Measuring Forecasting Skill from Text

Shi Zong[1] **Alan Ritter**[1] **Eduard Hovy**[2]
[1]Department of Computer Science and Engineering, The Ohio State University
[2]Language Technologies Institute, Carnegie Mellon University
{zong.56, ritter.1492}@osu.edu, hovy@cs.cmu.edu

Abstract

People vary in their ability to make accurate predictions about the future. Prior studies have shown that some individuals can predict the outcome of future events with consistently better accuracy. This leads to a natural question: what makes some forecasters better than others? In this paper we explore connections between the language people use to describe their predictions and their forecasting skill. Datasets from two different forecasting domains are explored: (1) geopolitical forecasts from Good Judgment Open, an online prediction forum and (2) a corpus of company earnings forecasts made by financial analysts. We present a number of linguistic metrics which are computed over text associated with people's predictions about the future including: uncertainty, readability, and emotion. By studying linguistic factors associated with predictions, we are able to shed some light on the approach taken by skilled forecasters. Furthermore, we demonstrate that it is possible to accurately predict forecasting skill using a model that is based solely on language. This could potentially be useful for identifying accurate predictions or potentially skilled forecasters earlier.[1]

1 Introduction

People often make predictions about the future, for example meteorologists tell us what the weather might look like tomorrow, financial analysts predict which companies will report favorable earnings and intelligence analysts evaluate the likelihood of future geopolitical events. An interesting question is why some individuals are significantly better forecasters (Mellers et al., 2015b)?

Previous work has analyzed to what degree various factors (intelligence, thinking style, knowledge of a specific topic, etc.) contribute to a person's skill. These studies have used surveys or psychological tests to measure dispositional, situational and behavioral variables (Mellers et al., 2015a). Another source of information has been largely overlooked, however: the language forecasters use to justify their predictions. Recent research has demonstrated that it is possible to accurately forecast the outcome of future events by aggregating social media users' predictions and analyzing their veridicality (Swamy et al., 2017), but to our knowledge, no prior work has investigated whether it might be possible to measure a forecaster's ability by analyzing their language.

In this paper, we present the first systematic study of the connection between language and forecasting ability. To do so, we analyze texts written by top forecasters (ranked by accuracy against ground truth) in two domains: geopolitical forecasts from an online prediction forum, and company earnings forecasts made by financial analysts. To shed light on the differences in approach employed by skilled and unskilled forecasters, we investigate a variety of linguistic metrics. These metrics are computed using natural language processing methods to analyze sentiment (Pang et al., 2002; Wilson et al., 2005), uncertainty (de Marneffe et al., 2012; Saurí and Pustejovsky, 2012), readability, etc. In addition we make use of word lists taken from the Linguistic Inquiry and Word Count (LIWC) software (Tausczik and Pennebaker, 2010), which is widely used in psychological research. By analyzing forecasters' texts, we are able to provide evidence to support or refute hypotheses about factors that may influence forecasting skill. For example, we show forecasters whose justifications contain a higher proportion of uncertain statements tend to make more accurate predictions. This supports the hypothesis that more open-minded thinkers, who have a higher tolerance for ambiguity tend to make

[1]We provide our code and dataset descriptions at: https://github.com/viczong/measuring_forecasting_skill_from_text.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5317–5331
July 5 - 10, 2020.

better predictions (Tetlock, 2005).

Beyond analyzing linguistic factors associated with forecasting ability, we further demonstrate that it is possible to identify skilled forecasters and accurate predictions based only on relevant text. Estimating the quality of a prediction using the forecaster's language could potentially be very beneficial. For example, this does not require access to historical predictions to evaluate past performance, so it could help to identify potentially skilled individuals sooner. Also, forecasters do not always provide an explicit estimate of their confidence, so a confidence measure derived directly from text could be very useful.

2 Linguistic Cues of Accurate Forecasting

In this section, we are interested in uncovering linguistic cues in people's writing that are predictive of forecasting skill. We start by analyzing texts written by forecasters to justify their predictions in a geopolitical forecasting forum. Linguistic differences between forecasters are explored by aggregating metrics across each forecaster's predictions. In §3, we analyze the accuracy of individual predictions using a dataset of financial analysts' forecasts towards companies' (continuous) earnings per share. By controlling for differences between analysts and companies, we are able to analyze intra-analyst differences between accurate and inaccurate forecasts.

2.1 Geopolitical Forecasting Data

To explore the connections between language and forecasting skill, we make use of data from Good Judgment Open,[2] an online prediction forum. Users of this website share predictions in response to a number of pre-specified questions about future events with uncertain outcomes, such as: *"Will North Korea fire another intercontinental ballistic missile before August 2019?"* Users' predictions consist of an estimated chance the event will occur (for example, 5%) in addition to an optional text justification that explains why the forecast was made. A sample is presented in Figure 1.

Preprocessing. Not all predictions contain associated text justifications; in this work, we only consider predictions with justifications containing more than 10 tokens. We ran `langid.py` (Lui and Baldwin, 2012) to remove forecasts with non-English text, and further restrict our data to contain only users that made at least 5 predictions with text.

Question: Will Kim Jong Un visit Seoul before 1 October 2019?
Estimated Chance: 5%
Forecast Justification: No North Korean leader has stepped foot in Seoul since the partition of the Koreas at the end of the Korean War. . . .

Figure 1: A sample prediction made by a user in response to a question posted by *the Economist*.

In our pilot studies, we also notice some forecasters directly quote text from outside resources (like Wikipedia, New York Times, etc.) as part of their justifications. To avoid including justifications that are mostly copied from external sources, we remove forecasts that consist of more than 50% text enclosed in quotation marks from the data.

Dataset statistics. We collected all questions with binary answers that closed before April 9, 2019, leading to a total of 441 questions. 23,530 forecasters made 426,909 predictions. During preprocessing steps, 3,873 forecasts are identified as heavily quoted and thus removed. After removing non-English and heavily quoted forecasts, forecasts with no text justifications or justifications less than 10 tokens, in addition users with fewer than 5 predictions with text, 55,099 forecasts made by 2,284 forecasters are selected for the final dataset.

