Service Quality Markers in Tax Preparation Software

Igor A. Podgorny
Intuit Inc.
San Diego, USA
igor_podgorny@intuit.com

Chris Gielow
Intuit Inc.
San Diego, USA
chris_gielow@intuit.com

Faraz Sharafi
Intuit Inc.
San Diego, USA
faraz_sharafi@intuit.com

ABSTRACT

In anticipation that conversational user interfaces will have a useful role in measuring user satisfaction in financial applications, we present a study of self-help in TurboTax, a tax preparation software supporting U.S. federal and state tax filers. TurboTax customers use self-help to find answers to the tax questions and solve product related problems in TurboTax while preparing their tax returns. The customers have an option to vote experience up or down providing a proxy metrics for an agreement between perceived and expected service in online tax preparation. The collected votes are subjective, noisy and often biased reflecting the relative importance of product and tax related problems, positive and negative emotions associated with expected tax refund and other factors. To address this problem, we present the results of a data analysis based on 647,448 user votes collected in TurboTax in 2017 and then develop models for predicting and attributing the votes. We demonstrate that votes grouped by probabilistic topics of search queries can be used as service quality markers not only for the self-help itself, but for the entire tax preparation experience in TurboTax. In the concluding part of the study, we present redesigned search and voting experiences that optimize collection of the service quality signals. These results can be extended to other online tax and financial applications where self-help and conversational user interfaces are integrated directly into the product.

CCS CONCEPTS

• **Information systems** → **Information System Applications**; *Online Banking* • **Information systems** → **Information Retrieval**; *Query Intent*

KEYWORDS

Service quality, TurboTax, CQA, community question answering, feedbacks, financial services, CUI

ACM Reference format:

I. Podgorny, C. Gielow, and F. Sharafi. 2018. Service Quality Model for Tax Preparation Software. In Proceedings of Workshop on Data Science in Fintech, *London, UK, August 2018 (DSF'18)*, 8 pages.

1 INTRODUCTION

Service quality can be described as a measure of agreement between perceived and expected quality of service [3; 5]. The customers form a (subjective) expectation before encounter with the service and then compare it with the (also subjective) measure of the service received. For the case of e-commerce, service quality can be defined as "extent to which a website facilitates efficient and effective shopping, purchasing, and delivery" [12; 18]. Finally, for the case of online tax preparation, this definition can be reformulated as the taxpayer's perception of the effectiveness of tax preparation software, electronic filing and fulfilling the promise of receiving the expected tax refund. The traditional approach to measuring service quality in tax applications is based on the analysis of comprehensive user surveys using multivariate analysis or structural equation modeling [8]. While this approach does provide good insights on the factors driving customer perception of service quality, it has a limited scalability for the case of offline surveys or may even disrupt the users preparing their tax filing in the case of online surveys embedded into the product. On the other hand, customers who may contribute the most valuable inputs to the service quality data are those who have encountered problems with online tax services and seek self-help or assisted support. A recent shift from human assistance, e.g. phone support or professional tax advice, to artificial intelligence, e.g. chat bots or conversational user interfaces (CUI), may add yet another dimension to evaluation of service quality.

TurboTax Online is a try-before-you-buy software developed by Intuit offering federal and state tax preparation and filing in U.S. and Canada. Besides, TurboTax provides phone and self-support and assists customers with electronic filing of tax return as well as helps tracking the status of their tax refund. TurboTax customers often prefer self-help to assisted measures and are often able to find and apply their solution faster. TurboTax self-help content includes curated FAQs and AnswerXchange, a social Q&A system for generating long-tail help content [9, 10, 11]. As the users step through the TurboTax interview pages, they can search self-help content or ask a question to seek an advice or address a problem encountered in TurboTax. The users then may leave an optional feedback on the self-help experience in the form of up or down votes. Upvote fraction (the ratio of upvotes and all votes) provides a proxy metrics of user satisfaction that is applicable not only to self-help, but also to overall tax preparation and filing experience. In is important that

