Assessing Objective Recommendation Quality through Political Forecasting

H. Andrew Schwartz[†] **Masoud Rouhizadeh**[†,‡] **Michael Bishop**[‡]
Philip Tetlock[‡] **Barbara Mellers**[‡] **Lyle H. Ungar**[⊙]
[†]Computer Science, Stony Brook University
[‡]Psychology, University of Pennsylvania
[⊙]Computer & Information Science, University of Pennsylvania
has@cs.stonybrook.edu, ungar@cis.upenn.edu

Abstract

Recommendations are often rated for their subjective quality, but few researchers have studied quality in terms of objective utility. We explore quality assessment with respect to both subjective (i.e. users' ratings) and objective (i.e., did it influence? did it improve decisions?) metrics in a *massive online geopolitical forecasting system*, ultimately comparing linguistic characteristics of each quality metric. Using a variety of features, we predict all types of quality with better accuracy than the simple yet strong baseline of recommendation length. For example, more complex sentence constructions, as evidenced by subordinate conjunctions, are characteristic of recommendations leading to objective improvements in forecasting. Our analyses also reveal rater biases; for example, forecasters are subjectively biased in favor of recommendations mentioning business deals and material things, even though such recommendations do not indeed prove any more useful objectively.

1 Introduction

Finding good recommendations is an integral part of a modern information-seeking life – from purchasing products based on reviews to finding answers to baffling questions. Following the tradition of sentiment analysis, many have proposed methods to automatically assess the quality of recommendations or comments based on subjective ratings of their usefulness (Liu et al., 2007; Siersdorfer et al., 2010; Becker et al., 2012; Momeni et al., 2013) or of persuasiveness (Wei et al., 2016). However, information thought to be useful does not always prove so, and subjective ratings may be driven by biases. Reviews which convince you to watch a movie or buy a product do not guarantee that you will enjoy the product, and the most convincing or highest rated answers to questions on sites like Stack Overflow or Yahoo Answers are not always the most accurate.

We explore recommendations in a unique dataset, an online forecasting competition, which offers a rare glimpse into both subjective and objective quality. In this competition, the users (forecasters) had a measurable goal — to forecast the outcomes of geopolitical events — and a need to effectively gather information in order to reach this goal. They viewed recommendations (or "tips") from other forecasters, rated them, and potentially updated their own forecast. This data not only allows us to access what information the forecasters *thought* useful based on their ratings, but also what was objectively useful based on (a) the rate at which forecasters change their prediction after viewing tips and, (b) the average improvement (or decrease) in their prediction accuracy after this change.

We seek to tease out subjective biases by distinguishing the linguistic characteristics of recommendations with high subjective ratings from those of objective utility. Since objectively good recommendations tend to get higher subjective assessments, detecting such differences is non-trivial. Past literature has suggested that subjective quality is well predicted by comment length (Agichtein et al., 2008; Beygelzimer et al., 2015); We seek differences beyond this. We build quality predictive models from linguistic features of recommendations — 1 to 3 word sequences, parts-of-speech, and mentions of concepts from a taxonomy — comparing to surface-level features (length and readability). We then explore the language which distinguishes high quality comments from low quality, controlling for comment length,

and ultimately what distinguishes high subjective quality from objective quality in order to reveal subjective biases.

Contributions. We see the key contribution of this paper, perhaps non-conventional for NLP, as presenting an evidence-based suggestion for the field to consider metrics of objective quality beyond that of subjective ratings. To the best of our knowledge this represents the first study of objective comment quality using randomized experimental data. Specific novel contributions include (a) the development of automated assessments fit to objective outcomes, (b) the identification of linguistic features distinguishing high- from low-quality comments, (c) the use of a new, important, real-world domain for NLP – geopolitical information, and (d), most consequentially, the identification of subjective biases manifested in the comments' text itself.

