Ask GRANDMA: Generating Stances and Arguments for every Stakeholder Group - for each Topic

Anonymous ACL submission

Abstract

Computational argumentation emerged as an important research area in the field of NLP. W.r.t. argument search engines, arguments are often listed as one PRO and CON block without gatting an overview *who* stand in for *which* position. Yet, it is indisputable that while an argument on a controversial topic such as abortion may be typical to one (e.g., very religious and more male) stakeholder group, it may not be compelling to another (e.g. to a less religious and predominantly female) one.

In this paper, we fill this research gap and present GRANDMA, a framework for **G**ene**RA**ting Stances a**ND** Argu**M**ents for every St**A**keholder Group - for each topic. To that end, we use a new and feature-rich dataset of the large and famous debate portal "debate.org". In our experiments, we present an architecture combining neural recommender systems with large language models as well as GPT4, resulting in the capability of predicting stances and generating arguments for persons of individual stakeholder groups.

1 Introduction

We as society face many controversial topics with opposing camps and stakeholder groups claiming their interests and goals. Being aware of those opposing camps and knowing their demands is crucial for a democracy and can be a step towards de-escalation and an openness to listen to each other and learn from each other. However, it is challenging to gather all conflicting camps or overview the stances and propositions of diverse groups due to a high degree of subjectivity in certain fields. Some phenomena are well-explored in social science, for example, the topic "abortion" (Ginsburg, 1998), where we have the pro-life camp and the pro-choice camp, involving ethical, moral, philosophical, medical, religious, economic, legal, and other claims. While one stereotype would be to assume religious males in the pro-life camp, Ginsburg (1998) already showed that there are also exceptions, and the stance towards abortion is influenced by several factors. To predict the variety of individuals and their opinions having certain common properties, also for more tailored topics lacking in the exploration of social science, we develop GRANDMA – a framework for **G**ene**RA**ting Stances a**ND** Argu**M**ents for every St**A**keholder Group, a tool inspired by Gordon et al. (2022) which models the opinion towards controversial hate speech detection of certain groups as a jury containing a number of ambassadors for each group. Instead of predicting a general opinion for an entire group, each individual ambassador was modeled and predicted. By collecting the opinions of each ambassador, one can result in an opinion distribution for each group. With GRANDMA, we aim to predict the opinions of persons in stakeholder groups on controversial topics in terms of stance and their major claim, refining the stance in a textual form.

To that end of training and evaluation, we utilize a dataset of the comprehensive debate portal "debate.org" (Durmus and Cardie, 2018, 2019). This dataset contains a variety of topics with a collection of opinions towards these, linked to profiles describing properties of the contributed as party, religion, and education (among others).

In our experiments, we present a fine-tuned architecture combining neural recommender systems with large language models based on Gordon et al. (2022) as well as a prompt-based method using GPT4, resulting in the capability of predicting stance distributions and generating major claims for individual stakeholder groups. While the fine-tuned approach has the capability to embed and represent thousands of individuals while fine-tuning, but is lacking in (political) common-sense-knowledge and reasoning capabilities, GPT4 can successfully align topics and ambassador properties and shows a higher precision in predicting the stance properly.

Hence, we make the following contributions:

1. We introduce GRANDMA, a framework for receiving the stance distribution and major claims of persons from stakeholder groups given a topic.

2. We provide and evaluate two approaches of a complete pipeline devoted to predicting stances and producing major claims for individual stakeholder groups based on arbitrary combinations of the mentioned person properties

3. We conduct a comprehensive automatic evaluation and a manual study of these two approaches, intended to stimulate ideas for future research.

2 Related Work

Summarizing debates and opinion analysis is a prominent task in the field of argument mining (Friedman et al., 2021). Starting with a dataset called PERSPECTRUM extracting different angles of views in a debate by Chen et al. (2019), Bar-Haim et al. (2020) introduced the field of key-point analysis, aiming to extract the core high-level arguments, given a (comprehensive) set of arguments, including the task of mapping the given arguments to the conducted set of key-points. While this prominent task of key-point analysis got attention due to a shared task (Friedman et al., 2021), these approaches are purely text-based and can only detect *which* core-points exist (including their quantity), but not *who* stands in for which key-point – in addition to the requirement of proving a set of collected arguments as an input.