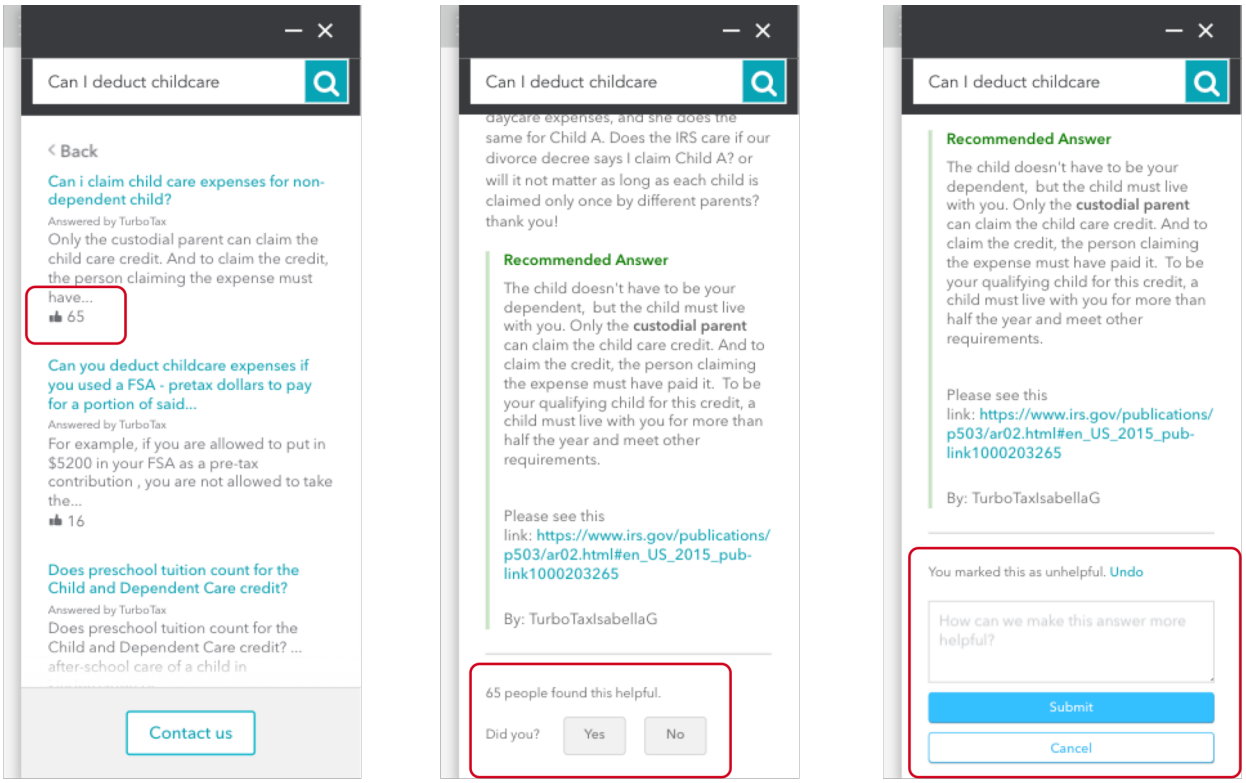

Figure 1: Self-help and voting experience in TurboTax. The current system is based on standard helpfulness rating and is not tailored to ask feedback relevant to customer intent.

Table 1: Model Attributes

Group of Attributes	Attribute	Attribute Type
I. In-product experience	TurboTax screen ID	Categorical
	TurboTax version	Categorical
	User agent	Categorical
	Day of the year	Categorical
II. Query	Text of the query	Tf-idf
	Probabilistic topic	Categorical
	Tax score	Numerical
	Query length	Categorical
	Correct spelling	Binary, tf-idf
	Non-alphanumeric characters	Categorical
III. Search relevancy	Texts of the question	Tf-idf
	Text of the answer	Tf-idf
	Relevancy score	Numerical
	Click page and position	Categorical
	Content type	Categorical
	Self-help interaction time	Categorical

the votes are collected from both customers who successfully filed tax returns and from those who did not. The first goal of this paper is to better understand customer's motivation for voting self-help up or down and, based on this knowledge, attribute the votes to the perceived service quality of the entire tax preparation experience in TurboTax. The second goal is more practical, how to use predictive models of user votes and the vote attribution model to optimize the existing search and voting experience in TurboTax self-help in order to minimize errors and biases in the measured service quality.

