only. Inspired by decision processes involving juries divided into different subgroups of individuals sharing a common characteristic (such as age, gender, ethnicity, political inclination, membership in a religious group, etc.), they determine the opinion of a whole group by taking the (predicted) decisions of their so-called *ambassadors* into account.

As a first attempt to predict the stances of single individuals, Toledo-Ronen et al. (2016) have considered the task of predicting the stances of prominent persons on a given topic. For this purpose, they provide a large-scale resource, the Expert Stance Graph from Wikipedia, obtaining background information about the persons from articles that refer to the topics. As a drawback, the method is limited because it only applies to famous people with a Wikipedia article. The approach can thus not predict stances for arbitrary persons based on information about socio-economic background. A step in this direction has been proposed by Jarrett et al. (2023) who have proposed the notion of a "digital representative" as a surrogate opinion generated by a fine-tuned LLM. Beyond this, Bakker et al. (2022) have proposed to generate statements by fine-tuned LLMs as a way to foster consensus.

Concerning the prediction of stance and personal opinions in the narrower sense, Alshomary et al. (2021) have shown how stance can be predicted on the basis of a topic and mainly socio-economic factors. In contrast, Argyle et al. (2023) have proposed a prompting-based approach to create "silicon samples" that have shown to successfully mirror human attitudes. In contrast to the above-mentioned approaches, the method proposed in this paper can predict the stance of an arbitrary person that is represented via a set of personal and demographic attributes. For this, we propose two approaches: one fine-tuned recommender approach and one prompting approach that we both evaluate on the debate.org dataset, discussing their performance and limitations on a wide range of unseen topics and persons.

3 Dataset

In this section, we describe the dataset used in our experiments that was provided by Plenz et al. (2024) and originates from debate.org. In particular, we describe the personal and demographic attributes we use for our model.

3.1 The debate.org Dataset as Source

The dataset provided by Plenz et al. (2024) is based on the now no longer available debate.org portal. The dataset comprises of controversial debates carried out in threads in which users can position themselves for or against a certain topic that labels the thread. The dataset has been used to feed argument search engines (Wachsmuth et al., 2017) and as a basis of shared tasks on argument retrieval such as Touché, organized at CLEF (Bondarenko et al., 2020, 2021, 2022).

While Durmus and Cardie (2018, 2019) published a debate.org-based dataset comprising of 78,376 debates that included personal profiles, we rely on the more comprehensive dataset provided by Plenz et al. (2024) that comprises of more data including opinions and poll votes in addition to all the necessary data for our experiments, including in particular topics, stances, user profiles, as well as the arguments exchanged in a topical thread.

3.2 Personal Characteristics

In order to characterize each user in terms of personal attributes, political orientation, membership to religious groups etc. we extract relevant properties from their profiles in debate.org in which they have shared this information in a voluntary and public fashion. While for some items users had to select values from drop-down lists such as for their EDUCATION LEVEL (among others with the options "*High School*", "*Bachelor Degree*", "*Post Doctoral*", etc.) or INCOME (among others with the options "*Less than $25,000*", "*More than $150,000*", "*$35,000 - $50,000*", etc.), other fields require to enter free text. All fields in the profile were optional, allowing persons to choose which attribute they wanted to disclose and which information they preferred to keep private. Besides the USER ID, in total there are 41 other properties that persons could select or were automatically computed (e.g. number of debates participated). Yet, some of them might not be meaningful for opinion predictions, such as persons' IDs.

In our work, we limit the 41 properties to a subset of 9 that we deem to be particularly relevant for the prediction of the stance towards a given topic. These properties are AGE, EDUCATION LEVEL, ETHNICITY, GENDER, INCOME, POLITICAL SPECTRUM, RELATIONSHIP, RELIGIOUS, and WORKING PLACE. For these 9 properties, we identified 17 key dimensions that we represented on a continuous interval to represent different nuances. For instance, for the property RELIGIOUS we committed to two dimensions, that is, the degree of religiousness and the form of theism, i.e., the number of assumed gods. For each dimension, we use the interval [-1,1] to capture a person' position within two extremes. The default value is represented by 0. Taking the example from above with the property RELIGIOUS and the derived dimension "form of theism", the extreme -1 implies no god (atheism), the default value 0 implies exactly one god (monotheism), and the extreme +1 implies several gods (polytheism). All properties, dimensions, and their descriptions can be examined in Table 3 of the Appendix. The data was labeled by four student assistants[1] trained on similar tasks, who independently mapped the profiles of users to values in the above interval for the 17 dimensions. To arrive at a ground truth value, we averaged the values across annotators per dimension[2].

4 Methods

To predict the stance and to generate a major claim for a given individual described in terms of personal attributes, we propose two different approaches: i) a fine-tuned LLM-based approach consisting of two models: one for predicting the stance, and another one for generating the textual major claim (Section 4.1), as well as ii) prompting-based approaches relying on a pre-trained LLM

[1] They were paid by the standard German pay scale for student assistants. Two of them were studying computer science (both M.Sc.), one mathematics (M.Sc.), and one linguistics (B.Sc.) at the time of the evaluation.

[2] One student assistant refused to rate the property RELIGIOUS, and so we took the mean of three ratings in this case.

to predict both stance and major claim within one model (Section 4.2)

4.1 Fine-tuned LLM

Following the idea of representing stakeholders as a group (jury) of individual persons having certain properties in common, we rely on a recommender-based architecture following Gordon et al. (2022). Since our goal is to predict the stance of a given person *and* to generate a major claim that can be regarded as an explanation for the stance, we introduce two major model components, the stance classifier and the major claim generator, trained in an end-to-end-fashion.