2 Data Set

Data were collected from the **m**assive **o**nline ge**o**political **f**orecasting system (MOOF) described by Mellers et al. (Mellers et al., 2015; Atanasov et al., 2016). Forecasters completed *tasks* where they indicated the probability of discrete outcomes for specific future geopolitical events around the world. For example, forecasters might be asked to forecast the likelihood of a coup in Venezuela within the next 12 months. Respondents indicated a probability for the event and, after the event resolved, they received a score (Brier, 1950) based on how close their probability reflected the outcome of the event.[1]

Recommendations generated from a another *parallel* forecasting system were presented as "tips". In the parallel system, the recommendation writers were asked "Why did you answer the way you did?" when making forecasts, and were given the option to mark their response as potentially being useful to others. Comments marked useful from the *parallel* system were then presented as tips within the main *MOOF* system, where they were evaluated for their subjective and objective usefulness. Below is an example of one such comment:

> *Predicting the foreign ministers will meet and state something along the lines of "hoping for a summit at the appropriate time." ... Predict this because of the cautious language here and elsewhere: http://www.china.org.cn/wap/2015-03/13/... If wrong, will have time to redress the damage done before any possible summit.*

The above forecaster identifies a reason for their decision, a source of information (a Chinese news website), and even a contingency plan if their reasoning doesn't seem to be planning out. Figure 1 shows a representative screen shot from the *MOOF* system.

We focus on one subjective and two objective quality metrics for these recommendations:
rating: subjective ratings on a 5-point scale by the forecasters in the parallel system.
influence: the rate at which MOOF forecasters update their predictions after reading the comment.
benefit: the mean *change* in MOOF forecaster accuracy resulting from updating their predictions after rating the comment. Because there were numerous confounding factors regarding the magnitude of change (i.e. forecaster quality, time until task resolution), we simply encoded the *change* as a binary indicator of whether it was positive or negative.

For the purposes of this study, we consider both *influence* and *benefit* as assessments of objective utility, though they each capture different aspects of utility (behavioral *influence* of others in the case of influence; and specifically positive influence in the case of *benefit*). The *MOOF* forecasters did not know who wrote the tip or any other information about it; their actions were based on the comment's content and not reputation of the comment author. *MOOF forecasters* were also linked directly to the form to update their forecast from the comment to minimize outside influences between the reading of the comment and prediction update. However, as the forecasters were not in a laboratory, other Web browsing behavior in other tabs could not be controlled.

In order to balance quality score reliability with quantity of recommendation data, we restricted the dataset to recommendations with at least 10 words that received at least 3 ratings. This resulted in $8,498$ comments with *ratings* and *influence* scores. Out of those, $4,317$ comments had at

[1] Data, in the form of quality ratings, POS features, and aggregate concept features is released at http://http://www3.cs.stonybrook.edu/~has/objective_quality/. Due to IRB privacy considerations we are unable to release the full data set.

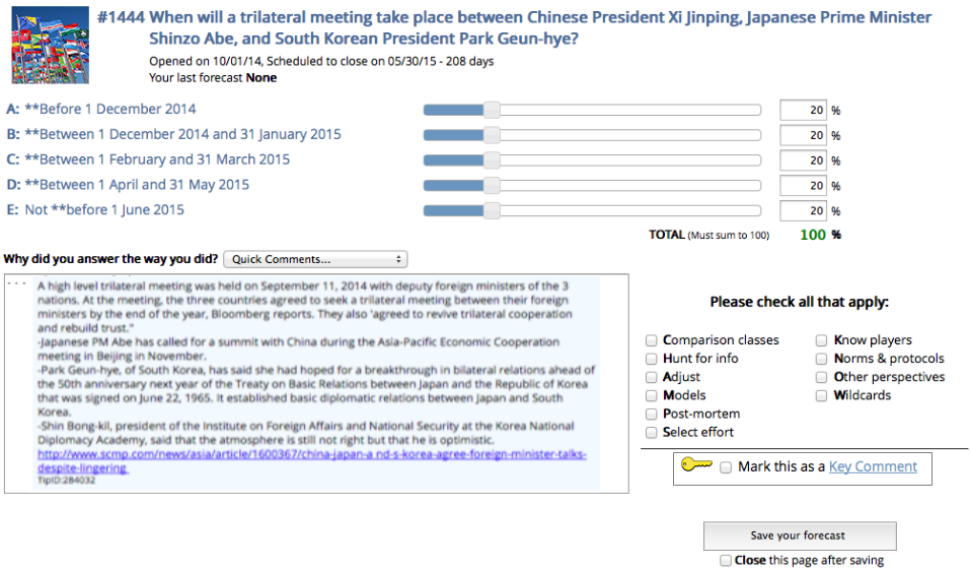

Figure 1: Screenshot of the forecasting system prompting users to justify their forecasts with recommendations. These recommendations are then passed on to the parallel *MOOF* system as "tips" which are rated and sometimes lead forecasters to change their predictions.

