2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Understanding Online Attitudes with Pre-Trained Language Models

William Power
Temple University
Philadelphia, USA
tug00038@temple.edu

Zoran Obradovic
Temple University
Philadelphia, USA
zoran.obradovic@temple.edu

***Abstract*—This work investigates how the rich semantic embeddings of pre-trained language models can be used to help understand the general attitudes of an online community. This work describes a novel prediction model that can ingest statements describing an arbitrary context and a piece of content, and output answers to a set of 'attitude questions' describing the relationship between them. Typically, annotating answers to questions like "Does this contain sarcasm?", or "Is this content positive with respect to this context?" requires costly human interaction. In this work, we consider the ability of large language models to answer these questions, while under the constraint of a small dataset using a novel prediction head. We show that this methodology can accurately answer these attitude questions, compare the model to off-the-shelf language model approaches, and describe a method for collecting and annotating attitude question data sets. The novel attitude question answering model achieves a 89% accuracy on the attitude question answering task, outperforming the ablated models (87%) as well as the off the shelf models using BERT-based Sequence Classification (13%), BART-based Natural Language Inference (88%), and RoBERTa-based Question-Answering (87%).**

***Index Terms*—Social Network Analysis, Data-mining, Question Answering, Attitude Modeling**

I. Introduction

How do we understand general attitudes about a concept? This is a broad idea, but there are many instances of such tasks; How can we understand and identify negative and dehumanizing attitudes towards Autistic people [1]? How can we understand the attitudes related to COVID vaccine hesitancy [2]? How can we understand the publics attitudes with respect to an increasing use of drones [3]? Questions like these have typically been answered by the workhorse of population research; the mail survey. However, the observed decline in both quantity and quality of useful mail responses suggests that new economical methods of meeting this task should be considered [4]. Online social networks represent a large body of user-created content from which we can draw observations and answer these questions. However, this requires addressing the tasks of selecting and filtering online populations of interest and mining content from them, and matching these data with models and pipelines appropriate to the quantity and quality of data.

We see a large body of work, across a variety of domains that fits this general structure of mining content and modeling attitude. Aggregate sentiment statistics calculated from geo-fenced tweets can provide insight into COVID vaccine attitudes [5]. Models of political opinion change can be built by combining topic extraction and lexicon-based frequency statistics [6]. Even, product reviews can be automatically parsed to extract aspect-opinion pairs to create targeted aspect sentiment analysis [7].

Where other works frame attitude as a function of frequency statistics or as a value based on sentiment, this work proposes a more holistic definition of attitude based on a set of representative 'attitude questions'. The set of these questions and their answers are taken as the 'definition' of an attitude, which would then provide the main information for downstream analysis. The questions are selected to mimic the language seen in Knowledge, Attitude, Practice surveys [8].

Modeling attitude then becomes a task of classifying pairs of context descriptions and user content by the answers to the attitude questions. Due to the costs (in time and money) for annotation, we sought to design a novel parameter-efficient prediction head that considers the embeddings of each of the three important pieces of natural language information; the context, the content, and the question. We will show that this proposed model is a parsimonious solution to the attitude question answering problem by comparing it to off the shelf, prompt based approaches, as well as simplified ablated versions of the proposed model. In these evaluations, we will show that our parameter efficient model outperforms both the prompt-based approaches, as well as the ablated versions.

This contribution outlines

- A methodology for answering attitude questions.
- A process for collecting and annotating attitude question data sets.

In support of these contributions, we will answer the following research questions.

1) Can off-the-shelf, prompt-based, pre-trained language

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
https://doi.org/10.1145/3625007.3627302

model pipelines accurately answer arbitrary-context attitude questions given a small dataset?
2) Can a model that operates over embeddings of the context, content, and questions strings outperform off the shelf models given a small dataset?
3) How important are each of the three language inputs (context, content, question) to the accuracy of the proposed model?

II. Prior Work

A. Analysis of Reddit Content

This work, and many others in the domain of social network data-mining and analysis, considers the online social network Reddit.com. Prior work has investigated and described this site as a source for topic-specific, user generated content [9]. Working with nuanced natural language content from an online social network like Reddit has been the target of a large body of work.