The distribution of predictions made by each forecaster is heavily skewed. 8.0% of forecasters make over 50 forecasts.[3] On average, each forecaster makes 10.3 forecasts, excluding those who made over 50 predictions. In Table 1, we also provide breakdown statistics for top and bottom forecasters.

2.2 Measuring Ground Truth

In order to build a model that can accurately classify good forecasters based on features of their language, we first need a metric to measure people's forecasting skill. For this purpose we use Brier score (Brier, 1950), a commonly used measure for evaluating probabilistic forecasts.[4] For questions

[2] https://www.gjopen.com/

[3] In our dataset, forecasters could even make over 1,000 forecasts with justifications.

[4] Other possible scoring rules exist, for example ranking forecasters by log-likelihood. For a log-likelihood scoring rule, however, we need to adjust estimates of 1.00 and 0.00, which are not uncommon in the data, to avoid zero probability events. There are many ways this adjustment could be done and it is difficult to justify one choice over another.

with binary answers, it is defined as:

$$\text{Forecaster's Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$$

Here f_i is the forecaster's estimated probability, o_i is a binary variable indicating the final outcome of the event, and N is the total number of forecasts. Brier scores can be interpreted as the mean squared error between the forecast probability and true answer; lower scores indicate better forecasts.

Ranking forecasters. Directly comparing raw Brier scores is problematic, because users are free to choose questions they prefer, and could achieve a lower Brier score simply by selecting easier questions. To address this issue, we standardized Brier scores by subtracting the mean Brier scores and dividing by the standard deviation within questions (Mellers et al., 2015a).

We construct a set of balanced datasets for training and evaluating classifiers by choosing the top K and bottom K forecasters respectively. In our experiments, we vary K from 100 to 1,000; when K=1,000, the task can be interpreted roughly as classifying all ∼2k users into the top or bottom half of forecasters.[5]

2.3 Linguistic Analysis

In §2.2, we discussed how to measure ground-truth forecasting skill by comparing a user's predictions against ground-truth outcomes. In the following subsections, we examine a selected series of linguistic phenomenon and their connections with forecasting ability. Statistical tests are conducted using the paired bootstrap (Efron and Tibshirani, 1994). As we are performing multiple hypothesis testing, we also report results for Bonferroni-corrected significance level 0.05/30.

As discussed in §2.1, the distribution of forecasts per user is highly skewed. To control for this, we compute averages for each forecaster and use aggregate statistics to compare differences between the two groups at the user-level. Analyses are performed over 6,639 justifications from the top 500 forecasters and 6,040 from bottom 500.

2.3.1 Textual Factors

Length. We first check the average length of justifications from different groups and report our results in Table 1. We observe that skilled forecasters normally write significantly longer justifications with more tokens per sentence. This suggests that good forecasters tend to provide more rationale to support their predictions.

Metric	Top 500	Btm 500	p
Forecasters statistics			
# users making ≥ 50 forecasts	20	14	-
Avg. forecasts (w/o above users)	9.4	9.2	-
Length & word counts			
Avg. # tokens per user	69.1	47.0	↑↑↑
% answers ≥ 100 tokens per user	18.5	8.3	↑↑↑
Avg. # tokens per sentence	20.9	19.2	↑↑↑

Table 1: Statistics of our dataset. p-values are calculated by bootstrap test. ↑↑↑: $p < 0.001$.

Readability. We compute two widely used metrics for readability: (1) Flesch reading ease (Flesch, 1948) and (2) Dale-Chall formula (Dale and Chall, 1948). Table 2 summarizes our results on average readability scores. We find good forecasters have lower readability compared to bad forecasters.

It is interesting to compare this result with the findings reported by Ganjigunte Ashok et al. (2013), who found a negative correlation between the success of novels and their readability, and also the work of Sawyer et al. (2008) who found award winning articles in academic marketing journals had higher readability. Our finding that more accurate forecasters write justifications that have lower readability suggests that skilled forecasters tend to use more complex language.

Emotion. We also analyze the sentiment reflected in forecasters' written text. Rather than analyzing sentiment orientation ("positive", "negative", or "neutral"), here we focus on measuring sentiment *strength*. We hypothesize that skilled forecasters organize their supporting claims in a more rational way using less emotional language. Many existing sentiment analysis tools (e.g., Socher et al. (2013)) are built on corpora such as the Stanford Sentiment Treebank, which are composed of movie reviews or similar texts. However, justifications in our dataset focus on expressing opinions towards future uncertain events, rather than simply expressing preferences toward a movie or restaurant, leading to a significant domain mismatch. In pilot studies, we noticed many sentences that are marked as negative by the Stanford sentiment analyzer on our data do not in fact express a negative emotion. We thus use Semantic Orientation CALculator (SO-

[5] Readers may wonder if there do exist differences between top and bottom forecasters. We provide justifications for our ranking approach in Appendix A.1.

Metric	p	Bonferroni
Textual Factors		
Readability		
Flesch reading ease	↓↓	
Dale-Chall	↑↑↑	∗
Emotion		
Absolute sentiment strength	↓↓↓	∗
Parts of Speech		
Cardinal	↑↑↑	∗
Noun	↑↑	
Preposition	↑↑↑	∗
Pronoun	↓↓↓	∗
1st personal pronoun	↑	
Verb	↓↓↓	∗
Cognitive Factors		
Uncertainty		
% uncertain statements	↑↑↑	∗
Tentative (LIWC)	↑↑↑	∗
Thinking style		
% forecasts with quoted text	↑↑↑	∗
Temporal orientation		
Focus on past	↑↑	
Focus on present & future	↓↓↓	∗

Table 2: Comparison of various metrics computed over text written by the top 500 and bottom 500 forecasters. Good forecasters tend to exhibit more uncertainty, cite outside resources, and tend toward neutral sentiment; they also use more complex language resulting in lower readability and focus more on past events. p-values are calculated by bootstrap test. The number of arrows indicates the level of p-value, while the direction shows the relative relationship between top and bottom forecasters, ↑↑↑: top group is higher than bottom group with $p < 0.001$, ↑↑: $p < 0.01$, ↑: $p < 0.05$. Tests that pass Bonferroni correction are marked by $*$.