rating v influence	***rating v benefit***	***influence v benefit***
.45*	.01	.06*

Table 1: Correlation (Pearson product-moment coefficient) between all subjective and objective quality metrics: rating: forecaster subjective rating, influence: forecaster update rate, benefit: change in forecast accuracy due to update. *significant at $p < .01$.

least one associated change in forecast, and thus had a score for the final metric, *benefit*.

Table 1 shows correlations (Pearson product-moment coefficients) between all subjective and objective quality metrics. *Rating* and *influence* are fairly predictive of each other ($r = 0.45$, $p < 0.01$), and neither correlates particularly strongly with *benefit*. ($r = 0.01$, $p > .01$ for *rating* v *benefit*, and $r = 0.06$, $p < .01$ for *influence* v *benefit*). This implies that while comments rated for subjective utility are also more likely to influence users, they are not necessary influencing them in a positive way. Further, it is possible (and indeed the case) that the characteristics of comments deemed subjectively useful may differ from those that objectively provide benefit. In Section 4 we consider the content that leads to these differences in the metrics.

3 Regression

Our goal is to predict the quality score of a comment from its content over each of the three quality metrics: *rating*, *influence*, and *benefit*. These 3 types of features were chosen in order to account for qualitatively different types of linguistic attributes. To capture linguistic variance at varying resolutions, we use a combination of open-vocabulary and taxonomic linguistic features:

ngrams: *1 to 3 word sequences*. These ngrams were recorded as binary variables indicating whether each ngram appeared in each comment. We limited ngrams to those mentioned in at least 0.1% of all comments.

parts-of-speech: *POS frequences* The Stanford Part-of-Speech tagger was used to identify parts of speech in each comment. The relative frequency of each tag (i.e., the probability of the tag, given the comment) was recorded.

concepts: *Nominal concepts within a hierarchy.* The WordNet noun ontology (Fellbaum, 1998) was used. For each comment, we tracked all of the hypernyms of each noun within the comment. As features per comment, we use the presence of each hypernym concept, limited to those concepts that appeared in at least 0.1% of all comments.

To control for *task*-specific language, features (n-grams) were restricted to those mentioned in at least 50% of forecasting *tasks*. The *n-grams*, intended to capture fine-grained lexical information,

were also restricted to those mentioned in at least 1% of comments, while the *parts-of-speech* and *concepts*, intended to capture more coarse-grained linguistic characteristics, were restricted to those appearing in 5% of comments. These thresholds were chosen such that total number of features was within the same order of magnitude as the number of recommendations with objective scores In the end, there were $1,202$ unique *n-grams*, 32 unique *parts-of-speech*, and 155 unique *concepts*.

3.1 Predictive Modeling

We built predictive models of all three quality metrics based on the linguistic features of the messages. To handle covariance and avoid over-fitting with so many features, we used a series of feature selection and dimensionality reduction techniques fed into a ridge regression to create the models. Specifically, we filtered the features (which were already restricted to those present in at least 50% of the forecasting tasks) to those with a small linear correlation with the outcome, defined as having a family-wise error rate < 60 (Toothaker, 1993). This feature selection was run independently on each type of feature (n-grams, parts-of-speech, and concepts). Similar to correcting for multiple hypothesis tests, family-wise error penalizes groups containing more features (i.e. n-grams) more stringently.

We then used randomized *singular value decomposition* (SVD) (Halko et al., 2011) to reduce the feature set to $^1/_5$ the number of original features. Randomized SVD uses stochastic sampling to more efficiently calculate the principal components (Martinsson et al., 2011).In this context, SVD functions as a type of regularization to reduce variance by removing lesser principal components (and thus helping to avoid overfit). Finally, the resulting dimensions were fit to the given quality metric using L2 penalized linear regression. All feature selection, dimensionality reduction, and L2 parameter tuning were done over a held-out portion of the training set.