This includes approaches related to mental health care. By scraping user generated content from suicide-related subreddits, and by using manually labeled training data, answers to clinically developed suicide and self harm diagnostic questionares can be implicitly generated by a sequence to sequence model [10]. The various sub-communities (specifically /r/depression, /r/suicidewatch) of reddit can be used as weak labels to help generate training data for distinguishing posts related to depression or suicidal ideation [11]. Additionally, content from finance-related subreddits and communities have been studied for their impact on major financial markets, using aggregate sentiment measurements to predict changes in share prices for certain stocks [12]. Reddit has been used to identify a body of users related to a population of interest. By looking at the content of mental health subreddits, and inspecting the content posts, a set of users that identify as having a mental health issue, like schizophrenia, can be built [13].

These works consider individual communities, with specific, targeted goals for collecting content to evaluate. In short, they consider specific contexts, with a specific set of questions. In comparison, the contributions outlined in this paper consider much more general contexts, over multiple domains.

B. Question Answering

Question answering is a large topic that addresses a varied set of question kinds and types [14]. Approaches in the field have considered a wide range of domains over which to ask and answer questions. Knowledge bases provide an interesting source of fodder for question answering systems, with a large body of current work using taxonomic or neural-semantic parsing to convert the various methods of instantiating knowledge to answer questions [15]. An example of such a method considers a network-based representation of facts within a knowledge base. A combination of message-passing graph neural networks and a novel subgraph selection process is used to build a model that can predict a chain of reasoning within the knowledge base [16].

Reddit has also been used as a source of training data for longer-form question answering tasks. The sub-community '/r/ELI5' contains posts of users requesting answers to questions, but with the assumption that any responses will be written as if intended for a child of the age of 5. These question posts are mined for highly-scored (by the reddit users) answers, and provide the training data for sequence-to-sequence models [17].

This work generalizes the question answering task to handle arbitrary contexts, where each input example consists of not only a question statement, but a description of a context. In addition, the proposed model addresses a binary question answering task, which, while a simpler task than many approaches, benefits from the smaller size of the data set used in this contribution.

C. Emotion and Attitude

Prior work has also considered the tasks of leveraging machine learning to extract, model, or predict the emotional element of natural language content. Cross-disciplinary efforts have been made to classify the possible psychological models of emotion that might be used as part of machine learning fueled text-based emotion detection methods [18]. Recent approaches have considered GPT, BERT, and LSTM based language models to predict the emotional content of a text, with common data sets like EmotionLines, SemEval, and EmoBank used to evaluate them [19].

The concept of attitude has also been studied with machine learning. Recent approaches have considered the ability to categorize the overall attitude of a region with respect to COVID-19 vaccines by considering the natural language content of mined tweets. Human annotators label vaccine related tweets with an attitude valance, which provides the training data for a BERT-based model capable of classifying a tweet as pro-vaccine [5].

Our study is enhancing published methodology for estimation of emotion and attitude by using a novel operationalization of the concept of attitude, where instead of being expressed or calculated as a function of a lexicon or single-label-annotation, it is expressed as a set of answers to a set of binary questions over the relationship between a piece of content and a context.

III. Data

In this section we will describe the collection of raw data from Reddit.com[1], as well as its annotation and processing for use in the subsequent evaluations.

A. Data Mining

To fully define the applied notion of attitude in this project, we consider the scope of possible content. By choosing a set of target contexts, we are choosing a set of natural language statements that represent some concept of interest. These are used to inform selection of sub-communities from which we mine content, as well as directly used as query strings when

[1]Using the PRAW (Python Reddit API Wrapper)

TABLE I
CONTEXTS CONSIDERED, AND THEIR QUERY STRINGS.

Context
”The military and its use of drones”
”The military and its use of tech”
”Technology and invasion of privacy”
”Technology and its role in fake news”
”The presidency of Joe Biden”
”The effectiveness of vaccines.”

searching for content. Ultimately, the selected contexts are used by annotators to label our collected data, as the attitude questions they answer consider the relationship between the context and content.

