The Importance of Written Explanations in Aggregating Crowdsourced Predictions

Abstract

This study demonstrates that incorporating the written explanations provided by individuals when making predictions enhances the accuracy of aggregated crowdsourced forecasts. The research shows that while majority and weighted vote methods are effective, the inclusion of written justifications improves forecast accuracy throughout most of a question's duration, with the exception of its final phase. Furthermore, the study analyzes the attributes that differentiate reliable and unreliable justifications.

1 Introduction

The concept of the "wisdom of the crowd" posits that combining information from numerous non-expert individuals can produce answers that are as accurate as, or even more accurate than, those provided by a single expert. A classic example of this concept is the observation that the median estimate of an ox's weight from a large group of fair attendees was remarkably close to the actual weight. While generally supported, the idea is not without its limitations. Historical examples demonstrate instances where crowds behaved irrationally, and even a world chess champion was able to defeat the combined moves of a crowd.

In the current era, the advantages of collective intelligence are widely utilized. For example, Wikipedia relies on the contributions of volunteers, and community-driven question-answering platforms have garnered significant attention from the research community. When compiling information from large groups, it is important to determine whether the individual inputs were made independently. If not, factors like group psychology and the influence of persuasive arguments can skew individual judgments, thus negating the positive effects of crowd wisdom.

This paper focuses on forecasts concerning questions spanning political, economic, and social domains. Each forecast includes a prediction, estimating the probability of a particular event, and a written justification that explains the reasoning behind the prediction. Forecasts with identical predictions can have justifications of varying strength, which, in turn, affects the perceived reliability of the predictions. For instance, a justification that simply refers to an external source without explanation may appear to rely heavily on the prevailing opinion of the crowd and might be considered weaker than a justification that presents specific, verifiable facts from external resources.

To clarify the terminology used: a "question" is defined as a statement that seeks information (e.g., "Will new legislation be implemented before a certain date?"). Questions have a defined start and end date, and the period between these dates constitutes the "life" of the question. "Forecasters" are individuals who provide a "forecast," which consists of a "prediction" and a "justification." The prediction is a numerical representation of the likelihood of an event occurring. The justification is the text provided by the forecaster to support their prediction. The central problem addressed in this work is termed "calling a question," which refers to the process of determining a final prediction by aggregating individual forecasts. Two strategies are employed for calling questions each day throughout their life: considering forecasts submitted on the given day ("daily") and considering the last forecast submitted by each forecaster ("active").

Inspired by prior research on recognizing and fostering skilled forecasters, and analyzing written justifications to assess the quality of individual or collective forecasts, this paper investigates the automated calling of questions throughout their duration based on the forecasts available each day. The primary contributions are empirical findings that address the following research questions:

* When making a prediction on a specific day, is it advantageous to include forecasts from previous days? (Yes) * Does the accuracy of the prediction improve when considering the question itself and the written justifications provided with the forecasts? (Yes) * Is it easier to make an accurate prediction toward the end of a question's duration? (Yes) * Are written justifications more valuable when the crowd's predictions are less accurate? (Yes)

In addition, this research presents an examination of the justifications associated with both accurate and inaccurate forecasts. This analysis aims to identify the features that contribute to a justification being more or less credible.

2 Related Work

The language employed by individuals is indicative of various characteristics. Prior research includes both predictive models (using language samples to predict attributes about the author) and models that provide valuable insights (using language samples and author attributes to identify differentiating linguistic features). Previous studies have examined factors such as gender and age, political ideology, health outcomes, and personality traits. In this paper, models are constructed to predict outcomes based on crowd-sourced forecasts without knowledge of individual forecasters' identities.

Previous research has also explored how language use varies depending on the relationships between individuals. For instance, studies have analyzed language patterns in social networks, online communities, and corporate emails to understand how individuals in positions of authority communicate. Similarly, researchers have examined how language provides insights into interpersonal interactions and relationships. In terms of language form and function, prior research has investigated politeness, empathy, advice, condolences, usefulness, and deception. Related to the current study's focus, researchers have examined the influence of Wikipedia editors and studied influence levels within online communities. Persuasion has also been analyzed from a computational perspective, including within the context of dialogue systems. The work presented here complements these previous studies. The goal is to identify credible justifications to improve the aggregation of crowdsourced forecasts, without explicitly targeting any of the aforementioned characteristics.