3.2 Evaluation

To evaluate our models out-of-sample, we use 10-fold cross-validation over the subjective ratings (*rating*) and update rates (*influence*). In this process, a random selection of 1/10th of the comments are held-out as a test set, while the other 9/10ths are used to train (estimate) the model. This model is then used to predict the quality of the comments in the 1/10th sample and compared to the true quality for those comments (using Pearson correlation in this case). However, many of our scores for change in forecaster accuracy (*benefit*) are based simply on one change and thus quite unreliable. While it is best to include such noisy data when training, it does not provide a very accurate assessment. Therefore, we use dedicated training and test sets, where the test set is a random sample of 500 comments with more than 3 updates and thus a more reliable mean change in forecaster accuracy.

	rating	***influence***	***benefit***
baseline	.59	.24	.02
our model	.76*	.37	.21*

Table 2: Predictive accuracy (out-of-sample Pearson correlation coefficient) of our content-based models across the subjective and objective measures. *baseline*: square-root number of words; *our model*: based on *ngrams*, *parts-of-speech*, and *concepts*. *significant reduction in error over the baseline at $p < .001$.

As a baseline, we use the square-root of the number of words in the comment. This may seem like weak measure of quality, but the history of automatic quality assessment is saturated with findings that length is the best predictor of quality. This holds true for both answers to questions (Agichtein et al., 2008; Surdeanu et al., 2011; Beygelzimer et al., 2015); as well as e-commerce reviews (Cao et al., 2011; Racherla and Friske, 2012). Of course, length is not as shallow as it may seem at first; given no strong incentive for authors to leave long comments, length is likely a proxy for thoroughness of the comment. Still, because we would like to understand the content distinguishing various metrics of quality, we view length as a baseline to move beyond.

Table 2 compares the accuracy of models built on content (*ngrams*, *parts-of-speech*, and *concepts*) to the baseline of length. In all cases, *our models*, based on content, perform significantly better than those based only on length. Further in the case of *benefit* (change in forecaster accuracy), length has virtually no predictive power.

measure	*readability*	*length & readability*
rating	.23	.59
influence	.08	.24

Table 3: Predictive accuracy (out-of-sample Pearson correlation coefficient) of the baseline of length and the baseline of readability and their combination across the subjective and objective measures. *length*: square-root number of words; *readability*: Flesch-Kincaid Scale.

3.3 Relation to Readability

Some previous work used on *readability* measures to evaluate comment quality (e.g. (Agichtein et al., 2008; Hsu et al., 2009)). Readability often refers to the difficulty or complexity scale of a comment, determining the minimum age group able to perceive it. Various readability measures have been suggested in the past such as Flesch-Kincaid Formula (Kincaid et al., 1975), Gunning-Fog Index (Gunning, 1952), and SMOG Grading (Mc Laughlin, 1969). All methods are based on the combination of the count of syllables or words in the comment (as a representation of syntactic complexity), and the number of sentences in the comments (representing the semantic complexity). Such measures of readability are often considered naive and questionable, however, they are commonly used and present a coarse evaluation of the comment's complexity.

We used Flesch-Kincaid scale to measure readability. This scale measures readability by the average number of syllables per word as well as the average number of words per sentence (Doak et al., 1996). We combined the baseline of length and the baseline of readability in order to measure the quality of comments. Table 3 shows results for readability and length plus readability. We find no significant improvement in the baseline and that our model based on content still adds significantly more predictive accuracy.[2]

4 Differential Analysis

The prediction results show promise for automated quality assessment, and that linguistic content can predict quality above-and-beyond length (a proxy for comprehensiveness). Next, we explore what content exactly it is that is predictive of each quality metric, and what content suggests biases distinguishing subjective versus objective quality.

4.1 Method

To identify distinguishing features, we use a series of multivariate linear regressions to find the relationship between each individual linguistic feature and the given quality metric, controlling for comment length – a process known as differential language analysis (Schwartz et al., 2013). Specifically, the individual linguistic feature along with the comment length (logged) are standardized and used as independent variables, and then fit to the standardized form of the given quality metric. The coefficient given is then taken as the standardized effect size of the relationship between that feature and the quality metric, holding length constant (Fox, 1997). In other words, it tells us how much the feature can explain the quality score, beyond what is explainable simply from length.