The input for predicting the stance and major claim of a person[3] thus consists of three parts: (i) the *topic*, (ii) a person identifier that is embedded to capture similarities across persons based on their *friendship network*[4], and (iii) all *personal properties of that person* (see Section 3). This input is processed once by the stance classifier for predicting a binary label and by a generative LLM (without the friendship connection) for generating the major claim. In order to integrate the stance prediction and friendship networks into the generation of the major claim, the hidden state of the stance prediction is fed into the LLM generating the major claim. Figure 1 provides an overview of the architecture. A detailed explanation of the major two components can be found in Appendix B.

4.2 Prompting approach

Our prompting approach exploits existing LLMs that have already acquired some common-sense knowledge and reasoning ability as part of their massive pre-training. Hereby, besides the general task instruction, the topic and personal attributes are provided as part of the prompt. We experiment both with a zero-shot setting and two few-show settings in which examples are provided to the LLM as input. The examples are automatically selected by the topic similarity in combination with the similarity of the requested personal properties towards the person of the example based on their attributes.

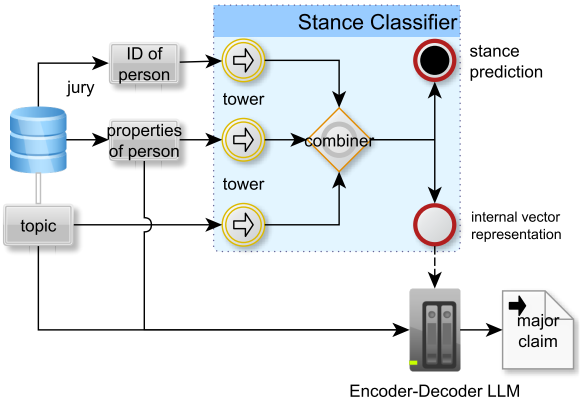


Figure 1: General overview of our architecture. For the inference, the user can input a topic and desired properties which are matched by existing profiles of persons in the database.

5 Experimental setup

We automatically evaluate the stances and major claim produced as output of our models. We rely on a 70-10-20 split of the dataset described in Section 3, corresponding to 5664, 847, and 1471 opinion arguments for the train, dev, and test set, respectively, without overlapping topics and persons. The dataset comprises of 1,253 unique persons. Each sample consists of a topic title, a person with their attributes, stance, and major claim (Plenz et al., 2024).

5.1 Implementation details

Fine-tuned approach In order to fine-tune existing LLMs on the task of predicting the stance and major claim, we rely on the Python `transformers`-library (Wolf et al., 2020). For the stance classification module, we use `all-MiniLM-L12-v2` for embedding the topic, 3 feed-forwarded layers for processing the attributes of the person, a graph neural network for processing the friendship graph, and a DeepCross-Network (Wang et al., 2021) to combine these three components. Details regarding the training are given in Appendix C.1. For the model predicting the major claim, we rely on a T5 base model (`t5-base`) with 8 beams and 4 beam pairs and a nucleus sampling of $p = 0.9$, generating texts of between 5 and 20 tokens in total length.

Prompting approach For our prompting approach, we rely on GPT4-turbo by OpenAI (2023) with 1.76 trillion parameters, queried with the Python library `openai`. We use the following base prompt:

[3] In inference (application case), the person is randomly sampled from the subset of all persons fulfilling desired person-specified properties

[4] Users on `debate.org` were able to be friends with other users on that portal. We mapped these relations to a friendship graph.

```
Your task is, given a person's profile
and a topic under discussion, to predict
the opinion of such a person regarding the
question. You should reply with a stance
(YES or NO) and a short single-sentence
argument explaining that opinion stance
from the viewpoint of that person.
```

We experimented with different settings:

- GPT4(zero-shot): using a standard zero-shot prompting approach providing the instruction, topic, and information about the user but no examples for the task.
- GPT4(3-shot$^{\text{coalition}}$): as GPT4(zero-shot) but additionally providing three examples for similar topics and users.
- GPT4(3-shot$^{\text{coalition}}_{\text{opposition}}$): as GPT4(3-shot$^{\text{coalition}}$) but additionally providing three examples for similar topics but dissimilar users.

In our few-shot settings, we rely on 3 examples as proposed by Yang et al. (2023) for each coalition/opposition. All samples were derived from the train-dev-split, excluding persons without any public properties to ensure that a minimum personal profile is available for each person. The detailed prompt template is given in Appendix C.2.1.

In order to measure the similarity of a sample topic given the query topic, we use SBERT (Reimers and Gurevych, 2019) (considering the model `all-MiniLM-L6-v2`). To compute the similarity between persons, we embed them by using the strategy proposed in the fine-tuned approach (see Section 4.1) and rely on the inverse of the Euclidean distance, i.e., the smaller the distance, the more similar the persons. Using these similarities, we rank all topics as well as all persons and sort them accordingly in ascending (for the coalition) or descending (for the opposition) manner, respectively. We take those samples in the train- and dev-set as examples for our few-shot approaches that minimize the multiplied ranks of their topic and person.

The usernames of all persons provided as part of the input for GPT4 are defamiliarized. We prompt each test sample three times, using a temperature of 0.8 with max. 255 output tokens. We apply no further restrictions or penalties for decoding. In order to yield a final stance we chose the majority vote across the three samples. Concerning the generated major claims, we select the first claim generated.

5.2 Evaluation setup

We report the settings for both our automatic evaluation as well as the manual evaluation carried out with the help of annotators.

Automatic evaluation For our automatic evaluation, we measure the performance of stance prediction compared to the original person' votes using accuracy and F_1. For measuring the quality of the generated major claims, we apply the BERTScore (Zhang et al., 2020), using the 18th layer of `microsoft/deberta-large-mnli`, rescaled with the baseline.