The methodology proposed in our study is evaluated on a set of contexts intersecting the domains of technology, the military, and politics. The list of considered contexts used in conducted experiments are listed in Table I. These contexts were selected to reflect current issues of concern, which would be widely discussed, and have a variety of possible opinions. These are USA-centric contexts, which was done to match the predominately US-based user population of Reddit.com.

To find related content, the contexts are first used to determine a set of related ‘subreddits’. These are the sub-communities that exist on the popular online social network, Reddit. This social network is composed of users self-organized into communities called subreddits. These sub communities are sub pages of the site, built around a common theme or topic, that allow users to submit posts for viewing by other users. These submissions also allow for users to comment under the post, contributing to a forest of nested comments. It is the comments written under these submissions that are mined for inclusion in the dataset.

In our study a list of related subreddits is manually constructed by considering the Reddit curated list of most popular existing subreddits (by number of users subscribed to the sub community), and choosing the highest populated subreddits most closely related to each context. Additionally, a set of ‘unrelated’ subreddits was also selected from the list of most-popular communities. These were chosen to provide a source of ‘negative’ responses for the data set. The contexts chosen roughly correspond to topics in politics, technology, and the military. We list the chosen subreddits by these categories in Table II.

With the set of contexts and sets of sub-communities in hand, actual content can be mined. This was done by searching within each of the listed subreddits, using the provided general query API, for each of the context strings. These search results were then ranked by amount of user up-votes. The top rated posts were then mined for content by collecting all of the top-level comments made by users on the post. The comments were tagged with the context string used to search for the post, and the author and subreddit IDs.

TABLE II
SUBREDDITS USED TO MINE CONTENT FROM, GROUPED BY CONTEXT.

Topic	Subreddits
Political	”r/conservative”, ”r/liberal”, ”r/politics”
Technology	”r/software”, ”r/computers”, ”r/tech”, ”r/apple”, ”r/android”
Military	”r/army”, ”r/navy”, ”r/airforce”, ”r/usmc”, ”r/military”, ”r/veterans”
Unrelated	”r/pics”, ”r/cats”, ”r/juggling”, ”r/woodworking”

TABLE III
BINARY ATTITUDE QUESTIONS

Question Text	Hypothesis Label
Is the content related to the context?	Related Content
Is the content positive with respect to the context?	Positive Relationship
Is the content negative with respect to the context?	Negative Relationship
Does the content contain sarcasm?	Sarcastic Content
Does the content contain an analogy?	Analogous Content
Is the content positive?	Positive Content
Is the content negative?	Negative Content

B. Attitude Question Definition

Before annotation could begin, the concept of attitude needed to be converted into a concrete set of questions that would provide the meat of an attitude-based analysis. What questions, if answered, would provide a useful set of metrics for investigating or modeling the attitudes of a group? To that end, a set of simple, direct questions were composed, intending to capture a notion of attitude. These questions are listed in Table III.

C. Data Annotation

To answer these questions, and provide useful training and evaluation data, undergraduate students were employed as human annotators. Annotators were tasked with answering the full set of questions for a provided set of context-content pairs. Annotators were trained on the task, which was structured around a custom database scheme and python scripts used to automatically create and ingest cohort task sheets. After training, annotators were grouped into cohorts of 3 members, with the goal of obtaining redundant triplicate labels for each context-content-question example. Each group of 3 annotators were presented with the same set of examples to label, with no overlap in the examples provided to the different cohorts. Annotators were trained to label provided examples as yes,

TABLE IV
ANNOTATOR AGREEMENT SUMMARY.

Annotators	Agreement Level	Count
2	2/2 Agree	2409
3	2/3 Agree	1019
3	3/3 Agree	913
2 and 3	'Maj+'	4341

no, or unrelated. This was done to limit the effect of an annotator being 'forced' to choose a label when no clear option is present. This is in line with best practices for the design of attitude surveys in the domain specific literature [8].