2 DATA AND MODEL ATTRIBUTES

2.1 Dataset Description

Shown in Fig. 1 are screenshots illustrating self-help experiences in TurboTax. They include query formulation, rendering of search result snippets and complete documents, and an option for leaving a structured (vote) and optional unstructured (text) feedbacks. In this study, we have employed 647,448 up and down votes collected in 2017 from 541,768 TurboTax Online U.S. customers. The vote engagement rate (i.e. fraction of searching users inside TurboTax Online who left feedback) was 7.4%. The average number of votes per user was 1.195 and the maximum number of votes per user was 35.

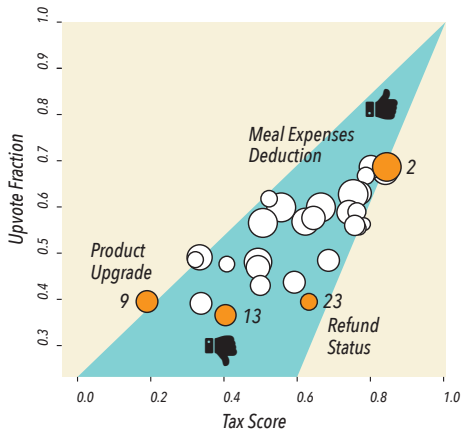

Figure 2: Feedback attribution model and number of votes (area of circle). Topics labels are listed in Appendix A.

The list of model attributes is presented in Table 1. The first group of attributes defines the in-product user experience, i.e. user interactions with TurboTax preceding the self-help request. TurboTax has several thousand interview pages focused on very specific tasks such as selecting right version of the product, importing financial information, entering deduction and printing and/or e-filing the tax return. As a result, the TurboTax interview screen where the search query was entered, page metadata and tags provide information about the type of the problem the user had. The product version (e.g. free vs. paid, Mac vs. Windows) may also contribute to both query intent and users' mental biases in voting content up or down and hence is also included to Group I in Table 1. Next, user agent includes information about user device, web browser and operating system. Finally, day of the year also reflects the type of the user problems in TurboTax reflecting tax filing seasonality factors. Less complex tax returns tend to be filed earlier in the season, while the users with more complex tax situations and those who owe payments to the U.S. Internal Revenue Service (IRS) tend to file closer to the tax filing deadline in April. There is another peak of tax filing activities in October (six months after tax filing deadline) from the customers who filed tax extension in April.

2.2 Query Intent

Query intent is what the users intend when they type search terms [2; 15]. In the case of TurboTax, the query intent can be better understood by attributing it to the specific sub-task in TurboTax that the user is focused on at the time preceding the query formulation. By design, self-help content in TurboTax is a combination of tax and product related topics. Tax related documents are semantically similar to publications by the IRS and by state tax authorities. Product related documents are TurboTax specific and deal with pricing, choice of product version, and software issues such as installation or e-filing. Some documents can be related to both types or can be related to federal or state authorities (e.g. question about status of tax refund) making it more difficult for the customers to decide who may be responsible the encountered problem. The tax vs. product content relationship can be captured by a one-dimensional tax score ranging from 0.0 (e.g. informational software question "Where do I enter W-2 in TurboTax?") to 1.0 (e.g. factoid tax question "What is AMT?") [10]. The users tend to down vote the product related questions more frequently than the tax related ones [9, 10], often irrespective of the quality of answers that can be measured independently [6; 16].

For the more granular view, probabilistic topic models can be applied to discover hidden topics in the self-help documents [1] and to group them by the discovered themes. For this study, the probabilistic topic model has been trained on the AnswerXchange self-help documents using Latent Dirichlet Allocation algorithm. Appendix A presents the complete list of 30 topics with the labels, examples of search queries for each topic, average one-dimensional tax score [9] and average upvote fraction per topic. The correlation (Pearson's r=0.77, Spearman's ρ=0.76) between average upvote fraction and tax score is statistically significant.

Fig. 2 illustrates data shown in Appendix A by plotting upvote fraction per topic as a function of one-dimensional tax score. Topic 2 (close to the top edge of the blue triangle) has the highest upvote fraction (0.69) and tax score (0.84), while topic 9 (close to the left bottom edge) has upvote fraction 0.39 and the lowest tax score (0.18). Topic 23 has a similar upvote fraction (0.39) as topic 9, but a relatively high tax score (0.63). To explain the observed pattern, Appendix A presents query examples for the four topics shown in Fig. 2. Topic 2 is about deductions and hence is a well-defined tax topic. On the contrary, topic 9 about changing the version of TurboTax, i.e. a well-defined product related topic. On the other hand, topic 23 which is about refund status for the already filed tax return has more to do with IRS than with TurboTax, yet it also triggers strong negative emotions. Finally, topic 13 is about rejection of the e-filed tax return and so the users searching for this information may be inclined to attribute their problem to both TurboTax and IRS.