CAL), a lexicon-based model proposed by Taboada et al. (2011) which has been demonstrated to have good performance across a variety of domains. The model generates a score for each justification by adding together semantic scores of words present in the justification, with a 0 score indicating a neutral sentiment. We then take the absolute values of scores from the model and calculate averages for each group. Results in Table 2 show that the top 500 forecasters have a significantly lower average sentiment strength compared to bottom 500 forecasters, indicating statements from skilled forecasters tend to express neutral sentiment.

Parts of Speech. As shown in Table 2, we observe that top forecasters use a higher percentage of cardinal numbers and nouns, while higher numbers of verbs are associated with lower forecasting ability.[6]

We also note the bottom 500 use a higher percentage of pronouns when justifying their predictions. To investigate this difference, we further separate first person pronouns[7] from second or third person pronouns. As presented in Table 2, first person pronouns are used more often by the top forecasters.

2.3.2 Cognitive Factors

We now evaluate a number of factors that were found to be related to decision making processes based on prior psychological studies (e.g., Mellers et al. (2015a)), that can be tested using computational tools. A number of these metrics are calculated by using the Linguistic Inquiry and Word Count (LIWC) lexicon (Tausczik and Pennebaker, 2010), a widely used tool for psychological and social science research.

Uncertainty. To test the hypothesis that good forecasters have a greater tolerance for uncertainty and ambiguity, we employ several metrics to evaluate the degree of uncertainty reflected in their written language. We use the model proposed by Adel and Schütze (2017) to estimate the proportion of uncertain statements made by each forecaster in our dataset. It is an attention based convolutional neural network model, that achieves state-of-the-art results on a Wikipedia benchmark dataset from the 2010 CoNLL shared task (Farkas et al., 2010); we use the trained parameters provided by Adel and Schütze (2017). After the model assigns an uncertainty label for each sentence, we calculate the percentage of sentences marked as uncertain. Results of this analysis are reported in Table 2; we observe that the top 500 forecasters make a significantly greater number of uncertain statements compared to the bottom 500, supporting the hypothesis mentioned above.

Thinking style. In §2.1, we discussed the issue that many forecasts contain quoted text. Although we removed posts consisting of mostly quoted text as a preprocessing step, we are interested in how people use outside resources during their decision making process. We thus calculate the portion of forecasts with quotes for the two groups. We notice skilled forecasters cite outside resources more frequently. This may indicate that skilled forecasters tend to account for more information taken from external sources when making predictions.

[6] POS tags were obtained using Stanford CoreNLP.
[7] "I", "me", "mine", "my" and "myself".

Temporal orientation. We make use of the LIWC lexicon (Tausczik and Pennebaker, 2010) to analyze the temporal orientation of forecasters' justifications. We notice good forecasters tend to focus more on past events (reflected by tokens like *"ago"* and *"talked"*); bad forecasters pay more attention to what is currently happening or potential future events (using tokens like *"now"*, *"will"*, and *"soon"*). We conjecture this is because past events can provide more reliable evidence for what is likely to happen in the future.

2.4 Predicting Forecasting Skill

In §2.3, we showed there are significant linguistic differences between justifications written by skilled and unskilled forecasters. This leads to a natural question: is it possible to automatically identify skilled forecasters based on the written text associated with their predictions? We examine this question in general terms first, then present experiments using a realistic setup for early prediction of forecasting skill in §2.5.

Models and features. We start with a log-linear model using bag-of-ngram features extracted from the combined answers for each forecaster. We experimented with different combinations of n-gram features from sizes 1 to 4. N-grams of size 1 and 2 have best classification accuracy. We map n-grams that occur only once to a ⟨UNK⟩ token, and replace all digits with 0. Inspired by our findings in §2.3, we also incorporate textual and cognition factors as features in our log-linear model.

We also experiment with convolutional neural networks (Kim, 2014) and BERT (Devlin et al., 2019). The 1D convolutional neural network consists of a convolution layer, a max-pooling layer, and a fully connected layer. We minimize cross entropy loss using Adam (Kingma and Ba, 2015); the learning rate is 0.01 with a batch size of 32. We fine-tune BERT on our dataset, using a batch size of 5 and a learning rate of 5e-6. All hyperparameters were selected using a held-out dev set.

Model performance. Results are presented in Table 3. As we increase the number of forecasters K, the task becomes more difficult as more forecasters are ranked in the middle. However, we observe a stable accuracy around 70%. All models consistently outperform a random baseline (50% accuracy), suggesting that the language users use to describe their predictions does indeed contain information that is predictive of forecasting ability.

The n-grams with largest weights in the logistic regression model are presented in Table 4. We find that n-grams that seem to indicate uncertainty, including: *"it seems unlikely"*, *"seem to have"* and *"it is likely"* are among the largest positive weights.

	K	100	200	300	500	1000
LR	Bag-of-ngrams	69.5	74.2	72.5	69.2	64.8
	Textual	66.0	60.8	62.0	59.3	57.4
	Cognitive	69.0	68.0	67.3	65.5	61.0
	All above	70.5	73.5	73.3	69.8	64.7
Neural	CNN	71.5	75.0	72.0	69.6	64.0
	BERT-base	74.5	77.3	74.3	69.7	65.1

Table 3: Accuracy (%) on classifying skilled forecasters when choosing the top K and bottom K forecasters. For logistic regression (LR), we experiment with different sets of features: bag of $\{1, 2\}$-grams, textual factors in §2.3.1, cognitive factors in §2.3.2, and combination of all above. For neural networks (Neural), we use convolutional neural network (CNN) and BERT-base. All results are based on 5-fold cross validation.