In order to move towards the relation of content and personalities, recent research in the field of data collection encourages not overriding disagreeing labels of annotators in subjective annotation tasks by merging them into a majority vote (Plank, 2022; Romberg et al., 2022). Instead, approaches move in the direction of predicting the label distribution (Pavlick and Kwiatkowski, 2019; Peterson et al., 2019) or even labels for individuals (Gordon et al., 2022; Heinisch et al., 2023). In the field of argumentation, only a few datasets were explored in terms of individual opinions. Heinisch et al. (2023) experimented with two datasets related to argument quality and showed the advantages of architectures that embed individual annotators with recommender-inspired systems. However, due to the lack of personal information of the annotators, they did not explore group-based aspects. One prominent subjective task yielding datasets with information about the annotators in order to group them is the detection of toxic content in social media (Kumar et al., 2021). Gordon et al. (2022) used this dataset to train a recommender-inspired model. In addition to Heinisch et al. (2023), they used the given characteristics of the annotator to train the model with a third input component (additional to the text to be classified and the annotator-ID), namely those annotator characteristics. Inspired by decision processes involving juries divided into different subgroups of individuals sharing a common property (such as sex, or ethnicity), they abstract the opinion of a whole group by taking the (predicted) decisions of their ambassadors into account.

However, all these approaches involving individuals and (stakeholder) groups did approach the task of predicting the stances and opinions of certain stakeholder groups towards a given topic so far – without requiring an argument-author retrieval step before. This paper approaches this task and presents two approaches to how this task can be approached.

3 Dataset

In this section, we present our dataset. First, we present the source of our dataset, that is, `debate.org`. Then, we present its special feature, i.e., the self-reported person properties such as RELIGIOUS, INCOME, or GENDER and how we processed them.

As briefly mentioned before, our final dataset is based on the now no longer available debate portal `debate.org`. More precisely, the website was shut down by its operators in June 2022 due to low user traffic (Reddit, 2022; fandom, 2023; YouTube, 2023). Back when it was up, it enabled persons from all over the world to participate in controversial debates. Thereby, persons themselves started threads such as "*Should Abortion be legal?*", where others could then position themselves for or against these topics by using arguments expressed in posts, adding their stances with pro or con labels as well. The invaluable worth of these user-generated texts is underpinned by the fact that argument search engines like Args (Wachsmuth et al., 2017) draw their predefined arguments from `debate.org`, among others. Moreover, the underlying dataset of Args – which uses `debate.org` as the largest source for its arguments– has served as the source for argument

retrieval tasks of CLEF lab's Touché (Bondarenko et al., 2020, 2021, 2022) between the years 2020 and 2022.

Consider that Durmus and Cardie (Durmus and Cardie, 2018, 2019) published the DDO dataset 2018 with 78,376 debates and also provide the users with their properties. However, their dataset contains only a subset of the persons, i.e., ours is larger and covers more data such as opinions and poll votes. Our source database consists of the following 13 tables in total (the number in brackets shows the number of entries for each table): `arguments` (433k), `big_issues` (53k), `comments` (642k), `debates` (81k), `detailed_opinion_arguments` (892k), `detailed_opinion_questions` (32k), `friends` (182k), `opinion_arguments` (106k), `opinion_questions` (12k), `poll_topics` (19k), `poll_votes` (289k), `users` (53k), `votes` (193k).

Recall that our main intention is to predict stances on controversial topics of certain persons from stakeholder groups including the major claim, a justification of their point of view (i.e. their main arguments and motivations). For this purpose, we primarily rely on the data of the `detailed_opinion_arguments` and the persons properties extracted from `users`, which we will discuss in more detail in the following. `detailed_opinion_arguments` consists of fields such as `user`, `link` (to the debate), `argument title`, (detailed argument), and `text`. The special feature of `detailed_opinion_arguments` compared to the `opinion_arguments` is that the former retain the persons' stances. More details about the persons' properties can be found in the appendix.

4 Approaches

To predict the stance and major claim of a Person with input-defined attributes towards a given textual topic, we rely on two distinct approaches: once our fine-tuned approaches with two cores containing LLMs: one for predicting the stance, one for generating the textual major claim (Section 4.1). In Section 4.2, we describe our second approach relying on larger LLMs generating both stance and major claim with one model, queried with a prompting strategy.