Within the field of computational linguistics, the task most closely related to this research is argumentation. A strong justification for a forecast can be considered a well-reasoned supporting argument. Previous work in this area includes identifying argument components such as claims, premises, backing, rebuttals, and refutations, as well as mining arguments that support or oppose a particular claim. Despite these efforts, it was found that crowdsourced justifications rarely adhere to these established argumentation frameworks, even though such justifications are valuable for aggregating forecasts.

Finally, several studies have focused on forecasting using datasets similar or identical to the one used in this research. From a psychological perspective, researchers have explored strategies for enhancing forecasting accuracy, such as utilizing top-performing forecasters (often called "superforecasters"), and have analyzed the traits that contribute to their success. These studies aim to identify and cultivate superforecasters but do not incorporate the written justifications accompanying forecasts. In contrast, the present research develops models to call questions without using any information about the forecasters themselves. Within the field of computational linguistics, researchers have evaluated the language used in high-quality justifications, focusing on aspects like rating, benefit, and influence. Other researchers have developed models to predict forecaster skill using the textual justifications from specific datasets, such as the Good Judgment Open data, and have also applied these models to predict the accuracy of individual forecasts in other contexts, such as company earnings reports. However, none of these prior works have specifically aimed to call questions throughout their entire duration.

3 Dataset

The research utilizes data from the Good Judgment Open, a platform where questions are posted, and individuals submit their forecasts. The questions primarily revolve around geopolitics, encompassing areas such as domestic and international politics, the economy, and social matters. For this study, all binary questions were collected, along with their associated forecasts, each comprising a prediction and a justification. In total, the dataset contains 441 questions and 96,664 forecasts submitted over 32,708 days. This dataset significantly expands upon previous research, nearly doubling the number of forecasts analyzed. Since the objective is to accurately call questions throughout their entire duration, all forecasts with written justifications are included, regardless of factors such as justification length or the number of forecasts submitted by a single forecaster. Additionally, this approach prioritizes privacy, as no information about the individual forecasters is utilized.

Table 1: Analysis of the questions from our dataset. Most questions are relatively long, contain two or more named entities, and are open for over one month.

Metric	Min	Q1	Q2 (Median)	Q3	Max	Mean
# tokens	8	16	20	28	48	21.94
# entities	0	2	3	5	11	3.47
# verbs	0	2	2	3	6	2.26
# days open	2	24	59	98	475	74.16

Table 1 provides a basic analysis of the questions in the dataset. The majority of questions are relatively lengthy, containing more than 16 tokens and multiple named entities, with geopolitical, person, and date entities being the most frequent. In terms of duration, half of the questions remain open for nearly two months, and 75% are open for more than three weeks.

An examination of the topics covered by the questions using Latent Dirichlet Allocation (LDA) reveals three primary themes: elections (including terms like "voting," "winners," and "candidate"), government actions (including terms like "negotiations," "announcements," "meetings," and "passing (a law)"), and wars and violent crimes (including terms like "groups," "killing," "civilian (casualties)," and "arms"). Although not explicitly represented in the LDA topics, the questions address both domestic and international events within these broad themes.

Table 2: Analysis of the 96,664 written justifications submitted by forecasters in our dataset. The readability scores indicate that most justifications are easily understood by high school students (11th or 12th grade), although a substantial amount (>25%) require a college education (Flesch under 50 or Dale-Chall over 9.0).

	Min	Q1	Q2	Q3	Max
#sentences	1	1	1	3	56
#tokens	1	10	23	47	1295
#entities	0	0	2	4	154
#verbs	0	1	3	6	174
#adverbs	0	0	1	3	63
#adjectives	0	0	2	4	91
#negation	0	0	1	3	69
Sentiment	-2.54	0	0	0.20	6.50
Readability					
Flesch	-49.68	50.33	65.76	80.62	121.22
Dale-Chall	0.05	6.72	7.95	9.20	19.77

Table 2 presents a fundamental analysis of the 96,664 forecast justifications in the dataset. The median length is relatively short, consisting of one sentence and 23 tokens. Justifications mention named entities less frequently than the questions themselves. Interestingly, half of the justifications contain at least one negation, and 25% include three or more. This suggests that forecasters sometimes base their predictions on events that might not occur or have not yet occurred. The sentiment polarity of the justifications is generally neutral. In terms of readability, both the Flesch and Dale-Chall scores suggest that approximately a quarter of the justifications require a college-level education for full comprehension.

Regarding verbs and nouns, an analysis using WordNet lexical files reveals that the most common verb classes are "change" (e.g., "happen," "remain," "increase"), "social" (e.g., "vote," "support," "help"), "cognition" (e.g., "think," "believe," "know"), and "motion" (e.g., "go," "come," "leave"). The most frequent noun classes are "act" (e.g., "election," "support," "deal"), "communication" (e.g., "questions," "forecast," "news"), "cognition" (e.g., "point," "issue," "possibility"), and "group" (e.g., "government," "people," "party").