Due to the nature of undergraduate employment, there were differences in the amount of labelling done by each annotator, with some annotators only completing a subset of their assigned tasks. This resulted in a varying degree of label quality amongst the annotated data, whereby an annotated example may have labels provided by 1, 2 or 3 of the assigned annotator.

Ultimately, this work made use of the annotated data that was given labels by 2 or 3 annotators. This data was further filtered to only consider examples that had at least a 'majority' of agreement amongst the labels provided by annotators. That is, the training data used in this work is composed of example triples that have been annotated by 2 or 3 annotators, where either 2 out of 2 agree, 2 out of 3 agree, or 3 out of 3 agree. We denote this type of training data as 'Maj+'. The total amount of data found and annotated, broken down by annotation level, is shown in Table IV.

D. Annotation Analysis

To evaluate the degree of agreement across annotators within this 'Maj+' data, a modified version of Fleiss' kappa is used [20]. This value is used to evaluate agreement by capturing the ratio of observed agreement to the amount one would expect to appear by chance. The canonical formula for the kappa is given by equation 1.

$$\kappa = \frac{\bar{P} - \bar{P}_c}{1 - \bar{P}_c} \tag{1}$$

where $\bar{P}$ is a measure of the observed agreement, and $\bar{P}_c$ represents the measure of agreement one would expect by random chance. For comparison, we make a small change to the typical formulation for these values, where we consider the size of a cohort (3) instead of the total number of annotators when scaling proportional values. The forms used in this work are shown in equation 2. The observed agreement is calculated by averaging the agreement over each of the N labeled examples. The agreement over a single example is measured by squaring the counts of each observed labels (in our case, this means counting how many yes and no labels are given for a question), and subtracting the number of annotators. To determine the agreement one would expect by chance, the proportion of examples, or rows, that contain a particular label are calculated and normalized (the p_j values in the following equations).

$$\begin{aligned}
\bar{P} &= \frac{1}{N}\sum_{i}^{N} P_i \\
&= \frac{1}{Nn^*(n^*-1)}\left(\sum_{i}^{N}(n_{0,i}^2 + n_{1,i}^2 - n^*)\right) \\
\bar{P}_c &= \sum_{j\in 0,1} p_j^2 \\
&= \sum_{j\in 0,1}\left(\frac{1}{Nn^*}\left(\sum_{i}^{N} n_{j,i}\right)\right)^2
\end{aligned} \tag{2}$$

When we use the value of $n^* = 3$, the size of our annotating cohorts, we see a Fleiss' kappa score of 0.018, non-negative, but only barely suggesting agreement. However, when we consider the size of the n^* to be equal to the average annotators-per example ($n^* = 2.42$), which considers the fact that some cohorts only had partially active members, we see a kappa value of 0.76, suggesting a strong agreement amongst the active annotators on their answers to the attitude questions.

IV. PROMPT-BASED APPROACHES

To support the utility of our proposed method, we must compare it to existing language models that might be directly suited for our task. Due to the limited size of the training data set, we consider off-the-shelf models that can be near-directly used with our data. In each of the following cases, we construct prompts out of our data examples, and annotate them with labels that follow the same structure as the off-the-shelf model. In the following sections, we outline the models, how we generate prompts for them with our data, and explain how we modify our labels to fit their pipeline.

These three methods are all based on the transformer architecture, with various pipelines and novel prediction layers used to solve specific problems. We first consider a BERT-based sequence classification pipeline, which is tasked with classifying a sentence based on predicting the polarity of a piece of text. Then, we consider a BART based Multi-genre Natural Language Inference (MNLI) model that is tasked with classifying a context and a hypothesis statement together into a label of 'contradiction', 'neutral', or 'entailment'. Finally, we look at a RoBERTa based question and answer pipeline trained to select the correct answer for a context-question pair, by predicting the location of the answer withing the context tokens. The following sections further describe how our data set was reformulated to work with each pipeline, as well as provide additional background for individual methods.

A. BERT Sequence Classification

The first of prompt-based model considered is a BERT-based model [21] fine-tuned using the yelp-polarity data set. This model operates on the tokenization of a framed input

statement, which contains a review and a user comment. The only pre-processing done to support this pipeline was to construct framed prompts using our context, content, and question strings. These prompts simply separate the three pieces of natural language content with the unique '[SEP]', or separator, token. Each of these prompts is then labeled with the yes (1) or no (0) value provided by the annotators.