4.1 Fine-tuned approach

Following the idea of representing stakeholders as a group (jury) of individuals having specified properties in common, modeled as a recommender system for each of the individuals (Gordon et al., 2022), we shape our architecture as a deep recommender-inspired architecture. Since our goal is to predict stance *and* major claims, i.e., explanation for the stance (argument) towards a given topic, we introduce two major components, the stance classifier and the major claim generator, trained in an end-to-end-fashion.

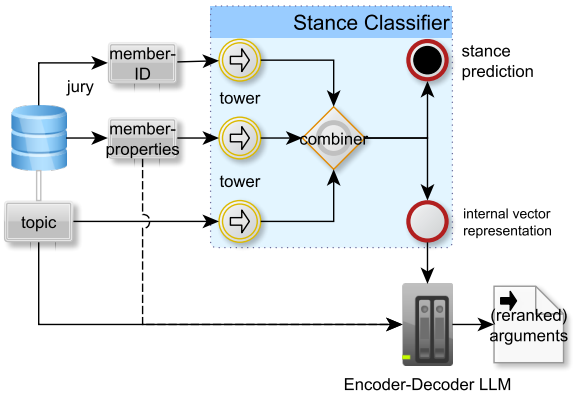

Figure 1: General overview of our architecture. Our database consists of topics and major claims. For the inference, the user can input a topic and desired person properties which are matched by existing profiles in the database.

Hence, our input for receiving the stance and major claim of one member of the jury consists of three parts: (i) the topic, (ii) a person identifier (in inference randomly sampled from the subset of all persons fulfilling the person-specified properties) embedded in their friendship relations, and (iii) all properties of that person (see Section 3). This input is processed by the stance classifier, producing a stance prediction and an internal vector representing the traits of the person towards the topic, which is used by the argument generator to generate a member-tailored argument. Figure 1 depicts this high-level overview. In the following two subsections, we will provide a deeper overview of the two major components.

4.1.1 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.

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 is randomly initialized). To embed the person properties, we implement two approaches. One approach concatenates all information to a single string processed once again by SBERT.

Another approach is to use common-sense background knowledge, mapping all relevant declarations for each property into a low-dimensional continuous space. One example is the mapping of the party information from right (-1) to left (+1) and from minor parties (-1) to dominant parties (+1). The mappings are created with the help of an annotation study (further details in Appendix A.1). The low-dimensional vectors for each profile, resulting from several profile fields, are concatenated and processed by a shallow feed-forward neural net to receive the final embedding of the person profile.

Overview of the combiners In order to combine the three final embeddings (resulting from topic, person friendships, and profile information), we implemented three combiners that can be ensembled for later use. 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 net. The third implementation is a Deep-& Cross-Network by Wang et al. (2021), also used in the Jury-learning system by Gordon et al. (2022).

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.

4.1.2 Argument Generator

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

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

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 pretrained 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 jury persons processed by the stance classifier.

4.2 Prompting approach

For our prompting approach, we rely on larger LLMs with the assumption that those already comprehensively feature common-sense knowledge and reasoning capabilities during pre-training which has to be only prompted now. Besides the general task description, the prompt lists all attributes of the queried person and the topic. In addition to this zero-shot setting, we experiment with few-shot prompts by introducing samples from the train- and dev-split of our data. The samples are automatically selected by the similarity of the query topic and the sample topic in combination with the similarity of the query person and the sample person based on their attributes.

5 Experimental setup

5.1 Dataset

In this section, we examine the implementation and the evaluation setup, carrying out an automatic as well as a manual analysis. We first drew a 70-10-20 sample from our dataset containing 5664, 847, and 1471 opinion arguments for the train, dev, and test set, respectively, without overlapping topics and persons. Each opinion argument consists of a topic title, a person with their attributes, their stance, their argument title and their argument. To condense the major claim (our textual target) using chatGPT by providing the topic, stance and argument title. In a manual study evaluating 99 of those major claims based on the original argument, 94/99 conclusions were labeled as appropriate.

5.2 Implementation

Fine-tuned approach For performing the fine-tuning, 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 GNN for processing the friendship graph and a DeepCross-

Network Wang et al. (2021) to combine these three towers. For the model used for generating the major claim we used `t5-base` using 8 beams and 4 beam pairs with a nucleus sampling with $p = 0.9$.

Prompting approach For our prompting approach, we rely on a LLM, that is, GPT4-turbo by OpenAI (2023) with 1.76 trillion parameters, queried with the Python library `openai`. We introduced the prompt with "*our 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.*" We experimented with three variances: GPT4(zero-shot), GPT4(3-shot^{coalition}), and GPT4(3-shot$^{\text{coalition}}_{\text{opposition}}$).