TurboTax users may view the IRS as the tax-authority they need to comply with, and therefore consider it an objective source of truth. For example, these people would tend to "thank" the answer for providing the truth about "Meal expenses deduction." People may view Intuit as a non-authority they can

choose to use, and therefore consider it a subjective source of opinion, arbitrary decisions and poor service. These people would tend to "argue" with the answer providing an opinion about "Refund status" or "Product upgrade". As a result, tax questions tend to get upvotes because those questions are "objective" answers, whereas product questions are "subjective" answers. It might suggest the real reason for the voting behavior, and that product score is just a proxy to detect it.

2.3 Query Writing Style

The length of the query provides a proxy metrics of query intent complexity. Short queries are typically related to general informational or navigational queries (e.g. query "AMT" indicates that user is looking for a definition of "Alternative Minimum Tax"), while longer queries may indicate the need for a detailed instruction or opinion specific to the long-tail tax situation. Shown in Fig. 3 is upvote fraction vs. query length measured in number of characters. The query length data is split into 10 equal deciles and shown in log-scale. As seen from Fig. 3, the upvote fraction decreases with query length. This can be explained by a gradual change in the query intent from tax to product related topics or by user dissatisfaction with the search results for the long-tail search queries for which search engine may be failing to retrieve relevant results or long-tail content may be missing in the AnswerXchange database.

Next, user's ability to formulate good queries can be modeled by language dimensions such as misspellings, usage of special characters, proper or excessive punctuations, use of question words from question taxonomies [13], stop word fraction and the use of vernacular language (e.g. using "dad" and "mom" instead of "parent"). The level of domain knowledge can be derived from the proper use of tax specific terminology (e.g. "filing with my wife" instead of using more formal term of "filing jointly") and tax forms (e.g. "1099-MISC" instead of "miscellaneous 1099"). Finally, the use of personal pronouns provides psycholinguistic information related to user's mental state [17]. Tax forms, personal pronouns, question words, common misspellings and vernacular language words are accounted for by adding them to the tf-idf statistics. The latter was generated with all tokens extracted from the queries including stop words. The special characters and punctuation marks have been included by using binary attributes.

The long-tail misspellings are accounted for using a custom-built character level neural network binary classifier. The classifier was trained with an open source machine learning library PyTorch (http://pytorch.org/) based on 2017 search queries. The misspellings included the ones typically detectable with edit distance metrics (e.g. "recieve" vs. "receive"), incorrect contractions typical for SMS type language (e.g. "im" instead if "I'm"), incorrectly typed tax forms (e.g. "1099 misc" instead of "1099-misc"), and casual abbreviations (e.g. "fed" instead of "federal"). Shown in Fig. 3 is the misspelling score defined as the fraction of search queries that have at least one misspelling. The misspelling fraction increases with query length, but this increase is not gradual.

2.4 Search Relevancy

The search relevancy metrics account for the user experience related to consumption of the rendered search results by measuring semantic similarity between search query and document title. One can expect that when relevant search results would result in an increased customer satisfaction, and vice versa. Shown in Fig. 3 is average relevancy score vs. query length. The score was computed with duplicate content classifier trained based on the set of 5,000 labeled question pairs using Python machine learning library "scikit-learn" [11]. The model predicts class label (0 for a non-duplicate and 1 for duplicate pair) and also the duplicate score (i.e. probability of the question pair to belong to either class ranging from 0.0 to 1.0. As seen from in Fig. 3, the average relevancy score reaches the maximum when the search query is about 30 characters long. The relevancy is low when query is short due to often ambiguous query intent, then it reaches the maximum and then goes down with an increasing query length. The latter may be because the search engine cannot retrieve relevant results for the long-tail customer situations. The remaining attributes in search relevancy group (Table 1) are text of the question and answer, type of the document (i.e. if the document is an official FAQ created by TurboTax or user generated content) and interaction time (i.e. time spent from submitting the query to leaving up or down vote).