Using regression to relate variables, although rarely done in Computer Science domains, is standard practice in social and political science (Fox, 1997), though typically not over thousands of potential independent variables as we do here. Therefore, we also correct for multiple hypotheses by using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) over our significance tests.

Differential language analysis allows us to observe and test the unique relationship between each feature and each metric, holding length constant. In addition, we use the difference between standardized metric scores to find the features that distinguish high quality comments in one metric versus another. All methods were implemented within the package, dlatk (Schwartz et al., 2017).

4.2 Quality Comment Features

Figure 2 shows the n-grams most highly correlated with each of our quality metrics. Size indicates correlation strength while color represents overall frequency. Across both subjective ratings and objective update rates, we see discussion of news plays an important role (e.g. "news", "article", and "www.reuters.com"). We do not see the same from comments resulting in positive changes of forecaster accuracy (benefit), which seemed to be distinguished by negation (e.g. "no", "unlikely"). For influence, we see other features indicating probabilistic reasoning (e.g. "%"); the individual

[2] Adding length and readability together to the full model had no benefit.

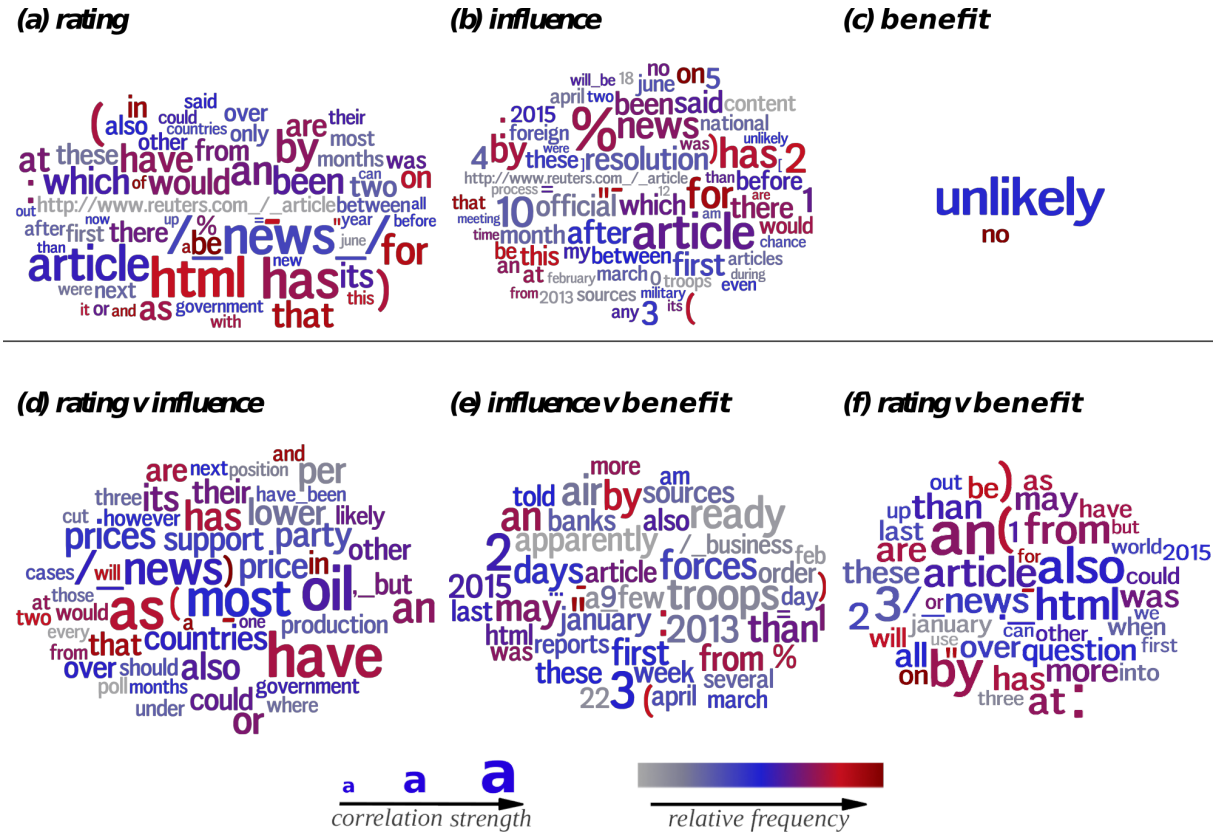

Figure 2: Top: *ngrams* (words and phrases) most distinguishing high quality comments based on (a) subjective ratings, (b) objective forecaster update rates, and (c) objective changes in forecaster accuracy. Bottom: *ngrams* (words and phrases) most distinguishing (d) subjective ratings from objective update rates and (e) objective update rates from objective changes in forecaster accuracy. All correlates are significant at $p < .05$ after a Benjamini-Hochberg false-discovery rate correction.

numbers in influence actually represent numbered lists of signal. These features are more indicative of a comment that convinces one to update their prediction rather than one that highly rated.

We can directly observe the differences in what the metrics capture by looking at the final two visualizations in Figure 2, *rating v influence* and *rating v benefit*. *Rating v influence* tells us which ngrams were predictive of high quality comments that were less likely to result in a forecast update. Discussion of energy topics (e.g. "oil", "prices", "cut", "production") are predictive of comments subjectively rated higher than their update rates would suggest. Further, discussion of dates (e.g. "january", "may", "days", "2015", "2013") seems to predict comments that lead forecasters to update but which do not actually result in better predictions (*influence v benefit*). Discussion of news and articles ("article", "news", "html", "question" appears to predict subjectively top-rated comments from those comments that actually result in better predictions (*influence v benefit*).