Top15 (High-weight)	in the next / . also , / . however , / based on the / there are no / . according to / of time . / . based on / they wo n't / there is no / it seems unlikely / do n't see / it is likely / more of a / seem to have
Bottom15 (Low-weight)	will continue to / it will be / the world . / . it 's / there is a / is not a / the west . / to be on / to be the / . yes , / he 's a / there will be / in the world / will still be / . he will

Table 4: High and low-weight n-gram features from the logistic regression model trained to identify good forecasters (K=500 with only 3-gram features for interpretability). Positive features indicate some uncertainty (e.g., *"it is likely"*, *"seem to have"* , *"it seems unlikely"*), in addition to consideration of evidence from many sources (e.g., *"based on the"*, *". according to"*).

2.5 Identifying Good Forecasters Earlier

With the model developed in §2.4, we are now ready to answer the following question: using only their first written justification, can we foresee a forecaster's future performance?

Setup. Our goal is to rank forecasters by their performance. We first equally split all 2,284 forecasters into two groups (top half versus bottom half) based on their standardized Brier scores. We then partition them into 60% train, 20% validation, and 20% test splits within each group. We combine all justifications for each forecaster in the training set. For forecasters in the validation and test sets,

we only use their single earliest forecast.

We use forecasters' final rank sorted by averaged standardized Brier score over all forecasts as ground truth. We then compare our text-based model to the following two baselines: (1) a random baseline (50%) and (2) the standardized Brier score of the users' single earliest forecast.

Results. We calculate the proportion of good forecasters identified in the top N, ranked by our text-based model, and report results in Table 5. We observe that our models achieve comparable or even better performance relative to the first prediction's adjusted Brier score. Calculating Brier scores requires knowing ground-truth, while our model can evaluate the performance of a forecaster *without* waiting to know the outcome of a predicted event.

	P@10	P@50	P@100
Brier score	60	64	62
Text-based (LR)	70	70	65
Text-based (CNN)	90	68	64
Text-based (BERT-base)	80	70	67

Table 5: Precision@N of identifying skilled forecasters based on their first prediction.

3 Companies' Earnings Forecasts

In §2, we showed that linguistic differences exist between good and bad forecasters, and furthermore, these differences can be used to predict which forecasters will perform better. We now turn to the question of whether it is possible to identify which *individual* forecasts, made by the same person, are more likely to be correct. The Good Judgment Open data is not suitable to answer this question, because forecasts are discrete, and thus do not provide a way to rank individual predictions by accuracy beyond whether they are correct or not. Therefore, in this section, we consider numerical forecasts in the financial domain, which can be ranked by their accuracy as measured against ground truth. In this paper, we analyze forecasts of companies' earnings per share (EPS). Earnings per share is defined as the portion of a company's profit allocated to each share of common stock. It is an important indicator of a company's ability to make profits. For our purposes, EPS also supports a cleaner experimental design as compared to stock prices, which constantly change in real time.

Data. We analyze reports from the Center for Financial Research and Analysis (CFRA).[8] These reports provide frequent updates for analysts' estimates and are also organized in a structured way, enabling us to accurately extract numerical forecasts and corresponding text justifications.

We collected CFRA's analyst reports from the Thomson ONE database[9] from 2014 to 2018. All notes making forecasts are extracted under the *"Analyst Research Notes and other Company News"* section. The dataset contains a total of 32,807 notes from analysts, covering 1,320 companies.

3.1 Measuring Ground Truth

We use a pattern-based approach (in Appendix B.1) for extracting numerical forecasts. After removing notes without EPS estimates, 16,044 notes on 1,135 companies remain (this is after removing analysts who make fewer than 100 forecasts as discussed later in this section). We next evaluate whether the text can reflect how accurate these predictions are.

Forecast error. We measure the correctness of forecasts by absolute relative error (Barefield and Comiskey, 1975; Dreman and Berry, 1995). The error is defined by the absolute difference between the analyst's estimate e and corresponding actual EPS o, scaled by the actual EPS:

$$\text{Forecast Error} = \frac{|e - o|}{|o|}$$

Low forecast errors indicate accurate forecasts.[10]

Ranking individual forecasts. As our goal is to study the intra-analyst differences between accurate and inaccurate forecasts, we standardize forecast errors within each analyst by subtracting the analyst's mean forecast error and then dividing by the standard deviation. To guarantee we have a good estimate for the mean, we only include analysts who make at least 100 forecasts (19 analysts are selected). We notice most forecast errors are smaller than 1, while a few forecasts are associated with very large forecasting errors.[11] Including these outliers would greatly affect our estimation for analysts' mean error. Thus, we only use the first 90% of the sorted forecast errors in this calculation.

[8] https://www.cfraresearch.com/

[9] https://www.thomsonone.com/

[10] Other methods for measuring the forecasting error have been proposed, for example to scale the relative error by stock price. We do not take this approach as stock prices are dynamically changing.

[11] For example, one analyst estimated an EPS for Fiscal Year 2015 of Olin Corporation (OLN) as $1.63, while the actual EPS was $-0.01, a standardized forecast error of 164.

3.2 Predicting Forecasting Error from Text

Our goal is to test whether linguistic differences exist between accurate and inaccurate forecasts, independently of who made the prediction, or how difficult a specific company's earnings might be to predict. To control for these factors, we standardize forecasting errors within analysts (as described in §3.1), and create training/dev/test splits across companies and dates.

Setting. We collect the top K and bottom K predictions and split train, dev and test sets by time range and company. All company names are randomly split into 80% train and 20% evaluation sets. We use predictions for companies in the train group that were made in 2014-2016 as our training data. The dev set and test set consist of predictions for companies in evaluation group made during the years 2017 and 2018, respectively. All hyperparameters are the same as those used in §2.4. When evaluating the classifier's performance, we balance the data for positive and negative categories.