In our few-shot settings (GPT4(3-shot^{coalition}), and GPT4(3-shot$^{\text{coalition}}_{\text{opposition}}$)), we propose three samples with similar topics and persons and, in case of GPT4(3-shot$^{\text{coalition}}_{\text{opposition}}$), three further samples with similar topics but dissimilar persons. All samples were derived from the train-dev-split, but we exclude persons with less than 4 known properties to provide only informative examples without narrowing the available samples too much. We based on the work of Yang et al. (2023) and also use 3-shot prompting.

For measuring the similarity of a sample topic given the query topic, we use SBERT (Reimers and Gurevych, 2019) (considering the model `all-MiniLM-L6-v2`). For measuring the similarity of persons, we embed them by using the strategy proposed in the fine-tuned approach (see Section 4.1) and measure the Euclidean distance. Using these similarities, we rank all topics and also rank all persons in a decent and ascend manner for the coalition and opposition, 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 sent to GPT4 are defamiliarized. We prompt each test sample three times, using a temperature of 0.8 with max. 255 output tokens. We applied no further restrictions or penalties for decoding. In order to yield a final stance receiving three replies per sample, we select the majority vote for the stance. Furthermore, we select one of the generated major claims as the final one. In our standard experiments, we select the first one.

5.3 Setup of Evaluation

We performed once an automatic and once a manual evaluation by asking four 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-scores. 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

Setup 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 decided to draw a sample of 30 topics with 5 persons at most for 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]).[1] To determine the performance of the three approaches (fine-tuned, GPT4(zero-shot), and GPT4(3-shot^{coalition})), we engaged three student assistants, two of whom had already contributed to the dataset (Section 3).

Annotation Process We designed and provided our own annotation tool written in Python using the `tkinter` library. A visualization of the tool can be viewed 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 points. 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.[2] 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 predicted stance did not seem plausible or unlikely, 2 if it seemed rather implausible,

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

[2] Since this is of course a very complex annotation task, we rather intend to ensure that the predictions are consistent. 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.

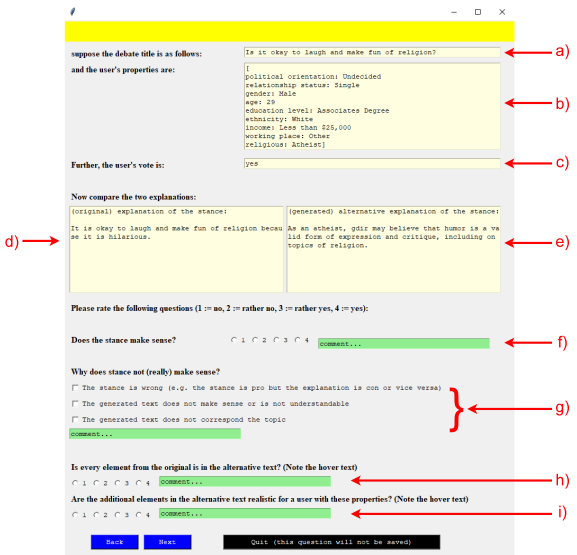

Figure 2: Annotation tool for evaluating the generated arguments.

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., (g) or (h) and (i). To be precise, in the case of values 1 and 2 in (f) (i.e. little to no plausible stance), the annotators should now tick (g) to explain why this is the case. To do this, they should only 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) said something like "*it is ok to make fun of religion but it is not ok to make fun of religion*". (3) would be eligible if (e) in this example said "*solar energy is better than nuclear energy*", i.e., complete off the topic. However, if the value 3 or 4 was selected in (f), the answer in (h) should be whether the generated text in (e) also contains all of the main text from (d). Furthermore, in (i) they should select whether additional elements in the generated text (e) are plausible or rather unlikely for users with the properties from (b) using the same scale from 1 to 4 as for the stance.