Table 4 shows differential language of quality based on part-of-speech. We observe that some patterns from the *ngram* results are generalized. Examples of these patterns include more numbers, parentheses, and quotes in highly rated and influential comments. Other results are somewhat novel, such as the use of more adverbs and subordinate conjunctions (e.g. "though", "since", "whereas") in comments leading to better forecast accuracy, and that both ratings and influence favored quotes.

We notice that including explanation or afterthought (using more parentheses) can predict subjectively high rated comments from the comments leading to updated ratings. The use of quotes and numbers along with mentioning proper nouns can predict influential comments that do not help in better predictions. Including explanation, quotes, and numbers as well as reporting past events seems to predict comment that convinces

	Parts-of-speech				
rating	parentheses	number	line-break	quote	verb, past-tense
influence	line-break	number	parentheses	verb, past-tense	quote
benefit	adverb	sub-conjunction	-	-	-
rating v influence	parentheses	-	-	-	-
influence v benefit	quotes	proper noun	number	-	-
rating v benefit	parentheses	quotes	number	verb, past-tense	-

Table 4: Top: Most distinguishing parts-of-speech correlating with the three metrics of quality: subjective rating, forecaster updates (*influence*), and forecaster accuracy (*benefit*). Bottom: *parts-of-speech* most predictive of differences in quality metric scores (*rating v influence* and *rating v benefit*). All correlates are significant at $p < .05$ after a Benjamini-Hochberg false-discovery rate correction.

one to update their prediction as opposed to helping better forecasting accuracy.

Distinguishing *concepts*, in Table 5 offer a different perspective. Discussions of documents and written material characterize highly rated and influential comments, while changes in accuracy (*benefit*) were characterized by more discussion of abstract attributes or states of being. Highly rated comments were more likely to discuss concepts related to transactions and materials than influential comments, but they are more likely to discuss concepts about creation compared to comments resulting in better predictions.

5 Related Work

While no prior work has focused on recommendation quality in terms of how readers change or improve decisions (i.e. objective metrics), there is an extensive body of literature on the automated analysis of subjective comment quality from which we build. Such work typically uses subjective assessments specific to their application domain (e.g. thumbs up/ thumbs down over YouTube comments, answer ratings in Yahoo Answers, or helpfulness ratings of Amazon product reviews). Below we discuss such work organized into three main categories of subjective comment quality: *comment usefulness*, *the quality of answers in QA platform*, and *comment helpfulness*.