Results. Table 6 shows the performance of our classifier on the test set. We observe our classifiers consistently achieve around 60% accuracy when varying the number of top and bottom forecasts, K.

	K	1000	2000	3000	5000
LR	Bag-of-ngrams	63.9	62.5	61.9	59.3
	Linguistic	56.3	59.2	55.4	55.5
	All above	64.3	64.1	61.5	59.7
Neural	CNN	66.7	67.8	64.7	64.0
	BERT-base	70.8	66.7	65.8	64.4

Table 6: Accuracy (%) for classifying accurate predictions when using top K and bottom K analysts' predictions. We choose n-gram sizes to be 1 and 2. All reported results are on the test set.

3.3 Linguistic Analysis

We present our linguistic analysis in Table 7. The same set of linguistic features in §2.3 is applied to top 4,000 accurate and bottom 4,000 inaccurate analysts notes, excluding readability metric and quotation measure in thinking style metric. Analysts' notes are written in a professional manner, which makes readability metric not applicable. The notes do not contain many quoted text so we exclude quotation measure from the analysis. We also replace the emotion metric with a sentiment lexicon specifically tailored for financial domain and provide our discussions. The Bonferroni-corrected significance level is 0.05/15. We defer discussions to §4 for comparing across different domains. On average, each forecast contains 132.2 tokens with 5.5 sentences.

Financial sentiment. We make use of a lexicon developed by Loughran and Mcdonald (2011), which is specifically designed for financial domain. The ratio of positive and negative sentiment terms to total number of tokens is compared. Our results show that inaccurate forecasts use significantly more negative sentiment terms.

Metric	p	Bonferroni
Parts of Speech		
Cardinal	↑↑	
Noun	↑↑	
Verb	↓↓↓	*
Uncertainty		
% uncertain statements	↓↓	*
Temporal orientation		
Focus on past	↑↑	*
Focus on present & future	↓↓↓	*
Financial sentiment		
Positive	↑↑	
Negative	↓↓↓	*

Table 7: Comparison of various metrics over top 4,000 accurate and bottom 4,000 inaccurate forecasts. Only hypotheses with $p < 0.05$ are reported. See §3.3 for detailed justifications. We follow the same notation as in Table 2, ↑↑↑: $p < 0.001$, ↑↑: $p < 0.01$, ↑: $p < 0.05$.

4 Comparison of Findings Across Domains

In §2 and §3, we analyze the language people use when they make forecasts in geopolitical and financial domains. Specifically, these two sections reveal how language is associated with accuracy both within and across forecasters. In this section, we compare our findings from these domains.

Our studies reveal several shared characteristics of accurate forecasts from a linguistic perspective over geopolitical and financial domains (in Table 2 and Table 7). For example, we notice that skilled forecasters and accurate forecasts more frequently refer to past events. We also notice accurate predictions consistently use more nouns while unskilled forecasters use more verbs.

We also note one main difference between two domains is uncertainty metric: in Good Judgment Open dataset, we observe that more skilled forecast-

ers employ a higher level of uncertainty; while for individual forecasts, less uncertainty seems to be better. It makes us consider the following hypothesis: within each forecaster, people are more likely to be correct when they are more certain about their judgments, while in general skilled forecasters exhibit a higher level of uncertainty. To test this hypothesis, we calculate the Spearman's ρ between the financial analysts' mean forecasting errors and their average portion of uncertain statements. Results show that these two variables are negative correlated with ρ=-0.24, which provides some support for our hypothesis, however the sample size is very small (there are only 19 analysts in the financial dataset). Also, these mean forecasting errors are not standardized by the difficulty of companies analysts are forecasting.

5 Related Work

Many recent studies have analyzed connections between users' language and human attributes (Hovy et al., 2015; Nguyen et al., 2013; Volkova et al., 2014; Tan et al., 2016; Althoff et al., 2014). Son et al. (2018) developed a tool for discourse analysis in social media and found that older individuals and females tend to use more causal explanations. Another example is work by Schwartz et al. (2015), who developed automatic classifiers for temporal orientation and found important differences relating to age, gender in addition to Big Five personality traits. Eichstaedt et al. (2015) showed that language expressed on Twitter can be predictive of community-level psychological correlates, in addition to rates of heart disease. Demszky et al. (2019) analyzed political polarization in social media and Voigt et al. (2017) examined the connections between police officers' politeness and race by analyzing language. A number of studies (De Choudhury et al., 2014; Eichstaedt et al., 2018; Benton et al., 2017; Park et al., 2017) have examined the connection between users' language on social media and depression and alcohol use (Kiciman et al., 2018). Other work has analyzed users' language to study the effect of attributes, such as gender, in online communication (Bamman et al., 2014; Wang and Jurgens, 2018; Voigt et al., 2018). In this work we study the relationship between people's language and their forecasting skill. To the best of our knowledge, this is the first work that presents a computational way of exploring this direction.

Our work is also closely related to prior research on predicting various phenomenon from users' language. For example Tan et al. (2014) study the effect of wording on message propagation, Gillick and Bamman (2018) examine the connection between language used by politicians in campaign speeches and applause and Pérez-Rosas and Mihalcea (2015) explored linguistic differences between truthful and deceptive statements. Ganjigunte Ashok et al. (2013) show linguistic cues drawn from authors' language are strong indicators of the success of their books and Tsur and Rappoport (2009) presented an unsupervised model to analyze the helpfulness of book reviews by analyzing their text.

There have been several studies using data from Good Judgment Open or Good Judgment Project (Mellers et al., 2015b). One recent study examining the language side of this data is Schwartz et al. (2017). Their main goal is to suggest objective metrics as alternatives for subjective ratings when evaluating the quality of recommendations. To achieve this, justifications written by one group are provided as tips to another group. These justifications are then evaluated on their ability to persuade people to update their predictions, leading to *real* benefits that can be measured by objective metrics. Prior work has also studied persuasive language on crowdfunding platforms (Yang et al., 2019). In contrast, our work focuses on directly measuring forecasting skill based on text justifications.