B. Zero-shot Classification with BART-MNLI

The BART Multi-genre Natural Language Inference (MNLI) model is a BART based architecture [22] trained on the Multi-genre Natural Language data set which consists of statements and hypothesis labeled with three types of inference classes; neutral, contradiction, and entailment [23]. The combination of the BART embeddings, along with the labeled MNLI data, allows for the zero-shot BART MNLI pipeline to classify a statement into a set of categories that are provided at the same time as the prompt. This is done by providing each label, inside of a framing statement like "the example is a _____". The pre-trained MNLI model then processes the context and hypothesis to output a probability for the three inference classes. A high value in the 'entailment' dimension indicates that the label likely applies to the provided context.

To utilize the MNLI pipeline for our task, our data is pre-processed into prompts that contain a framing statement composed of the context and the user content, along with one of the possible hypothesis statements (as listed in Table III). These hypothesis labels are slightly reworded versions of the questions answered by annotators. The MNLI model provides an output that contains probabilities for each of the three labels; entailment, neutral, and contradiction. To match these, we encode the yes/no answers in the language of these three labels. To do this, we define the 'yes' label as a $[1.0, 0.0, 0.0]$ vector (only entailment has a non-zero value) and a 'no' label as a $[0.0, 0.0, 1.0]$ vector (only contradiction has a non-zero value).

The format of the prompt used can be found in Table V. The MNLI model takes in this prompt, with the context and content strings inserted, along with one of the hypothesis statements. The result of the model for each of the hypothesis statements is used to build the full set of answers to the attitude questions.

C. RoBERTa Question-Answering

The final prompt-based model considered is a RoBERTa based pipeline for question answering [24]. In this formulation, the model takes in the text of a question, and the text of a context, from which an answer is selected. The model outputs an answer in the form of start and end locations for text within the provided context. The evaluated RoBERTa model was fine-tuned using the SqUAd question-answering benchmark data set [25].

To prepare our data to a format that can be operated on by the RoBERTa pipeline, we convert each context-content-question example into a framed prompt. This prompt contains the text of the context string, the text of the user content, and the phrase "The answer is expected to be a yes or a no.". This final phrase is included to ensure that the model always has a possible location to provide for either answer. Each of these prompts is then labeled with an answer object, which contains the location of either the yes or no for that prompt.

Similar to the MNLI pipeline, the prompt is filled in with the context and content string, and then provided to the question and answering model along with one of the question strings.

TABLE V
THE THREE PRE-TRAINED LANGUAGE MODELS AND THEIR ASSOCIATED PROMPT FORMATS.

Model	Prompt
BERT Sequence Classifier	"{CONTEXT STRING} [SEP] {CONTENT STRING} [SEP] {QUESTION STRING}"
RoBERTA MNLI Classifier	"Posts were searched for using this context: '{CONTEXT STRING}'. A user commented the following content on one of the returned posts: '{CONTENT STRING}'."
BART Question-Answering	"Posts were searched for using this context: '{CONTEXT STRING}'. A user commented the following content on one of the returned posts: '{CONTENT STRING}'. Questions about the context and content should be answered with a yes or no."

V. PROPOSED MODEL

To address the limitations of the large language models, and lack of sufficient data to train a bespoke one, we propose a parameter-efficient prediction head that can make use of the already-learned embeddings captured by a large language model. In section A, we outline the full 'Context Content Question' based CCQ model, which considers the embeddings of the context, content, and questions. In section B, we discuss the ablated versions of the model, which consider subsets of the three natural language inputs; 'Context-Content' only, 'Context-Question' only, and 'Content' only based models.