Comment usefulness concerns the acceptance (vs. non-acceptance) of comments by a community (Siersdorfer et al., 2010). Some previous work has focused on the usefulness of YouTube comments. For example, Siersdorfer et. al. (Siersdorfer et al., 2010) have used support vector machines to identify the acceptance of comments by the community in 6 million comments on 67,000 YouTube videos. They showed that community feedback and term weight features can be good predictors of comment acceptance. Other work such as (Momeni et al., 2013) predicted comment usefulness on YouTube and Flickr, and found that comments rated as useful usually include named entities and "insight"-related terms (think, know, consider, etc.), whereas non-useful comments contain emotional and affective expression, and "certainty"-related (always, never, etc.).

Others working on comment quality assessment focus on user-generated answers in social media and QA platforms. (Bian et al., 2008) present a general ranking approach for finding the answers from 1,250 TREC factoid questions containing at least one similar question from Yahoo! Answers. They found that various features including textual (e.g. word overlap, length ratio), and community (e.g. total points, total answers) are important in retrieving factual answers, whereas statistical features (e.g. length, lifetime, popularity) are not very effective. Exploring if additional features could outperform answer length in predicting the best answer, Beygelzimer et al. (2015) considered a wide variety of features including functional, linguistic, questioner and answerer personalization, and "superlative" features, but were unable to overcome the length baseline.

Helpfulness is mainly defined in the context of online reviews and represents the number of users indicating a particular review was helpful. Using structural features like sentence tokens, length,

	Concepts	
rating	*document, written document, papers* – writing that provides information (especially information of an official nature).	*gathering, assemblage* – a group of persons together in one place.
influence	*writing, written material, piece of writing* – the work of a writer; anything expressed in letters of the alphabet (especially when considered from the point of view of style and effect).	*auditory communication* – communication that relies on hearing.
benefit	*attribute* – an abstraction belonging to or characteristic of an entity.	*state* – the way something is with respect to its main attributes.
rating v influence	*transaction, dealing, dealings* – the act of transacting within or between groups (as carrying on commercial activities).	*material, stuff* – the tangible substance that goes into the makeup of a physical object.
rating v benefit	*creation* – an artifact that has been brought into existence by someone.	

Table 5: Top: Select concepts correlating with the three metrics of quality: subjective rating, forecaster updates (*influence*), and forecaster accuracy (*benefit*). Bottom: concepts most predictive of differences in qualtric metric scores (*rating* v *influence* and *rating* v *benefit*). All correlates are significant at $p < .05$ after a Benjamini-Hochberg false-discovery rate correction.

proportion of question sentences along with lexical and syntactic features, Kim et. al. (2006) could achieve rank correlations of up to 0.66 with helpfulness votes of Amazon reviews. Ghose and Ipeirotis (2011) analyzed length, readability, and subjective and objective information on Amazon.com reviews finding that reviews with objective, and highly subjective sentences are rated more helpful. Similar findings were reported by Mudambi and Schuff (2010), finding that review extremity, review depth, and product type affect the perceived helpfulness of the review.

Most studies listed thus far found length of comment to be the dominant predictor, with other features providing minimal benefit. However, a few studies (including our own) have found this baseline can be overcome. For example, Racherla and Friske (Racherla and Friske, 2012) investigated perceived usefulness of consumer reviews on Yelp and found that reputation and expertise were more important than total number of words on perceived usefulness.

All of these prior works focus on assessing subjective aspects of comments (usefulness, quality, and helpfulness); Perhaps the study coming closest in spirit to our own was Ghose et al. Ghose et al. (2007) who quantified quality of reviews by the economic change they produced. However, they still were not dealing with a randomized experiment and so conclusions were correlational and the objective was better sales of the product rather than benefit to the reader (i.e. leading to a better decision).

6 Conclusion

Our results suggest three key findings. First, what one writes in a comment is more important than simply how much one writes; this is true across both subjective and objective outcomes, though length had virtually no predictive ability for improving forecaster accuracy. Second, we found many linguistic features characteristic of quality, many of which seemingly align with attributes of strong forecasters (Mellers et al., 2015). For example, high quality comments contained signals of probabilistic reasoning (e.g. "%", "unlikely", numerical parts-of-speech), inductive reasoning (e.g. justifications with "news" and documents), and cognitive flexibility (e.g. subordinate conjunctions which signal more complex sentence constructions used to relate two independent clauses or ideas).