Finally we note that there is a long history of research on financial analysts' forecasting ability (Crichfield et al., 1978; Chopra, 1998; Loh and Mian, 2006). Most work relies on regression models to test if pre-identified factors are correlated with forecasting skill (e.g., Loh and Mian (2006); Call et al. (2009)). Some work has also explored the use of textual information in financial domain. For example, Kogan et al. (2009) present a study of predicting companies' risk by using financial reports. We also note a recent paper on studying financial analysts' decision making process by using text-based features from earning calls (Keith and Stent, 2019). As far as we aware, our work is the first to evaluate analysts' forecasting skill based on their language.

6 Limitations and Future Work

Our experiments demonstrated it is possible to analyze language to estimate people's skill at making predictions about the future. In this section we

highlight several limitations of our study and ethical issues that should be considered before applying our predictive models in a real-world application. In our study, we only considered questions with binary answers; future work might explore questions with multiple-choice outcomes. Prior studies have found that people's forecasting skills can be improved through experience and training (Mellers et al., 2014). Our study does not take this into account as we do not have detailed information on the forecasters' prior experience. Finally, we have not investigated the differences in our model's outputs on different demographic groups (e.g., men versus women), so our models may contain unknown biases and should not be used to make decisions that might affect people's careers.

7 Conclusion

In this work, we presented the first study of connections between people's forecasting skill and language used to justify their predictions. We analyzed people's forecasts in two domains: geopolitical forecasts from an online prediction forum and a corpus of company earning forecasts made by financial analysts. We investigated a number of linguistic metrics that are related to people's cognitive processes while making predictions, including: uncertainty, readability and emotion. Our experimental results support several findings from the psychology literature. For example, we observe that skilled forecasters are more open-minded and exhibit a higher level of uncertainty about future events. We further demonstrated that it is possible to identify skilled forecasters and accurate predictions based solely on language.

Acknowledgments

We would like to thank the anonymous reviewers for providing valuable feedback on an earlier draft of this paper. This material is based in part on research sponsored by the NSF (IIS-1845670), ODNI and IARPA via the BETTER program (2019-19051600004) DARPA via the ARO (W911NF-17-C-0095) in addition to an Amazon Research Award. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, ARO, IARPA, DARPA or the U.S. Government.

References

Heike Adel and Hinrich Schütze. 2017. Exploring different dimensions of attention for uncertainty detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 22–34, Valencia, Spain. Association for Computational Linguistics.

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*.

Russell M. Barefield and Eugene E. Comiskey. 1975. The accuracy of analysts' forecasts of earnings per share. *Journal of Business Research*, 3(3):241–252.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*.

Andrew C. Call, Shuping Chen, and Yen H. Tong. 2009. Are analysts' earnings forecasts more accurate when accompanied by cash flow forecasts? *Review of Accounting Studies*, 14(2):358–391.

Vijay Kumar Chopra. 1998. Why so much error in analysts' earnings forecasts? *Financial Analysts Journal*, 54(6):35–42.

Timothy Crichfield, Thomas Dyckman, and Josef Lakonishok. 1978. An evaluation of security analysts' forecasts. *The Accounting Review*, 53(3):651–668.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. 2018. Constructing a lexicon of English discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365, Melbourne, Australia. Association for Computational Linguistics.

Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota. Association for Computational Linguistics.

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

David N. Dreman and Michael A. Berry. 1995. Analyst forecasting errors and their implications for security analysis. *Financial Analysts Journal*, 51(3):30–41.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Johannes C Eichstaedt, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. 2015. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*.

Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A. Asch, and H. Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 1–12, Uppsala, Sweden. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1764, Seattle, Washington, USA. Association for Computational Linguistics.

Jon Gillick and David Bamman. 2018. Please clap: Modeling applause in campaign speeches. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Emre Kiciman, Scott Counts, and Melissa Gasser. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280, Boulder, Colorado. Association for Computational Linguistics.

Roger K. Loh and G. Mujtaba Mian. 2006. Do accurate earnings forecasts facilitate superior investment recommendations? *Journal of Financial Economics*, 80(2):455 – 483.

Tim Loughran and Bill Mcdonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational linguistics*.

Barbara Mellers, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S Emlen Metz, Lyle Ungar, Michael M Bishop, Michael Horowitz, Ed Merkle, and Philip Tetlock. 2015a. The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, 21.

Barbara Mellers, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, et al. 2015b. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*.

Barbara Mellers, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E. Scott, Don Moore, Pavel Atanasov, Samuel A. Swift, Terry Murray, Eric Stone, and Philip E. Tetlock. 2014. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, 25(5):1106–1115. PMID: 24659192.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. " how old do you think i am?" a study of language and age in twitter. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics.

Gregory Park, H Andrew Schwartz, Maarten Sap, Margaret L Kern, Evan Weingarten, Johannes C Eichstaedt, Jonah Berger, David J Stillwell, Michal Kosinski, Lyle H Ungar, et al. 2017. Living in the past, present, and future: Measuring temporal orientation with language. *Journal of personality*.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*.

Alan G. Sawyer, Juliano Laran, and Jun Xu. 2008. The readability of marketing journals: Are award-winning articles better written? *Journal of Marketing*, 72(1):108–117.