Manual evaluation For our manual evaluation, we first divided the persons into three well-balanced buckets according to the number of their known properties: [1-3], [4-5], and [6-9]. We then randomly draw 5 persons from each bucket. As there were not enough person-topic combinations in all buckets as some questions are less frequently answered than others, we obtained 843 samples to annotate (264 for bucket [1-3], 288 for bucket [4-5], and 291 for bucket [6-9]).[5] In order to evaluate the performance of the three approaches (fine-tuned, GPT4(zero-shot), and GPT4(3-shot$^{\text{coalition}}$)), we hired three student assistants, two of whom had already contributed to the dataset (Section 3)[6]. They annotated the data relying on a custom annotation tool written in Python using the `tkinter` library. The GUI of the tool is shown in Figure 2. For the following explanation of the annotation task, we have labeled the figure with letters from (a) to (i). The annotators should work on several sub-tasks. First of all, they should only look at the debate title (a) and the person properties (b) in order to then assess in (f) whether the predicted person stance (c) is plausible for a person with these properties w.r.t. that debate title.[7] Note that the annotators were not allowed to consider the explanations in d) and e). They should assign a value between 1 and 4. The value 1 was assigned if the

[5] We obtained 843 instead of 1350 samples to annotate (3 buckets · 3 methods · 5 persons· 30 questions).

[6] These were almost the same annotators as for the task in Section 3, i.e., a linguist (B.Sc) and two computer scientists (M.Sc. and B.Sc.).

[7] Since this is of course a very complex annotation task, our main intention was to ensure consistency of annotation. An inconsistent sample would be, e.g., a case where a person belongs to an extreme left-wing party but has extreme right-wing party standpoints.

predicted stance was implausible or unlikely, 2 if it seemed rather implausible, 3 if it seemed rather likely, and 4 if it seemed very reasonable.

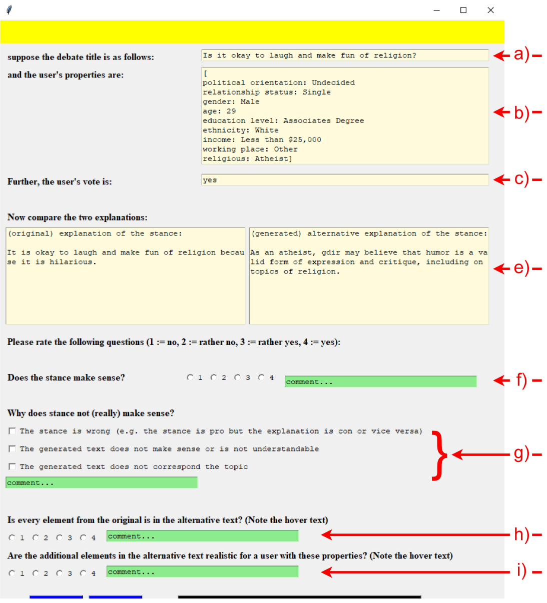
Depending on which value was selected in (f), various subsequent annotation tasks appeared for this test case, i.e., either (g) or (h) and (i). In the case of values 1 and 2 in (f) (i.e. little to no plausible stance), the annotators were asked to indicate (g) why this is the case. To do this, they were asked to look at the generated answer in (e), the stance in (c), and the debate title in (a) and choose between the options (1) "*The stance is wrong*", (2) "*the generated text does not make sense or is not understandable*", and (3) "*the generated text does not correspond to the topic*". (1) could be selected, for example, if the debate title in (a) is "*Is it ok to laugh and make fun of religion*", the stance in (f) is "*yes*", but the explanation in (e) is "*It is not ok to make fun of religion because someone could be hurt*". (2) might be chosen, e.g. if (e) was in itself inconsistent. (3) would be eligible if (e) was completely off-topic. However, in the case of selecting the values 3 or 4 in (f), the answer in (h) should indicate whether the generated text in (e) also contains all of the main text from (d). Furthermore, in (i) they were expected to indicate whether statements in the generated text (e) that are not covered by the original argument are plausible for users with the properties from (b) using the same scale from 1 to 4 as for the stance. The annotators reported that they needed between 30 and 90 seconds per sample, resulting in a total amount of about 10-11 hours.

6 Experiments and Evaluation

6.1 Comparison between approaches

In our first experiment, we predict all stances and major claims for all topic-person combinations in our test set, ensuring that there is no overlap between topics between our data splits as well as no overlap between persons. Therefore, the investigated approaches are expected to generalize to unseen topics and persons. Table 1 reports the scores regarding our automatic evaluation.

Results for stance prediction: For the binary prediction of the stance, we observe mediocre scores for our fine-tuned approach with an accuracy of 0.521 (0.517 F$_1$), only slightly above a random baseline (0.503 accuracy). Note that the prediction probabilities for both CON and PRO are quite balanced. In contrast, the prompting approach has much higher accuracies, that is, GPT4(zero-shot) has an accuracy of 0.682 (0.673 F$_1$). Adding examples to the prompt increases the performance slightly: GPT4(3-shot$^{\text{coalition}}$) yields an accuracy and F$_1$ of 0.698 and 0.695, respectively. However, looking at GPT4(3-shot$^{\text{coalition}}_{\text{opposition}}$), adding not only positive but also examples from the opposition worsens the F$_1$ and accuracy values compared to GPT4(zero-shot) and GPT4(3-shot$^{\text{coalition}}$). One possible explanation might be that GPT4 was confused by the opposite stances made by the opposition.



suppose the debate title is as follows: Is it okay to laugh and make fun of religion?

and the user's properties are:
[
political orientation: Undecided
relationship status: Single
gender: Male
age: 29
education level: Associates Degree
ethnicity: White
income: Less than $25,000
working place: Other
religious: Atheist]

Further, the user's vote is: yes

Now compare the two explanations:

(original) explanation of the stance:

It is okay to laugh and make fun of religion becau
se it is hilarious.

(generated) alternative explanation of the stance:

As an atheist, gdir may believe that humor is a va
lid form of expression and critique, including on
topics of religion.

Please rate the following questions (1 := no, 2 := rather no, 3 := rather yes, 4 := yes):

Does the stance make sense? ○ 1 ○ 2 ○ 3 ○ 4 comment...

Why does stance not (really) make sense?