4 Experiments and Results

Experiments are conducted to address the challenge of accurately calling a question throughout its duration. The input consists of the question itself and the associated forecasts (predictions and justifications), while the output is an aggregated answer to the question derived from all forecasts. The number of instances corresponds to the total number of days all questions were open. Both simple baselines and a neural network are employed, considering both (a) daily forecasts and (b) active forecasts submitted up to ten days prior.

The questions are divided into training, validation, and test subsets. Subsequently, all forecasts submitted throughout the duration of each question are assigned to their respective subsets. It's important to note that randomly splitting the forecasts would be an inappropriate approach. This is because forecasts for the same question submitted on different days would be distributed across the training, validation, and test subsets, leading to data leakage and inaccurate performance evaluation.

4.1 Baselines

Two unsupervised baselines are considered. The "majority vote" baseline determines the answer to a question based on the most frequent prediction among the forecasts. The "weighted vote" baseline, on the other hand, assigns weights to the probabilities in the predictions and then aggregates them.

4.2 Neural Network Architecture

A neural network architecture is employed, which consists of three main components: one to generate a representation of the question, another to generate a representation of each forecast, and an LSTM to process the sequence of forecasts and ultimately call the question.

The representation of a question is obtained using BERT, followed by a fully connected layer with 256 neurons, ReLU activation, and dropout. The representation of a forecast is created by concatenating three elements: (a) a binary flag indicating whether the forecast was submitted on the day the question is being called or on a previous day, (b) the prediction itself (a numerical value between 0.0 and 1.0), and (c) a representation of the justification. The representation of the justification is also obtained using BERT, followed by a fully connected layer with 256 neurons, ReLU activation, and dropout. The LSTM has a hidden state with a dimensionality of 256 and processes the sequence of forecasts as its input. During the tuning process, it was discovered that providing the representation of the question alongside each forecast is more effective than processing forecasts independently of the question. Consequently, the representation of the question is concatenated with the representation of each forecast before being fed into the LSTM. Finally, the last hidden state of the LSTM is connected to a fully connected layer with a single neuron and sigmoid activation to produce the final prediction for the question.

4.3 Architecture Ablation

Experiments are carried out with the complete neural architecture, as described above, as well as with variations where certain components are disabled. Specifically, the representation of a forecast is manipulated by incorporating different combinations of information:

* Only the prediction. * The prediction and the representation of the question. * The prediction and the representation of the justification. * The prediction, the representation of the question, and the representation of the justification.

4.4 Quantitative Results

The evaluation metric used is accuracy, which represents the average percentage of days a model correctly calls a question throughout its duration. Results are reported for all days combined, as well as for each of the four quartiles of the question's duration.

Table 3: Results with the test questions (Accuracy: average percentage of days a model predicts a question correctly). Results are provided for all days a question was open and for four quartiles (Q1: first 25% of days, Q2: 25-50%, Q3: 50-75%, and Q4: last 25% of days).

	Days When the Question Was Open				
Model	All Days	Q1	Q2	Q3	Q4
Using Daily Forecasts Only					
Baselines					
Majority Vote (predictions)	71.89	64.59	66.59	73.26	82.22
Weighted Vote (predictions)	73.79	67.79	68.71	74.16	83.61
Neural Network Variants					
Predictions Only	77.96	77.62	77.93	78.23	78.61
Predictions + Question	77.61	75.44	76.77	78.05	81.56
Predictions + Justifications	80.23	77.87	78.65	79.26	84.67
Predictions + Question + Justifications	79.96	78.65	78.11	80.29	83.28
Using Active Forecasts					
Baselines					
Majority Vote (predictions)	77.27	68.83	73.92	77.98	87.44
Weighted Vote (predictions)	77.97	72.04	72.17	78.53	88.22
Neural Network Variants					
Predictions Only	78.81	77.31	78.04	78.53	81.11
Predictions + Question	79.35	76.05	78.53	79.56	82.94
Predictions + Justifications	80.84	77.86	79.07	79.74	86.17
Predictions + Question + Justifications	81.27	78.71	79.81	81.56	84.67

Despite their relative simplicity, the baseline methods achieve commendable results, demonstrating that aggregating forecaster predictions without considering the question or justifications is a viable strategy. However, the full neural network achieves significantly improved results.