H Andrew Schwartz, Gregory Park, Maarten Sap, Evan Weingarten, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Jonah Berger, Martin Seligman, et al. 2015. Extracting human temporal orientation from facebook language. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

H. Andrew Schwartz, Masoud Rouhizadeh, Michael Bishop, Philip Tetlock, Barbara Mellers, and Lyle Ungar. 2017. Assessing objective recommendation quality through political forecasting. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2348–2357, Copenhagen, Denmark. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Youngseo Son, Nipun Bayas, and H Andrew Schwartz. 2018. Causal explanation analysis on social media. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Sandesh Swamy, Alan Ritter, and Marie-Catherine de Marneffe. 2017. "i have a feeling trump will win.....": Forecasting winners and losers from user predictions on twitter. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1583–1592, Copenhagen, Denmark. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 613–624, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*.

Philip Tetlock. 2005. Expert political judgment: How good is it? how can we know?

Oren Tsur and Ari Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *Third International AAAI Conference on Weblogs and Social Media*.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. Rtgender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Brussels, Belgium. Association for Computational Linguistics.

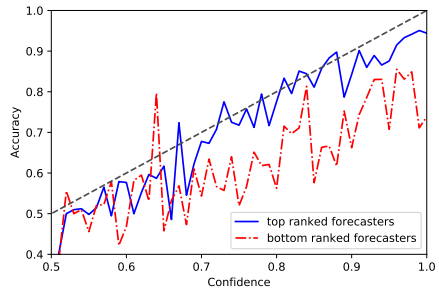
Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Lets make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. 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)*.

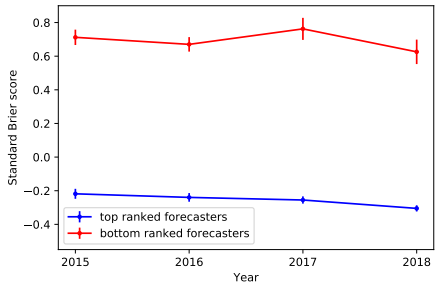
A Additional Experiments on Good Judgment Open Dataset

A.1 Differences Between Top and Bottom Forecasters?

Figure 2 presents calibration curves and averaged standardized Brier scores across years for the top and bottom 500 forecasters. We observe the differences between these two groups are persistent over time. Controlled lab experiments from psychology have also demonstrated that top forecasters ranked by Brier scores consistently have better forecasting performance than bottom forecasters (Mellers et al., 2015a).



(a) Calibration curves by using all forecasts.



(b) Aggregated forecasting performance across years.

Figure 2: Comparison of forecasting skill between the top 500 and bottom 500 forecasters ranked by averaged standardized Brier scores. (a) Calibration curves for each group calculated using all forecasts (with and without justifications). The diagonal dotted line indicates a perfect calibration. (b) Trends of average standardized Brier scores over years. Negative values indicate better forecasting skill.

A.2 Additional Metrics and Examples for Linguistic Analysis

Uncertainty. We present examples of sentences with uncertainty scores from our dataset in Table 9.

Discourse connectives. We further investigate the portion of discourse connectives used between sentences within each group. For this purpose, we use a lexicon developed by Das et al. (2018), which collects connectives from PDTB corpus connective list, RST Signalling Corpus and RST-DT relational indicator list. The lexicon contains 149 English connectives, divided into 4 categories: comparison, contingency, expansion, and temporal.[12] Our results show that skilled forecasters tend to use discourse connectives more frequently compared to unskilled forecasters, which may indicate that they tend to make more coherent arguments.

Thinking style. Analytical thinking score in LIWC (Tausczik and Pennebaker, 2010) ranks the level of a person's thinking skill. A high score correlates with formal, logical, and hierarchical thinking, while low scores are associated with informal, and narrative thinking. As shown in Table 8, good forecasters appear to demonstrate better analytical thinking skills.

Metric	p	Bonferroni
Discourse connectives		
Comparison	$\uparrow\uparrow\uparrow$	$*$
Contingency	$\uparrow\uparrow$	
Expansion	$\uparrow\uparrow$	$*$
Temporal	$\uparrow\uparrow\uparrow$	$*$
Thinking style		
Analytical thinking	$\uparrow\uparrow$	$*$

Table 8: Comparison of various metrics computed over text written by the top 500 and bottom 500 forecasters. p-values are calculated by bootstrap hypothesis test. The number of arrows indicates the level of p-value, while the direction shows the relative relationship between top and bottom forecasters, $\uparrow\uparrow\uparrow$: top group is higher than bottom group with $p < 0.001$, $\uparrow\uparrow$: $p < 0.01$, $\uparrow$: $p < 0.05$. Tests that pass Bonferroni correction are marked by $*$.