□ The stance is wrong (e.g. the stance is pro but the explanation is con or vice versa)

□ The generated text does not make sense or is not understandable

□ The generated text does not correspond the topic

comment...

Is every element from the original is in the alternative text? (Note the hover text)

○ 1 ○ 2 ○ 3 ○ 4 comment...

Are the additional elements in the alternative text realistic for a user with these properties? (Note the hover text)

○ 1 ○ 2 ○ 3 ○ 4 comment...

Figure 2: Annotation tool for evaluating the generated arguments.

Results for major claim generation: Regarding the evaluation of the generated major claims, the fine-tuned approach seems to outperform the prompting approach, reaching a BERTScore of 0.644. The fine-tuning approach is able to learn the specific patterns behind the type of argumentative claims that are typically found in the dataset. Indeed, very often the claims represent rephrased versions of the topic (often formulated as a question), stating a declarative sentence (negated in the case of CON stance), often extended by an explaining or refined subclause. The prompting approach without further guidance about the expected surface of the major claim in the case of GPT4(zero-shot) yields only a low BERTScore of 0.157. Adding examples to the prompt increases the score to 0.321 and 0.344 for GPT4(3-shot$^{\text{coalition}}$) and GPT4(3-shot$^{\text{coalition}}_{\text{opposition}}$), respectively, which, nevertheless, is quite far off from the score of the fine-tuned approach.

6.1.1 Quantitative manual study

Soundness of stance prediction Regarding the stance predictions, measuring a fair inter-annotator agreement of $\kappa = 0.38$, the manual ratings correlate with the automatic metrics, confirming that the fine-tuned approach (∅2.59/4) is outperformed by GPT4(zero-shot) (∅2.87/4), which is outperformed by GPT4(3-shot$^{\text{coalition}}$) (∅2.93/4). According to the majority vote of the annotators regarding the stance predictions, 51.6% are rated as (rather) sound for the fine-tuned approach while this ratio increases to 64.4% and 67.6% using GPT4(zero-shot) and GPT4(3-shot$^{\text{coalition}}$), respectively. In the latter case, 40.2% of all stance predictions are rated as very reasonable. However, even GPT4(3-shot$^{\text{coalition}}$) outputs very implausible stances in 25.6% of all cases. In 4.5% of the cases, the annotators claimed that the provided (known) information is too sparse to decide on the soundness of the predicted stance.

Manual evaluation of generated major claims The manual evaluation of major claims is split into two cases, depending on the soundness of the predicted stance. In case of a (rather) plausible stance, we ask for the *coverage*, i.e. the degree to which the generated major claims contain all elements from the original argument (observing a moderate agreement of $\kappa = 0.50$) and ask for the *precision*, i.e. whether the additional elements included in the generated major claims are plausible (observing a moderate agreement of $\kappa = 0.36$). In opposite to the automatic evaluation, here, we observe mediocre ratings for the fine-tuned approach (coverage of ∅1.65/4, precision of ∅1.30/4). The generated major claims by the prompt-based approaches were equally preferred (coverage of ∅2.3/4, precision of ∅2.5/4 for both GPT4(zero-shot) and GPT4(3-shot$^{\text{coalition}}$)). Regarding the coverage, 32.6%, 44.3%, and 40.2% of the generated major claims in case of a (rather) plausible stance yield the highest rating (4) for the fine-tuned approach, GPT4(zero-shot), and GPT4(3-shot$^{\text{coalition}}$), respectively. Looking at the precision, we measure ratios of 30.5%, 59.0%, and 57.3% receiving the highest rating, for the fine-tuned approach, GPT4(zero-shot), and GPT4(3-shot$^{\text{coalition}}$), respectively. However, GPT4(3-shot$^{\text{coalition}}$) generated more major claims which are rated with (3) in coverage and precision than GPT4(zero-shot), indicating a more conservative generation behavior.

Regarding the cases with a wrongly predicted stance, the generated major claims are often not plausible, too, especially for the prompt-based approaches. In the fine-tuned approach, the manual investigation reveals generated major claims where the conveyed stance contradicts the predicted binary stance label. However, the ratio of non-understandable or unrelated major claims is low in all approaches. In the cases of wrong stance predictions, ≈ 4% of all major claims are broken, regardless of the approach. However, especially for the fine-tuned approach, 5.4% of the generated major claims are so vague or general that they are rated as not helpful to back the stance. On the other hand, the prompt-based approaches tend to generate major claims unrelated to the actual topic in rare cases (5%).

In summary, according to the manual study, the fine-tuned approach often fails to generate meaningful and reasonable major claims, but follows simple patterns to maximize the automatic scores. Although high BERTScores are obtained, looking at the generated major claims manually, the prompting approach delivers clearly better results. While the injection of *examples of coalition* seems to improve the stance prediction slightly, the effect is negligible on the generation of major claims.

6.1.2 Case study

In this case study, we compare three approaches (our fine-tuned approach, GPT4(zero-shot), and GPT4(3-shot$^{\text{coalition}}$)) considering two topics. We consider the prominent controversial topic "*Is abortion wrong?*" and, in addition, "*Is Barack Obama doing a good job as president?*". The two major parties in the USA tend to have opposite opinions regarding both topics. Regarding abortion, Republicans tend to represent the Pro-Life position and Democrats are often positioned as Pro-Choice.[8] Regarding the topic "*Is Barack Obama doing a good job as president?*", we expect a clear PRO stance from Democrats and a clear CON stance from Republicans. Therefore, we created two juries with 10 persons each from our dataset by taking only their political orientation as a filter, i.e., randomly selecting 10 persons favoring the Democratic Party (JURY$_{\text{DEMOCRATS}}$) and randomly selecting 10 persons favoring the Republican party (JURY$_{\text{REPUBLICANS}}$).