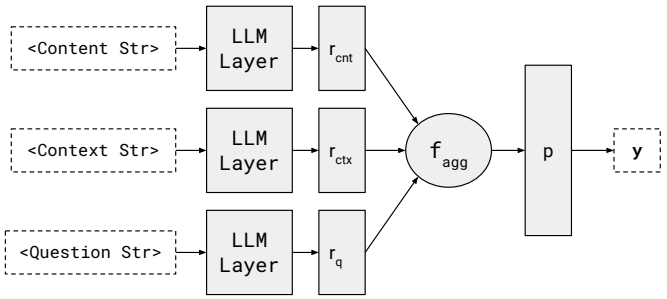

Fig. 1. CCQ Model - The main model from which described approaches are derived. The CCQ Model considers all three elements of the input, with other ablated versions removing one or more of the three input sequences.

A. CCQ Model

The motivating hypothesis for this model is that there is useful information in the semantic signal of the context, content, and question strings. To access this information, we propose a prediction head that passes each of the three pieces of embedded content through their own series of learned dimensionality reducing layers, which are then aggregated through a concatenation function. The output of this aggregation is passed to another set of fully connected layers that ultimately produce a prediction. This output is interpreted as the answer to the attitude question over the context and content pair. The overall architecture can be seen in Figure 1. During training, the pre-trained model's weights are allowed to update, and the same pre-trained model layer is used to embed all three of the input sequences.

The goal with this model is to balance the limited amount of data with the parameter count of the prediction head. Additionally, we seek to show that this combination of parsimonious complexity and small data is sufficient to outperform the more general, but more complex, off-the-shelf models.

B. Ablated Models

To further investigate the capability of the proposed approach, ablated versions of the model are considered. These versions are built to consider the removal of different input elements from the prediction flow and come in two main groups; multi-label and binary.

We consider two multi-label models; a Content-Only (CO) model and a Context-Content (CC) mode. These two models do not have a notion of the question being asked to them, as they are not provided the sequence for the attitude question. To account for this, these models output a multi-label vector. This allows them to answer each question, but removes the semantic information in the question sequence from the model itself.

The final ablated model is a binary model similar to the proposed CCQ model, a Content-Question (CQ) model. This considers an aggregation of the embeddings of the sequences, but only includes the content and question embeddings in the model. This model outputs a single binary value representing the answer to the attitude question. However, it is attempting to make this answer without knowledge of the semantic information in the context description.

VI. Evaluation

In these experiments, we compare the accuracy for the defined attitude-question-answering for the proposed model, vs the ablated versions and prompt-based alternatives. The models are trained and evaluated using data composed of annotations with two or three annotators (Maj+). We include three-annotator data where either 2 out of 3, or 3 out of 3 annotators agree. We include two-annotator data where both annotators agree. Evaluating the dataset, it became clear that there was a data imbalance across the contexts and their labeled answers. To account for this, we utilized an oversampling data balancing approach. The dataset was oversampled to include an amount of data from each context, such that each context is represented by the same amount of individually labeled examples. Additionally, when building batches and folds, a balanced subsampler was used to ensure a smooth distribution of labels was seen during training.

For the multi-label models, the labels are aggregated and converted to be multi-hot tensors. For this comparison, all of the compared models use concatenation as their aggregation function. Models are then trained using a 5-fold cross validation scheme, with each fold being trained for 50 epochs, using an SGD optimizer over a binary-cross-entropy loss function. The datasets contain examples built from roughly 670 Reddit posts, with 7 questions asked over each.

The proposed AQA model, as well as the sequence classification model, provide binary outputs which can be reported using a straightforward metric of accuracy. The MNLI approach outputs a multi-label classification tensor, which we then interpret as a binary value, allowing for another simple accuracy metric. For the question-answering pipeline, we must consider an actual text output; provided as the start and end tokens for an answer. In addition to the full proposed AQA model, we also report the accuracy of the three additional, ablated models.

VII. Results

The test accuracies for the considered approaches can be found in Tables VI and VII. Additionally, the test accuracy for each approach is broken down by the individual question and context strings in Table VIII.

The poor performance of the OTS model suggests that a custom prediction layer is needed. This was to be expected, as the structure of the task used to train the polarity-based BERT model assumed only two subsections of input text, while this task must consider three. The added structure of the polarity prediction layer prevents the model from fully accommodating the slightly different task. We can see that when the underlying BERT embeddings are used to compose a new prediction layer that does consider the three distinct embeddings, that we can achieve high accuracies. Additionally,

TABLE VI
PROMPT-BASED MODEL ACCURACIES.