Using Daily or Active Forecasts Incorporating active forecasts, rather than solely relying on forecasts submitted on the day the question is called, proves advantageous for both baselines and all neural network configurations, except for the one using only predictions and justifications.

Encoding Questions and Justifications The neural network that only utilizes the prediction to represent a forecast surpasses both baseline methods. Notably, integrating the question, the justification, or both into the forecast representation yields further improvements. These results indicate that incorporating the question and forecaster-provided justifications into the model enhances the accuracy of question calling.

Calling Questions Throughout Their Life When examining the results across the four quartiles of a question's duration, it's observed that while using active forecasts is beneficial across all quartiles for both baselines and all network configurations, the neural networks surprisingly outperform the baselines only in the first three quartiles. In the last quartile, the neural networks perform significantly worse than the baselines. This suggests that while modeling questions and justifications is generally helpful, it becomes detrimental toward the end of a question's life. This phenomenon can be attributed to the increasing wisdom of the crowd as more evidence becomes available and more forecasters contribute, making their aggregated predictions more accurate.

Table 4: Results with the test questions, categorized by question difficulty as determined by the best baseline model. The table presents the accuracy (average percentage of days a question is predicted correctly) for all questions and for each quartile of difficulty: Q1 (easiest 25%), Q2 (25-50%), Q3 (50-75%), and Q4 (hardest 25%).

	Question Difficulty (Based on Best Baseline)				
	All	Q1	Q2	Q3	Q4
Using Active Forecasts					
Weighted Vote Baseline (Predictions)	77.97	99.40	99.55	86.01	29.30
Neural Network with Components...					
Predictions + Question	79.35	94.58	88.01	78.04	58.73
Predictions + Justifications	80.84	95.71	93.18	79.99	57.05
Predictions + Question + Justifications	81.27	94.17	90.11	78.67	64.41

Calling Questions Based on Their Difficulty The analysis is further refined by examining results based on question difficulty, determined by the number of days the best-performing baseline incorrectly calls the question. This helps to understand which questions benefit most from the neural networks that incorporate questions and justifications. However, it's important to note that calculating question difficulty during the question's active period is not feasible, making these experiments unrealistic before the question closes and the correct answer is revealed.

Table 4 presents the results for selected models based on question difficulty. The weighted vote baseline demonstrates superior performance for 75

5 Qualitative Analysis

This section provides insights into the factors that make questions more difficult to forecast and examines the characteristics of justifications associated with incorrect and correct predictions.

Questions An analysis of the 88 questions in the test set revealed that questions called incorrectly on at least one day by the best model tend to have a shorter duration (69.4 days vs. 81.7 days) and a higher number of active forecasts per day (31.0 vs. 26.7). This suggests that the model's errors align with the questions that forecasters also find challenging.

Justifications A manual review of 400 justifications (200 associated with incorrect predictions and 200 with correct predictions) was conducted, focusing on those submitted on days when the best model made an incorrect prediction. The following observations were made:

* A higher percentage of incorrect predictions (78%) were accompanied by short justifications (fewer than 20 tokens), compared to 65% for correct predictions. This supports the idea that longer user-generated text often indicates higher quality. * References to previous forecasts (either by the same or other forecasters, or the current crowd's forecast) were more common in justifications for incorrect predictions (31.5%) than for correct predictions (16%). * A lack of a logical argument was prevalent in the justifications, regardless of the prediction's accuracy. However, it was more frequent in justifications for incorrect predictions (62.5%) than for correct predictions (47.5%). * Surprisingly, justifications with generic arguments did not clearly differentiate between incorrect and correct predictions (16.0% vs. 14.5%). * Poor grammar and spelling or the use of non-English were infrequent but more common in justifications for incorrect predictions (24.5%) compared to correct predictions (14.5%).

6 Conclusions

Forecasting involves predicting future events, a capability highly valued by both governments and industries as it enables them to anticipate and address potential challenges. This study focuses on questions spanning the political, economic, and social domains, utilizing forecasts submitted by a crowd of individuals without specialized training. Each forecast comprises a prediction and a natural language justification.

The research demonstrates that aggregating the weighted predictions of forecasters is a solid baseline for calling a question throughout its duration. However, models that incorporate both the question and the justifications achieve significantly better results, particularly during the first three quartiles of a question's life. Importantly, the models developed in this study do not profile individual forecasters or utilize any information about their identities. This work lays the groundwork for evaluating the credibility of anonymous forecasts, enabling the development of robust aggregation strategies that do not require tracking individual forecasters.