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, our approaches have to generalize for unseen topics and novel persons. Table 1 reports the scores regarding our automatic evaluation. W.r.t. the binary prediction of the stance, we observe mediocre scores for our fine-tuned approach with an accuracy of 52.1 % (51.7 % F_1), only slightly above the random baseline and often balanced predicted probabilities for both CON and PRO. Here, we observe that the fine-tuned approach lacks common-sense knowledge and (logical) reasoning capabilities to generalize to unseen (political) topics in combination to infer for unseen persons where the model has to rely on the profile properties and friendship network. This contrasts our prompting approaches – these predicted either CON or PRO in most cases, and are more precise in addition. GPT4(zero-shot) predicts the stance in 68.2 % correctly (F_1-score of 67.3 %). Adding examples to the prompt increases the performance slightly: GPT4(3-shot^{coalition}) yields an accuracy and F_1-score of 69.8% and 69.5%, 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^{coalition}). Only the BERTscore of the major claims increases, arguably due to the additional examples. One possible explanation might be that GPT4 was confused by the opposite stances made by the opposition.

However, looking at the predicted major claim, the fine-tuned approach is outperforming with a BERTscore of 64.4 %. This is due to the successfully learned pattern of the major claim which is often the topic (often expressed as a question), rephrased as a declarative sentence (negated in the case of CON stance), often added with an explaining or refined subclause. The prompting approach without further guidance about the expected surface of the major claim, GPT4(zero-shot), has only a low BERTscore of 15.7%. Adding a few examples to the prompt guides the model towards this common surface pattern, yielding a better score of 32.1%.

Manual study Since the automatic evaluation uses the BERTSCORE to measure the major claims in particular and the prompting approaches performance is lower than the fine-tuned one, we aim to get more insights going beyond surface structures, and thus perform a manual study. In our first annotation task, we ask for the soundness of the predicted stance for each approach. In this task, we measure a fair agreement of $\kappa = 0.38$. For our fine-tuned approach, the majority vote of our annotators confirms the weakness of this approach by rating 41.1 % of all stance predictions as completely unrealistic and only 29.6 % of all stance predictions are confirmed as surely sound, resulting in an average score of 2.59 out of 4. For GPT4(zero-shot), already 63.7 % of the predicted stances are (or tend to be) sound with an average score of 2.87, and GPT4(3-shot^{coalition}) is again slightly better with 65.1 % (tending) sound stances (average of 2.93). In 5 % of the annotated cases, the annotators claimed that the provided (known) information is too sparse to decide the soundness of the predicted stance.

Regarding the generated major claim, although the fine-tuned approach generated good ones at the surface, the manual study confirms the drawback of the weak stance prediction using this approach resulting in many major claims convoy an unrealistic stance (35.4 %), are too general and do not explain the provided stance (2.8 %). Only in rare cases, the generated major claim was considered as broken and not understandable (2 %). The prompting approaches weaken these unwanted effects, generating claims with a non-fitting stance in 22.9 % and 20.7 % of the cases, for GPT4(zero-shot) and GPT4(3-shot^{coalition}), respectively. Regarding the question of the recalled textual elements of the generated major claims towards the original explanation, having a moderate inter-annotator-agreement of $\kappa = 0.50$, the fine-tuned approach has the lowest recall with an average of 1.64 out of 4. Considering the major claims with the correct stance, the majority vote considers 17 % of them without any recall but 33 % with a perfect recall, i.e. every element from the original explanation is covered. The prompting approaches are better with respect to recall without an observable increase by adding examples, having an average score of 2.27 for both GPT4(zero-shot) and GPT4(3-shot^{coalition}). However, GPT4(3-shot^{coalition}) tends to cover only a fraction (> 50 %) of the original elements (44 % of all major claims with a correct stance). The question whether opinion elements which are exclusively expressed in the generated major claims are probable is more subjective, having an inter-annotator-agreement of $\kappa = 0.36$. Again, the fine-tuned approach has mediocre ratings with an average score 1.3, providing 20.5 % major claims with (rather) likely additional elements regarding the majority vote. The prompting approaches are gain next to each other, having an average of 2.44 and 2.49 for GPT4(zero-shot) and GPT4(3-shot^{coalition}). Both prompting approaches yield 58.4 % and 61.3% major claims with (rather) likely additional elements, showing a positive inspiration for the model by introducing examples.

To conclude, the fine-tuned approach failed to generalize for unseen topics on general. Although the BERTscore is outperforming using this approach, looking at the generated major claims, the prompting approaches are more promising. While the injection of coaliting examples improves the stance production slightly, the effect on the generated major claim vanishes.