Model	Type	Accuracy
BERT Sequence Classifier	Off-the-shelf	13%
RoBERTA NLI Classifier	Off-the-shelf	87.09%
BART Question-Answering	Off-the-shelf	86.57%

TABLE VII
PROPOSED MODEL AND ABLATED VERSIONS ACCURACIES.

Semantic Information Used			
Content	Context	Question	Acc
X			88.09%
X	**X**		88.10%
X		**X**	88.16%
X	**X**	**X**	**89.15**

conducted experiments provide evidence that the semantic information of the question itself is useful to the predictive model. This is seen in the difference between the multi-label and binary approaches average performance across the entire dataset.

When we look at the break down of accuracy by question, we see that the binary models can better handle the relevancy and context-content polarity better than the multi-label models. While the multilabel models outperform on the sarcasm, analogy, and gross sentiment (positive or negative) questions, the binary models outperform on all contexts except for one. This suggests that binary formulation, as implemented here, may be more robust to arbitrary contexts than the multilabel approaches.

VIII. CONCLUSIONS

These results suggest that that current off-the-shelf models may be able to handle the task of general question answering, but that a simple prediction head can also be fine-tuned to fully leverage the semantic content of the context, content, and question strings and improve on their accuracies. Additionally, it shows that the provided data set is a useful source of ground truth for the task of understanding online attitudes.

In the future, the work in this contribution could be integrated with methods that automatically extract named entities or approaches that rank content by some metric, to create systems for automated monitoring of developing contexts and a populations related attitudes. Such systems could observe content from online social networks and specific sub communities within them, for a variety of tasks, such as identifying growing political issues by observing political and news communities or developing an understanding of market needs by observing market-related communities.

Additionally, it would be useful to explore the integration of social features, based on the posting and commenting behaviour of the user that has authored the observed data. This would be improved further by expanding the set of considered social networks beyond Reddit.com

TABLE VIII
PER-QUESTION AND PER-CONTEXT ACCURACIES FOR CONTENT-CONTEXT, CONTENT-QUESTION, AND FULL CCQ MODELS.

	Content Context	Content Question	Content Context Question (CCQ)
Question String			
Is the content related to the context?	66.0%	**90.02%**	88.4%
Is the content positive with respect to the context?	89.4%	84.1%	**90.0%**
Is the content negative with respect to the context?	91.9%	92.0% 0	**93.9%**
Does the content contain sarcasm?	**95.7%**	96.7%	96.7%
Does the content contain an analogy?	**98.4%**	97.7%	97.7%
Is the content positive?	**89.0%**	82.1%	86.5%
Is the content negative?	**81.1%**	80.4%	81.0%
Context String			
The military and its use of drones	85.6%	90.7%	**93.0%**
The military and its use of tech	87.9%	88.5%	**90.9%**
Technology and invasion of privacy	84.9%	94.5%	**96.3%**
Technology and its role in fake news	87.5%	**95.5%**	**95.5%**
The presidency of Joe Biden	88.5%	91.2%	**92.5%**
The effectiveness of vaccines.	**88.3%**	87.3%	87.0%

IX. ACKNOWLEDGMENT

Research was sponsored by the DEVCOM Analysis Center and was accomplished under Cooperative Agreement Number W911NF-22-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This research is also supported in part by Temple University office of the Vice President for Research 2022 Catalytic Collaborative Research Initiative Program. AI & ML Focus Area.

REFERENCES

[1] E. Cage, J. Di Monaco, and V. Newell, "Understanding, attitudes and dehumanisation towards autistic people," *Autism*, vol. 23, no. 6, pp. 1373–1383, 2019.

[2] K. Pogue, J. L. Jensen, C. K. Stancil, D. G. Ferguson, S. J. Hughes, E. J. Mello, R. Burgess, B. K. Berges, A. Quaye, and B. D. Poole, "Influences on attitudes regarding potential covid-19 vaccination in the united states," *Vaccines*, vol. 8, no. 4, p. 582, 2020.