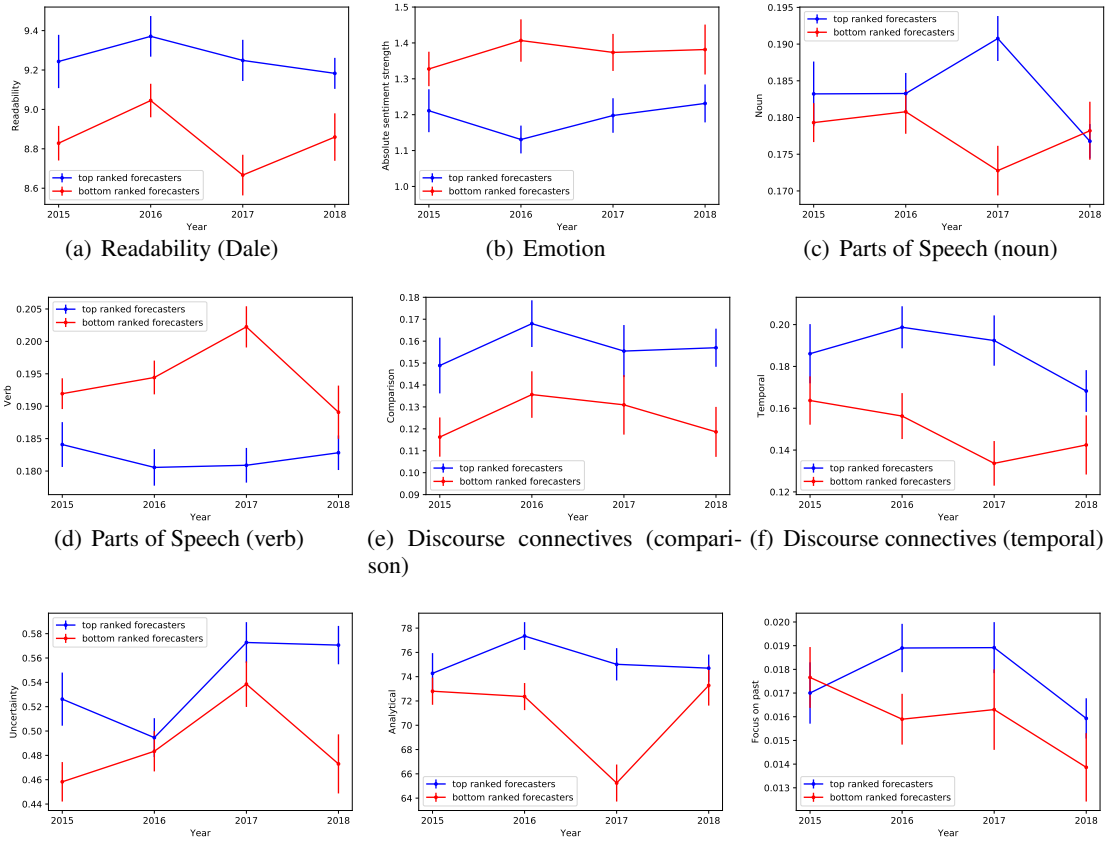
A.3 Linguistic Cues over Time

We are interested in whether our observed linguistic differences are consistent over time. To answer this question, we select the top 500 and bottom 500 forecasters based on their final ranking and evaluate aggregated metrics for the two groups in different years. Our results are shown in Figure 3. We observe the same pattern for all linguistic metrics. For example, skilled forecasters consistently exhibit a higher level of uncertainty and past temporal orientation, and a lower readability compared to unskilled forecasters.

[12] As some connectives are listed under more than one category, we restrict the list to those belonging to one or two categories.

Sentence	Uncert. Score
Merkel is probably least prone to political scandals among the Western leaders and candidates .	1.00
It seems unlikely that the court would transfer the terms of that contract to Uber .	0.99
My assumptions : - Sturgeon will not set a date for indyref2 before the UK elections on June 8 .	0.05
To date , Toyota has distributed only 100 of the 300 Mirais preordered in California ...	0.02

Table 9: Examples of sentences in our dataset with uncertainty scores estimated by the model proposed by Adel and Schütze (2017). A higher uncertainty score indicates a higher level of uncertainty.



(a) Readability (Dale) (b) Emotion (c) Parts of Speech (noun)

(d) Parts of Speech (verb) (e) Discourse connectives (comparison) (f) Discourse connectives (temporal)

(g) Uncertainty (h) Thinking style (analytical score) (i) Temporal orientation (focus on past)

Figure 3: Linguistic features in different years for top 500 and bottom 500 forecasters. The plots show how readability (Dale), emotion, Parts of Speech (noun and verb), discourse connectives (comparison and temporal), uncertainty, thinking style (analytical score), and temporal orientation (focus on past) change in different years. We observe nearly consistent trends for all metrics over time, which indicates that linguistic differences are stable. Error bars represent standard errors.

B Experimental Details on Companies' Earning Forecasts

B.1 Extracting Numerical Forecasts from Text

Not all analysts' notes in our dataset are associated with structured earnings forecasts (in tables). Instead, the analysts' numerical predictions for future earnings are directly reported in the text of their notes, which also contain additional language justifying their predictions. Therefore, our first goal is to extract structured representations of analysts' EPS estimates in a ⟨TIME, VALUE⟩ format. We noticed that analysts have a highly consistent style when writing this section of the report, we therefore use a set of lexico-syntactic patterns to extract the forecasts from text; as described below. We found this approach to have both high precision and high recall.

We randomly sampled 60% of the notes in our dataset for developing patterns. Before generating the rules, we replaced entities indicating time

Sentence	We trim our 12-month target price to $20 from $23 , 10X our '16 EPS estimate of $2.01 -LRB- trimmed today from $2.10 -RRB- .
Pattern	⟨TIME⟩ EPS estimate of ⟨MONEY⟩
Extracted	⟨'16, $2.01⟩
Sentence	We raise '18 and '19 EPS estimates by $4.61 and $5.72 to $19.85 and $25.95 .
Pattern	⟨TIME⟩ and ⟨TIME⟩ EPS estimates ⟨BY-MASK⟩ to ⟨MONEY⟩ and ⟨MONEY⟩
Extracted	⟨'18, $19.85⟩, ⟨'19, $25.95⟩
Sentence	We raise our FY 17 EPS estimate to $3.23 from $2.96 and set FY 18 's at $3.43 .
Pattern	⟨TIME⟩ EPS estimate to ⟨MONEY⟩ ⟨FROM-MASK⟩ and set ⟨TIME⟩ at ⟨MONEY⟩
Extracted	⟨FY 17, $3.23⟩, ⟨FY 18, $3.43⟩

Table 10: Examples of earnings forecasts extracted from analysts' notes. Only sentences mentioning the earnings forecast are shown; the notes also contain additional analysis to justify the forecast. All sentences from notes are used to classify accurate versus inaccurate forecasts as described in §3.2.

and money with special ⟨TIME⟩ and ⟨MONEY⟩ tokens. To evaluate the generalization of our patterns, we randomly sampled 100 sentences containing 136 numerical forecasts from the remaining 40% of notes and manually checked all of them. We estimate that our pattern-based approach extracts numerical forecasts with 0.91 precision and 0.82 recall. Table 10 shows examples of numerical forecasts extracted using our approach. In a few cases we found that an analyst's note can contain more than one forecast. For simplicity, we only consider the earliest forecast that is made within the 2014-2018 time range.