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

Many persons who should be predicted provided only sparse information, for example, persons where only the gender is known but nothing else. The assessment based on this information is harder than for a person who provided all the information. Hence, 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 are in Table 2, confirming this positive correlation for our prompt-based approaches. While dealing with lots of "wildcards" in the low-informative group and average-informative group leads to an F_1 stance score of 0.6 with GPT4(3-shot^{coalition}), we gain 0.16 F_1-points moving to the high-informative group. Having high-informative persons leads to more precise examples in the case of GPT4(3-shot^{coalition}), which leads to a larger performance gain towards GPT4(zero-shot) in this group (+11 % in opposite to +7 % in the low-informative group). However, looking at our fine-tuned approach, we observe a slight negative correlation, mainly due to the number of training sam-

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

Table 1: Table showing the (macro F_1 and accuracy for stance prediction and F_1 BERTscore for argument generation. The original person votes serve as ground truth for the stances whereas the 'conclusion' is the basis for the generated argument. W.r.t. the fine-tuned approach, the values show the average performance of the five runs.

bucket	method	(macro) F_1	accuracy
[1-3]	fine-tuned	.534	.524
[1-3]	GPT4(zero-shot)	.563	.58
[1-3]	GPT4(3-shot^{coalition})	**.6**	**.614**
[4-5]	fine-tuned	.438	.479
[4-5]	GPT4(zero-shot)	**.602**	**.615**
[4-5]	GPT4(3-shot^{coalition})	.601	.604
[6-9]	fine-tuned	.449	.495
[6-9]	GPT4(zero-shot)	.689	.691
[6-9]	GPT4(3-shot^{coalition})	**.763**	**.763**

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.

ples 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.

Looking at our manual study, the quality of the major claims correlates also positively with the provided information. For example, looking at the majority votes, in the case of a correctly predicted stance, GPT4(3-shot^{coalition}) generates in 97 % of all cases major claims that cover (almost) all aspects of the original claim, and 91 % of the claims are likely to be supported by such persons in every stated aspect. These numbers drop to 69 % in terms of coverage for the low-informative group but increase to 96 % in terms of likelihood, 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. It is obvious that the annotation tasks are difficult. Though, the values do not differ a lot and show very often acceptable agreements.

7 Conclusion and Future Work

In this paper, we present GRANDMA as an approach for predicting the stance and major claims towards a given controversial topic of single persons by providing a list of attributes. Repeating those predictions many times for a group of persons having one or more attributes in common, one can abstract the opinion (spectrum) of a group toward this issue similar to a jury. While our approach of fine-tuning a model based on specific entities failed, i.e. those models failed to generalize to unseen topics that require comprehensive common-sense knowledge and inference abilities in terms of politics, news, and society trends, we present a prompting approach based on the huge LLM GP4 which serves as a first step towards the direction of an automatic opinion analysis without requiring comprehensive surveys or retrieval steps. We show that including examples from related opinion polls can guide our prompted model to a small extend, resulting in a stance accuracy of 70 %. In additional analysis, we show the importance of providing detailed information about the person to be predicted to receive precise and expressive predictions.

With this work, we open the field of automatic opinion analysis. In our future research, we want to combine our fine-tuned approach with comprehensive common-sense knowledge base and retrieval techniques in order to yield a more controllable model with capabilities to infer stances and major claims in a more precise and diverse way.

Limitations

Our approaches do not always predict correctly. This is due to the variety of persons, including have surprising opinions which do not fit into the major opinion landscape of a stakeholder group. Hence, one interpretation of our sometimes low scores is not to pigeonhole individual persons but respect divergent opinions. However, this might not devalue statistical studies about phenomena in society. Another limitation of our approaches are that novel events as the Corona epidemic can lead to worse predictions in case of lacking current training data.

A third limitation of our approaches is that they only look for statistical correlations, but not for reasons or causalities. For example, our fine-tuned

approach predicts sometimes the same stance for a person once for a topic "We should do X" and once for the negated topic "We should do not X". While GPT4(zero-shot) and GPT4(3-shot^{coalition}) reduce these fallacies, such large language models lacks in a meta-level reflecting why they predict a certain opinion and why different persons come to different conclusions under different circumstances. In conclusion, our approaches should not replace social studies and talking to persons but should lower the barrier to get a first overview about controversial topics and their stakeholders.