[3] B. Aydin, "Public acceptance of drones: Knowledge, attitudes, and practice," *Technology in society*, vol. 59, p. 101180, 2019.

[4] R. C. Stedman, N. A. Connelly, T. A. Heberlein, D. J. Decker, and S. B. Allred, "The end of the (research) world as we know it? understanding and coping with declining response rates to mail surveys," *Society & Natural Resources*, vol. 32, no. 10, pp. 1139–1154, 2019.

[5] Q. G. To, K. G. To, V.-A. N. Huynh, N. T. Nguyen, D. T. Ngo, S. Alley, A. N. Tran, A. N. Tran, N. T. Pham, T. X. Bui *et al.*, "Anti-vaccination attitude trends during the covid-19 pandemic: A machine learning-based analysis of tweets," *Digital Health*, vol. 9, p. 20552076231158033, 2023.

[6] P. Sobkowicz, M. Kaschesky, and G. Bouchard, "Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web," *Government information quarterly*, vol. 29, no. 4, pp. 470–479, 2012.

[7] F. Tang, L. Fu, B. Yao, and W. Xu, "Aspect based fine-grained sentiment analysis for online reviews," *Information Sciences*, vol. 488, pp. 190–204, 2019.

[8] C. Andrade, V. Menon, S. Ameen, and S. Kumar Praharaj, "Designing and conducting knowledge, attitude, and practice surveys in psychiatry: practical guidance," *Indian Journal of Psychological Medicine*, vol. 42, no. 5, pp. 478–481, 2020.

[9] A. N. Medvedev, R. Lambiotte, and J.-C. Delvenne, "The anatomy of reddit: An overview of academic research," *Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches 10*, pp. 183–204, 2019.

[10] A. Alambo, M. Gaur, U. Lokala, U. Kursuncu, K. Thirunarayan, A. Gyrard, A. Sheth, R. S. Welton, and J. Pathak, "Question answering for suicide risk assessment using reddit," in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. IEEE, 2019, pp. 468–473.

[11] P. Jain, K. R. Srinivas, and A. Vichare, "Depression and suicide analysis using machine learning and nlp," in *Journal of Physics: Conference Series*, vol. 2161. IOP Publishing, 2022, p. 012034.

[12] S. C. Long, B. Lucey, Y. Xie, L. Yarovaya *et al.*, ""i just like the stock": The role of reddit sentiment in the gamestop share rally," *The Financial Review*, vol. 58, no. 1, pp. 19–37, 2023.

[13] J. Zomick, S. I. Levitan, and M. Serper, "Linguistic analysis of schizophrenia in reddit posts," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 74–83.

[14] A. Rogers, M. Gardner, and I. Augenstein, "Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–45, 2023.

[15] B. Fu, Y. Qiu, C. Tang, Y. Li, H. Yu, and J. Sun, "A survey on complex question answering over knowledge base: Recent advances and challenges," *arXiv preprint arXiv:2007.13069*, 2020.

[16] R. Das, A. Godbole, A. Naik, E. Tower, M. Zaheer, H. Hajishirzi, R. Jia, and A. McCallum, "Knowledge base question answering by case-based reasoning over subgraphs," in *International Conference on Machine Learning*. PMLR, 2022, pp. 4777–4793.

[17] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, "Eli5: Long form question answering," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3558–3567.

[18] S. Zad, M. Heidari, H. James Jr, and O. Uzuner, "Emotion detection of textual data: An interdisciplinary survey," in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0255–0261.

[19] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of bert-based approaches," *Artificial Intelligence Review*, pp. 1–41, 2021.

[20] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and psychological measurement*, vol. 33, no. 3, pp. 613–619, 1973.

[21] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019. [Online]. Available: http://arxiv.org/abs/1910.13461

[23] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 1112–1122. [Online]. Available: http://aclweb.org/anthology/N18-1101

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[25] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *arXiv e-prints*, p. arXiv:1606.05250, 2016.