[8] https://news.gallup.com/poll/246278/abortion-trends-party.aspx

Starting with the analysis of the topic *Is abortion wrong?*, we note that in the dataset, 8 out of 10 Jury$_{\text{Democrats}}$ are CON, that is, claiming that abortion is not wrong, while 8 out of 10 Jury$_{\text{Republicans}}$ are PRO. Jury$_{\text{Democrats}}$ tend to emphasize freedom of choice and denying any moral responsibility, while Jury$_{\text{Republicans}}$ tend to point out the moral implications and the right to life. Interestingly, both groups feature two "outliers" each. Two persons in Jury$_{\text{Democrats}}$ claiming "*Abortion is wrong*" with respect to human dignity are also Christians.

Turning to the analysis of the model predictions, we observe that the fine-tuned approach makes predictions yielding a more balanced perspective, with only 50% of Jury$_{\text{Democrats}}$ being PRO and 40% of Jury$_{\text{Republicans}}$ being PRO. Nevertheless, the outliers in both groups are correctly predicted. In general, the generated major claims by the fine-tuned approach are not overly specific, and read as "*Abortion is [not] wrong*".

The prompt-based approach yields higher intra-group consistency at the extreme of producing an almost unanimous vote. GPT4(zero-shot) predicts 9 times the stance CON for Jury$_{\text{Democrats}}$ (by admitting that one Christian could vote for "yes") and 10 times the stance PRO for Jury$_{\text{Republicans}}$ (full agreement). GPT4(3-shot$^{\text{coalition}}$) yields even a perfectly unanimous stance per group.

Here, we observe that the prompt-based methods, especially when prompted with examples, enforce the stereotypes and minimize the diversity, overfitting to the majority position. Regarding the generation of the major claims, the prompt-based approaches generate verbose claims such as *"NO, because as a Democrat, the person likely supports a woman's right to choose what happens to her own body"* (GPT4(zero-shot)) and *"Abortion is a personal choice and a right that should be respected for individual autonomy and circumstances."* (GPT4(3-shot$^{\text{coalition}}$)) for Jury$_{\text{Democrats}}$. The generated major claims weakly correlate with the references regarding thoughts of Pro-Choice vs. Pro-Life, but are often far more verbose and more unemotional than the references in the dataset, e.g., *"Abortion is not inherently wrong and the ongoing debate surrounding it is absurd"*.

The importance of not overemphasizing a single property of a person (a tendency that is observable for the prompt-based approaches) becomes even more clear by analyzing the question *"Is Barack Obama doing a good job as president?"*. The groups are quite polarized on this question, with the Jury$_{\text{Democrats}}$ being PRO Obama with one single exception and the Jury$_{\text{Republicans}}$ being unanimously CON. The fine-tuned approach fails to capture the PRO stance by Jury$_{\text{Democrats}}$, arguably missing the common sense knowledge that Obama is a Democrat. The prompt-based approaches are far better, capturing the stances correctly in 19 out of 20 cases. However, the prompt-based approaches misclassify the case of one outlying member of Jury$_{\text{Democrats}}$ being a supporter of Stewart Alexander (Socialist Party USA, running against Obama) instead.

6.2 Analysis of the information-sparsity-effect of provided personal-information

A substantial share of people in the dataset provided only sparse personal information, e.g., limited to revealing their gender only in some cases. The assessment based on this information is harder than for a person who provided all the information. Thus, we hypothesize a positive correlation between the number of known properties of a person and the model performance. To this end, we separate the persons into three groups: one low-informative group with ≤ 3 known properties, an average-informative group with $4-5$ known properties, and a high-informative group with ≥ 6 known properties. The results in Table 2 corroborate this positive correlation for our prompt-based approaches. While the lack of personal information in the low-informative group yields an F$_1$ stance score of 0.652 with GPT4(3-shot$^{\text{coalition}}$), the higher density of personal information in the high-informative group raises the score by 0.123 F$_1$-points. Having high-informative persons leads to more close examples included in the prompt in the case of GPT4(3-shot$^{\text{coalition}}$), which leads to a larger performance gain towards GPT4(zero-shot) in this group (+16% in opposite to +14% compared with the low-informative group). However, looking at our fine-tuned approach, we observe a slight negative correlation, mainly due to the number of training samples which is much higher for the low-informative group than for the high-informative group. Hence, often the fine-tuned approach concentrates only on a few selective properties and compensates "wildcards" with general opinion trends or additional information from the friendship network. Note that the scores for the fine-tuned approach differ from those in Table 1 as the fine-tuned approach also includes users who have not specified any properties;

Table 1: Table showing the (macro F_1 and accuracy for stance prediction and F_1 BERTScore for major claim generation.

method	(macro) F_1	accuracy	BERTScore
random baseline	.500	.503	.000
fine-tuned	.517	.521	**.644**
GPT4(zero-shot)	.673	.682	.157
GPT4(3-shot$^{\text{coalition}}$)	**.695**	**.698**	.321
GPT4(3-shot$^{\text{coalition}}_{\text{opposition}}$)	.672	.676	.344

Table 2: Table showing the performance of the stance prediction measured with macro F_1 and accuracy. The predicted values are compared with the mean of the annotator labels.

bucket	method	(macro) F_1	accuracy
[1-9]	fine-tuned	.459	.483
[1-9]	GPT4(zero-shot)	.673	.682
[1-9]	GPT4(3-shot$^{\text{coalition}}$)	**.695**	**.698**
[1-3]	fine-tuned	.469	.474
[1-3]	GPT4(zero-shot)	.635	.652
[1-3]	GPT4(3-shot$^{\text{coalition}}$)	**.652**	**.661**
[4-5]	fine-tuned	.441	.489
[4-5]	GPT4(zero-shot)	.697	.702
[4-5]	GPT4(3-shot$^{\text{coalition}}$)	**.722**	**.723**
[6-9]	fine-tuned	.422	.473
[6-9]	GPT4(zero-shot)	.726	.731
[6-9]	GPT4(3-shot$^{\text{coalition}}$)	**.775**	**.776**

this is in contrast to the prompting approach, which considers only those users who have specified at least one property.