Code of Ethics

In this paper we investigate the possibility of predicting the stances and opinions of individuals belonging to different stakeholder groups. This may raise skepticism about whether this paper could lead to stigmatization. We take this concern very seriously and have therefore taken care at all times to avoid such opportunities for stigmatization. For example, we do not predict entire stakeholder groups, but only for individuals, which can then help to summarize a spectrum of opinions. Apart from that, GPT4 is also subject to ethical requirements, so that sensitive topics would already be prevented by this. Finally, we have anonymized the names of the users in order to prevent feedback about the real people.

References

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.

fandom. 2023. Debate.org. https://ddo.fandom.com/wiki/Debate.org. Online; accessed 03 August 2023.

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.

Deepak Kumar, Patrick Gage, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *SOUPS*. Usenix.

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.

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).

Reddit. 2022. What's the deal with debate.org shutting down its entire website? https://www.reddit.com/r/OutOfTheLoop/comments/v98rtp/whats_the_deal_with_debateorg_shutting_down_its/. Online; accessed 03 August 2023.

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.

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.

YouTube. 2023. Debate.org Has Officially Shut Down - R.I.P. Debate.org (2007 - 2022). https://www.youtube.com/watch?v=fD6SiZoZSME. Online; accessed 03 August 2023.

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 Appendix

A.1 Person Properties

Persons of debate.org were able to detail various properties in their accounts. While some of them offered drop-down options to select one like 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.), others offered free fields like favorite TV shows or sports teams where persons could enter arbitrary data. However, no property was mandatory, allowing persons to choose which attribute they wanted to disclose and which information they preferred to keep anonymous, which was then declared as "*private*". Besides the USER ID, in total there are 41 other properties that persons could select or were automatically mapped to (e.g. number of debates participated). Yet, some of them might not be meaningful for opinion predictions, such as persons' ids, private e-mail addresses, their real names, or statistics about their behavior on debate.org.

In this work, we therefore limited these properties to nine, which we found to be the most meaningful and useful for our task. Based on this, we extracted 17 dimensions as floating point values because some properties might need more than one dimension to be represented in a satisfactorily expressive manner. For instance, for the property RELIGIOUS we committed to two dimensions, that is, the grade of religiousness and the number of gods. This mapping was necessary since we had to conduct feature engineering and the various specifications of the properties were widely spread in quantity and degree of difference. This information could be lost if mapped to fixed classes. For example, we counted 72 different unique specifications of the property RELIGIOUS. For all dimensions, we used the interval [-1,1], where a value in the middle means (most of the time) that the property is neutral with respect to a class. All properties, dimensions, and their descriptions can be examined in Table 3 in the appendix. As an example, consider the property INCOME, where the floating number -1 means that the income is comparatively very low ("*Less than $25,000*"), while +1 means that it is comparatively very high ("*More than $150,000*").

To this end, we hired four student assistants with great expertise in computational argumentation, who were independently presented with the 17 dimensions in question, and all the grouped drop-down choices from the nine properties so that they could map them to a scale with the two extremes -1 and +1.[3] Note that before annotation, the four annotators were presented with the properties and dimensions so that they could give their feedback on our preliminary design. After discussion, we then modified them to the final dimensions as shown in Table 3 in the appendix. Thus, the effort for the annotators was manageable in most of the cases, since the number of choices for the property values were mostly small. However, for the properties POLITICAL SPECTRUM and RELIGIOUS, there were plenty of options to choose from, 55 and 72, respectively. Here we gave the instruction to ignore underrepresented religions and parties with less than 30 people avowing them if the annotators did not know them, because they would not have a substantial impact anyway. Hence, we assigned these to the default value of 0 in the end.

After completing the annotation, we took the mean values of the mappings and henceforth use these as our ground truth.[4] Be advised, that an inter-annotator agreement makes little sense with this annotation because the annotators could have used any floating-point number between -1 and +1 for their mapping, which would practically always infer a 'bad' but probably invalid agreement. However, manual inspection revealed that at least the tendency was often the same. In particular, due to the high number of annotators, we expect to find a good basis by taking the mean as we strongly emphasize the importance of having a large number of annotators to provide different perspectives.

[3] Note that although we recruited four annotators for this task, their annotations may still be biased due to their cultural and demographic backgrounds as all of them live in Germany.

[4] One student assistant refused to rate the property RELIGIOUS, and so we took the mean of three ratings in this case. In addition, skin color was ignored for the property ETHNICITY to avoid moral concerns and solved this in more elegant way by using longitude and latitude.

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 positon/ 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 recall.

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^{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^{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^{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^{coalition})	.356	**-.089**	.284	.33	.422