Most importantly, our results suggest a subjective bias: that what people believe to be useful does not always turn out to be truly useful. Subjective ratings were favorably biased towards comments containing energy-related content (e.g. "oil", "production", "prices"), news articles, and nouns of creation and materials (as opposed to abstractions or attributes).

The implications of identifying subjective biases in comment quality extend to many domains involving comment ratings. Consumers of comments, typically, desire information that ultimately leads to real utility benefits, and this domain is not the only one where objective quality can be obtained: For example, one could: (1) ask consumers of restaurant reviews to indicate if one convinces them to go to the restaurant and then follow up on their experience, (2) consider evaluating research paper quality – reviewer ratings versus citation count (influence), (3) determine whether the answer to the question about how to drive to conserve fuel lead to the reader actually using less gas? A review being convincing or being "liked" may correlate with better outcomes, but it is not equivalent.

7 Acknowledgments

Funding This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center (DoI/NBC) Contract No. D11PC20061. The U.S. government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation thereon. The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/ NBC, or the U.S. government. This work also supported, in part, from the Templeton Religion Trust, grant TRT-0048.

References

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, pages 183–194.

P. Atanasov, W. Chang, S. Patil, B. Mellers, and P Tetlock. 2016. Accountability and adaptive performance: The long-term view. *Under Review* .

Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. 2012. Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, pages 533–542.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 289–300.

Alina Beygelzimer, Ruggiero Cavallo, and Joel Tetreault. 2015. On yahoo answers, long answers are best. In *Proceedings of CrowdML: The ICML 15 Workshop on Crowdsourcing and Machine Learning*.

Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of the 17th international conference on World Wide Web*. ACM, pages 467–476.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78(1):1–3.

Qing Cao, Wenjing Duan, and Qiwei Gan. 2011. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems* 50(2):511–521.

Cecilia Conrath Doak, Leonard G Doak, and Jane H Root. 1996. Teaching patients with low literacy skills. *AJN The American Journal of Nursing* 96(12):16M.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

John Fox. 1997. *Applied regression analysis, linear models, and related methods.*. Sage Publications, Inc.

Anindya Ghose and Panagiotis G Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on* 23(10):1498–1512.

Anindya Ghose, Panagiotis G Ipeirotis, and Arun Sundararajan. 2007. Opinion mining using econometrics: A case study on reputation systems. In *annual meeting-association for computational linguistics*. volume 45, page 416.

Robert Gunning. 1952. {The Technique of Clear Writing} .

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review* 53(2):217–288.

Chiao-Fang Hsu, Elham Khabiri, and James Caverlee. 2009. Ranking comments on the social web. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*. IEEE, volume 4, pages 90–97.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006*

Conference on empirical methods in natural language processing. Association for Computational Linguistics, pages 423–430.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. 2007. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL*. pages 334–342.

Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. 2011. A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis* 30(1):47–68.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading* 12(8):639–646.

Barbara Mellers, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S Emlen Metz, Lyle Ungar, Michael M Bishop, Michael Horowitz, Ed Merkle, and Philip Tetlock. 2015. The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of experimental psychology: applied* 21(1):1.

Elaheh Momeni, Claire Cardie, and Myle Ott. 2013. Properties, prediction, and prevalence of useful user-generated comments for descriptive annotation of social media objects. In *ICWSM-2013: Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Susan M Mudambi and David Schuff. 2010. What makes a helpful review? a study of customer reviews on amazon. com. *MIS quarterly* 34(1):185–200.

Pradeep Racherla and Wesley Friske. 2012. Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications* 11(6):548–559.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9).

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Johannes C Eichstaedt, and Lyle Ungar. 2017. DLATK: Differential Language Analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web*. ACM, pages 891–900.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2011. Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics* 37(2):351–383.

Larry E Toothaker. 1993. *Multiple comparison procedures*. 89. Sage.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in the online forum. In *The 54th Annual Meeting of the Association for Computational Linguistics*. page 195.