Our manual study additionally reveals that the quality of the major claims correlates positively with the provided information. For example, looking at the majority votes, in the case of a correctly predicted stance in the high-informative group, GPT4(3-shot$^{\text{coalition}}$) generates major claims that cover (almost) all aspects of the original claim in 97% of all cases, and 91% of the major claims are (rather) likely to be supported by such persons (precision). These numbers drop to 69% in terms of coverage for the low-informative group but increase to 96% in terms of precision, basically due to the fact the annotators often refuse to rate “unlikely” when nearly no information about the person is known.

The inter-annotator agreement for all buckets as well as for the individual ones for several annotation sub-tasks measured with Fleiss κ can be seen in Table 4 in the appendix. While it has become clear that the annotation tasks are difficult and very subjective, surprisingly the values do not differ a lot and show very often acceptable agreements.

7 Conclusion and Future Work

We have presented a framework and new task consisting of predicting the stance and major claims of individual stakeholders towards a given controversial topic. The proposed approach relies on a set of personal attributes provided as input to predict the stance and argumentative behaviour of a certain stakeholder (group) that is defined by demographic variables, political inclination, socioeconomic background, membership to a religious group, etc. We have presented and experimentally evaluated two specific approaches toward this end: a fine-tuning-based approach and a prompting-based approach.

Our experiments on the `debate.org` dataset using an automatic evaluation have shown that the prompting-based approach delivers better results than the fine-tuning approach on the prediction of stance (Accuracy of .698 vs .521). In contrast, the fine-tuning approach seems to perform better than the prompting approach on the task of generating the actual arguments when compared to the ground truth arguments in terms of BERTScore. The manual evaluation has relativized this, showing that the fine-tuning approach in many cases yields arguments that are not consistent with the given stance in 35% of cases, while the prompting-based approach has an error rate of only about 20%.

We have nevertheless demonstrated that the approach is interesting from an application perspective as it enables to assemble juries of people that represent the perspectives of different stakeholders, thus allowing us to simulate the opinions and stances of a diverse set of people to understand their perspective on a given topic. This has the potential to support deliberation, allowing for the identification of potential issues, and to foster a better understanding of the rationale of different stakeholder groups. The approach has the potential to unveil the majority opinion of a certain stakeholder group as well as reveiling outliers.

Overall, our paper shows that the newly proposed task is indeed challenging. Future work could consider a retrieval-augmented architecture that extracts additional background and commonsense knowledge and integrates them into the input of models to increase their ability to make more accurate predictions based on knowledge about the topic.

Limitations

It has become clear from our experimental results that the performance of our approach for predicting the stance and major claim of stakeholders is clearly limited. In fact, the approach can be understood as providing a hypothetical synthetically generated argument that members of a stakeholder group characterized by the personal attributes in questions might have come up with. At the same time, it needs to be clear that the generated arguments might significantly misrepresent the perspective of the given group. Even worse, the generated stance and argument might create an impression that the group characterized by the input personal attributes might have a homogeneous perspective on the topic. In general, the approach might thus lead to emphasizing popular or majority views and even foster stigmatization of a certain group. As the approach clearly relies on statistical correlations as reflected in pre-trained models, it might lack sufficient knowledge about a topic as well as about the reality of the groups it is making predictions for. In particular, given that the models lack explicit common sense knowledge about causal relationships and feature a limited ability to reason logically, they might suffer from inconsistencies and conclusions that are not logical. For instance, we have observed in some cases that the same stance is predicted both for the topic "We should do X" as well as for the negated topic "We should not do X".

The approach is further limited in that it might not have the necessary data to make predictions about new topics or debates, while still pretending to be able to accurately predict stances and arguments on topics it has never seen. Finally, the dataset we rely on, based on `debate.org`, is mostly used by English-speaking persons from the USA (and from the UK, sometimes), so that the dataset might have a bias toward the Western culture and might be less representative of other cultures.

Overall, the generated stances and conclusions need thus to be used with caution, understanding their limitations. Our approach should thus in no case be seen as replacing surveys or studies involving real persons.

Ethical aspects

Our proposed approach predicts the stances and arguments of stakeholders on the basis of personal attributes including political inclination, socio-economic background, membership in a religious group, etc. In our experiments, users have been pseudo-anonymized to minimize the probability of identifying a real person or user. The considered attributes, if characterizing a specific individual, represent sensitive information that has to be protected, even if users have made these overtly public.

While our model has been fine-tuned with respect to the stances and arguments provided by individuals as training data, our intended use of the model is not to predict the argumentative behavior of specific individuals but rather of a group characterized by the attributes given. We thus do not intend the model to be used to make predictions for real persons. The personal attributes provided as input of the model should rather be understood as a hypothesized or fictive group the perspective of which one wants to learn something about. The main ethical risk involved in the use of the model is that it might misrepresent the plethora and diversity of opinions within a certain group, reducing it to a single perspective, thus suggesting homogeneity of opinion where there is none. This might lead to stigmatization of a group as well as the enforcement of cliches and stereotypes. The proposed model has thus to be used with extreme caution, being aware of the above-mentioned ethical risks.

Acknowledgements

We thank Mircea-Luchian Pojoni, Martin Hoppe, Leon Kamke, Kilian Bartz, and Björn Metzler for their invaluable help in doing the annotations in time and with great effort. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the projects ACCEPT, ReCAP, and ReCAP-II, Grant Number 455912133, and 375342983 - 2018-2024, respectively, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

References

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 224–233. Association for Computational Linguistics.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*, volume 35, pages 38176–38189. Curran Associates, Inc.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Barrett Academy for the Advancement of Human Values. 2023. The Stages of Psychological Development. `https://www.barrettacademy.com/stages-of-psychological-development`. Online; accessed 03 August 2023.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of touché 2020: Argument retrieval. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022. Overview of touché 2022: Argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, volume 13390 of *Lecture Notes in Computer Science*, pages 311–336. Springer.

Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2021. Overview of touché 2021: Argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings*, volume 12880 of *Lecture Notes in Computer Science*, pages 450–467. Springer.

Care Givers of America: Home Healthcare Services. 2023. 2022 Generation Names Explained. `https://caregiversofamerica.com/2022-generation-names-explained/`. Online; accessed 03 August 2023.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. 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 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1035–1045. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2019. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 602–607. Association for Computational Linguistics.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pages 154–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Faye D. Ginsburg. 1998. *Contested Lives: The Abortion Debate in an American Community, Updated edition*. University of California Press, Berkeley.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives! In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, Singapore. Association for Computational Linguistics.

Daniel Jarrett, Miruna Pislar, Michael Tessler, Michiel Bakker, Raphael Koster, Jan Balaguer, Romuald Elie, Christopher Summerfield, and Andrea Tacchetti. 2023. Language agents as digital representatives in collective decision-making. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

OpenAI. 2023. Gpt-4 technical report.

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

J. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky. 2019. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625, Los Alamitos, CA, USA. IEEE Computer Society.

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

Moritz Plenz, Philipp Heinisch, Anette Frank, and Philipp Cimiano. 2024. PAKT: Perspectivized argumentation knowledge graph and tool for deliberation analysis. In *Proceedings of the 1st International Conference on Recent Advances in Robust Argumentation Machines (RATIO-24)*, Bielefeld, Germany. Springer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julia Romberg, Laura Mark, and Tobias Escher. 2022. A corpus of German citizen contributions in mobility planning: Supporting evaluation through multidimensional classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2874–2883, Marseille, France. European Language Resources Association.

Orith Toledo-Ronen, Roy Bar-Haim, and Noam Slonim. 2016. Expert stance graphs for computational argumentation. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 49–59. Association for Computational Linguistics.

Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. Dcn v2: Improved deep& cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference 2021*, WWW '21, page 1785–1797, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. 2023. Representative demonstration selection for in-context learning with two-stage determinantal point process. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5443–5456. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Further Tables on Dimensions and their Impact

Table 3 shows the persons' properties and the dimensions we extracted to find similar persons. Table 4 shows the inter-annotator agreement for all buckets as well as for the individual ones for several annotation sub-tasks measured with Fleiss κ.

B Further Details of the Major Two Components of the Fine-tuned Approach

B.1 Stance Classifier

The stance classifier follows the structure of a recommender system, having units processing and embedding the parts of the input separately (called towers) and units processing and combining the embedded input parts, followed by a final classification head and internal-vector-head for the major claim generator, respectively.

Table 3: person properties and dimensions with descriptions.

		scale description		
property	**dimension**	**-1**	**0**	**+1**
AGE	generation based on their current age (Care Givers of America: Home Healthcare Services, 2023)	a young generation	unclear	very old generation
	psychological age stages (Barrett Academy for the Advancement of Human Values, 2023)	very young	unclear	for very old
EDUCATION LEVEL	education level	not educated	middle educated	for high education
ETHNICITY	(presumed) longitude of the ethnic origin	west longitude	central longitude	east longitude
	(presumed) latitude of the ethnic origin	south latitude	central latitude	north latitude
GENDER	gender certainty	unknown/hidden gender		known gender
	masculinity-femininity scale	female	undecided/something in between/mixture	male
INCOME	income power	income is very low		income is very high
POLITICAL SPECTRUM	left–right political spectrum	left-wing parties	centre/balanced parties	right-wing parties
	party size/political agenda	small party with a non-mainstream agenda	a medium party (but not a big player in the political landscape)	a major party covering the mainstream
RELATIONSHIP	relationship commitment	no relation	unclear/uncommitted relation	fixed committed relation
RELIGIOUS	(presumed) strength of faith	not religious	undecided/ no observable influence by faith	religious
	form of theism	atheism (0 gods)	monotheism (1 god)	polytheism ($\geq$ 2 gods)
WORKING PLACE	Current employment	currently not working for sure	unknown whether currently working	currently working for sure
	working experience	no working experience/lower-class working place	uncertain job position/ standard	has working experience/ upper-class working place
	job's systemic relevance	no system-relevant job	mediocre system-relevant job	system-relevant job
	governmental-industry job scale	pure governmental job	job in between/ uncertain	job in industry

Table 4: Table showing the inter-annotator agreement measured with Fleiss κ for several tasks: SR := stance rating, SC := stance comment, ANR := arguments that are rated negatively, APR := argument that are rated positively, PREC := arguments that are rated positively with precision, REC := argument that are rated positively by coverage.

bucket	method	SR	SC	ANR	PREC	REC
[1-9]	GPT4(zero-shot)	**.385**	**.11**	.283	**.393**	**.546**
[1-9]	fine-tuned	.377	-.108	**.287**	.308	.462
[1-9]	GPT4(3-shot$^{\text{coalition}}$)	.366	-.111	.272	.361	.497
[1-3]	GPT4 (zero-shot)	.384	-.159	**.262**	**.398**	**.593**
[1-3]	fine-tuned	**.385**	**.153**	.236	.285	.423
[1-3]	GPT4(3-shot$^{\text{coalition}}$)	.367	-.163	.242	.373	.571
[4-5]	GPT4(zero-shot)	.377	-.094	.29	**.404**	.56
[4-5]	fine-tuned	**.413**	**.97**	**.332**	.342	**.561**
[4-5]	GPT4(3-shot$^{\text{coalition}}$)	.369	.101	.282	.355	.47
[6-9]	GPT4(zero-shot)	**.392**	-.099	**.294**	**.367**	**.474**
[6-9]	fine-tuned	.334	-.093	.287	.295	.386
[6-9]	GPT4(3-shot$^{\text{coalition}}$)	.356	**-.089**	.284	.33	.422

Overview of the towers To embed the topic, we use SBERT-embeddings by Reimers and Gurevych (2019). To embed the person identifier within their friendships, we implement a simple graph neural net (each person as a node is randomly initialized).

To embed the person properties, while one approach concatenates all information to a single string processed by SBERT, we alternatively map all the properties into the 17 rational-valued dimension intervals as described in Section 3.2, providing these numerical values as input. The low-dimensional vectors for each profile, resulting from several profile fields as described in Table 3, are concatenated and processed by a shallow feed-forward neural net to receive the final embedding of the profile of the person.

Overview of the combiners In order to combine the three final embeddings (resulting from the topic, person friendships, and profile information), we use a Deep-& Cross-Network as proposed by Wang et al. (2021) which was also used in the Jury-learning system by Gordon et al. (2022)[9].

Classification head As classification head, we use a simple feed-forward neural net gathering the output of all applied combiners, predicting once the binary stance label and an internal vector representation for the argument generator.

[9] For completeness, we also implemented two other (but underperforming) approaches. The most simple one uses static algebraic operations and pooling methods to squeeze all three embeddings into a simple number. Our second implementation concatenates all final embeddings and processes them by a feed-forward neural network.

B.2 Argument Generator

For generating major claims, we rely on encoder-decoder-modeled LLMs, T5 (Raffel et al., 2020) in particular, providing the topic and, by default, the string-concatenated properties of the person as input. When calculating the internal vector representation of the input to the encoder, we introduce the internal vector representation $\mathcal{S}$ produced by the stance classifier by updating the vector of the encoder $\mathcal{E}$ optionally:

$$\tilde{\mathcal{E}} = \mathcal{E} + \lambda\mathcal{S} \tag{1}$$

Hereby, λ regulates how much the stance classifier is allowed to influence the output of the major claim generator. Since encoder-decoder models are not pre-trained with such an encoder output shift, the strategy is to slowly increase λ during training to adapt the language model to its new task, mainly producing differentiated generated texts while having the same (topic) input but different persons processed by the stance classifier.

C Further Experiments and details regarding experimental setup

To reproduce our results, we describe the details of our experimental setup (Appendix C.1) and provide additional insights into the outperformed fine-tuned approach (Appendix C.3). We release our code here: `https://github.com/phhei/ArGenSt`.

C.1 Fine-tuned approach

We train our fine-tuned approach for ≤ 8 epochs with a learning rate of $4e-5$ and $2e-5$ for the stance-classifier module and major claim-generating LLM, respectively, and use early stopping with respect to the stance-F_1 and BERTScore on the development split. Our batch size is $\leq$ 2 topics$\times$ $\leq$ 4 persons (the actual batch size depends on the topic-person-product so that we have training instances for all topic-persons-combinations in the batch). In case of our further experiments in Appendix C.3, while connecting the stance classifier with the generative LLM, we equalize the learning rate to $2e-5$ and set $\lambda = 1$ in Equation 1.

Each training and inference process was executed on one NVIDIA-A40-GPU with 48GB internal RAM. One overall run (training + test predictions) takes one to two hours.

Each configuration for training and testing the fine-tuned approach was run five times. We report the average scores across these five runs.

C.2 Prompting approach

For prompting GPT4, we used the API provided by OpenAI using their library `openai` (OpenAI, 2023), selecting `GPT-4 Turbo`, accessed on December 2023. We paid 44.53$ for all prompt-based experiments (8.94$, 13.70$, and 21.89$ for GPT4(zero-shot), GPT4(3-shot$^{\text{coalition}}$), and GPT4(3-shot$^{\text{coalition}}_{\text{opposition}}$), respectively)

C.2.1 Prompt template for the prompting approach

The prompt contains the task introduction first, then, for each available example in GPT4(3-shot$^{\text{coalition}}$) and GPT4(3-shot$^{\text{coalition}}_{\text{opposition}}$), a request followed by the reference reply, and, finally, the actual request for the instance that should be predicted. The returned reply is automatically divided into stance and major claim by string matching then.

Task introduction Your task is, given a person's profile and a question, to predict the opinion of such a person regarding the question. You should reply with a stance (YES or NO) and a short single-sentence argument explaining that opinion stance from the viewpoint of that person.

Content **Request:** `Person XXX [`

```
political orientation: XXX,
relationship status: XXX,
gender: XXX,
birthday: XXX,
education level: XXX,
ethnicity: XXX,
income: XXX,
working place: XXX,
religion: XXX
]
Question: XXX
```

Reply: `Stance: XXX`

```
Argument: XXX
```

C.3 Further experiments for the fine-tuned approach

Ablation: Disabling friendship network Since our prompt-based approaches such as GPT4(3-shot$^{\text{coalition}}$) are not able to process the friendship network in parallel due to the limited graph input capabilities and context window restrictions, we perform an ablation experiment to test the impact of encoding the friendship network of a user, leading to less information about the user being available for the fine-tuned approach. Leaving out the friendship graph networks leads to a worse stance and major claim-generation[10] ability, achieving a macro-F_1-score of 0.502 (-0.015) on the test split. This shows that our fine-tuned model is able to successfully use the friendship network to grasp the opinions and argumentation styles in community networks to some extent. This is especially important for persons with sparse profile information.

Ablation: impact of feeding the stance classifier vector into the LLM generating the argument Adding an architectural link between the stance classifier and the LLM by feeding the stance classifier vector into the LLM generating the argument as input leads to a decrease in BERTScore from 0.644 to 0.519. We observed that in doing this the LLM becomes more creative at the drawback of generating contradicting statements more often.

[10] The worse major claim-generation is observable in case of an information flow from the stance classifier to the LLM. Here, a missing friendship network leads to a decrease in BERTScore of 0.019. In the case of an unconnected stance-classifier and major claim-generating LLM, the LLM receives no information about friendships anyway