It is worth noting that search query intent, query writing style and search relevancy score are not completely independent attributes. For example, the writing style may be affected by the query topic since users may use questions starting from "why" more often for the price related topics ("Why I need to pay if it was free?") and use closed-ended question more often for the tax related topics ("Can I deduct my pet?"). Further, the way user formulates search query may affect the relevancy of the documents retrieved by the search engine and hence user satisfaction as expressed by the vote.

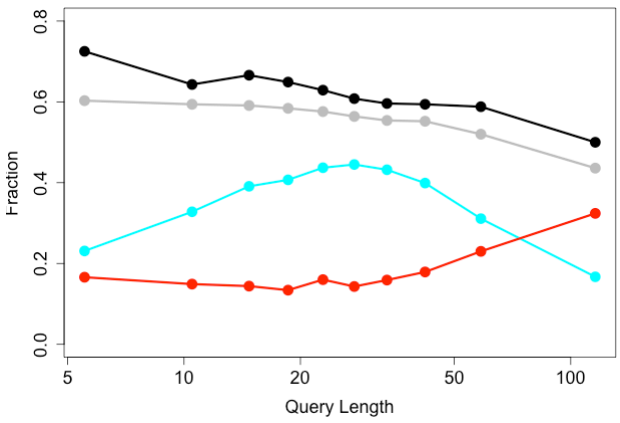

Figure 3: Upvote (grey), tax score (black), relevancy score (cyan) and misspelling fraction (red) vs. query length.

3 VOTE ATTRIBUTION

3.1 Predicting Vote Direction

Three predictive models of user votes have been trained based on the dataset described in Section 2 using "sklearn" logistic regression. The first version has been trained using model attributes from Group I only, i.e. only including in-product experience preceding search query formulation. The second version included model attributes from Groups I and II. The third version of the model has been trained including all model attributes listed in Table 1. The model performance was gradually increasing as more attributes were being added. The outputs of the models are (1) the class label (i.e. 0 or 1, corresponding to down or up vote, respectively) and (2) the class score (i.e. probability of the predicted vote to go down or up, uniformly distributed from 0.0 to 1.0). Shown in Table 2 are performance metrics for the vote predictive model. The metrics include the area under curve (AUC) for the receiver operating characteristic and F1 score computed using the conventional validation with the 33% holdout dataset. Table 2 also includes upvote fraction in the top 5% and bottom 5% segments. The fractions have been computed by ranking test instances by the model score and then retaining bottom and top segments.

Two most important attributes of the model are text of the search query and text of the answer. Recall from Fig. 1 that answer is included in the document that was clicked on. Adding search relevancy categorical attribute to the model does not result in extra information gains despite the fact that the relevancy varies substantially with the query length and hence with upvote fraction as shown in Fig. 3. Similarly, click page and position and time spent on consuming search results do not add extra information gains. In other words, the direction of vote is largely being determined at the query formulation stage and the answer itself serves as an extra factor in refining the original user intent. This confirms our earlier assumption that self-help experience itself does not substantially affect the vote direction and so the vote can serve as a proxy metrics of user experience inside TurboTax.

In addition to the task of better understanding user's motivation in voting up or down for a self-help experience, the predictive models described in this section can play a useful role in redesigning the voting experience. Specifically, identifying the most opinionated users in the top and bottom 5% segments in real time, can help customizing the experience to solicit less biased feedback in the structured or unstructured form as will be discussed in the concluding parts of the study in more detail.

Table 2: Performance Evaluation for the Predictive Models

Groups of Attributes	AUC	F1	Upvote in Top (Bottom) 5% Segments
I	0.654	0.681	0.743 (0.359)
I-II	0.725	0.723	0.820 (0.235)
I-III	0.755	0.741	0.853 (0.187)

3.2 Multivariate Analysis of User Votes

Multivariate analysis such as factor or principal component analysis can be used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved factors or principal components [4]. One application of multivariate analysis can be to examine if the directions of votes in different topics from the same user are correlated and, if so, to apply the correlation statistics to measuring service quality in TurboTax. Towards this goal, we have sampled a subset of the votes restricting it to the users who voted self-help experience in 25% or more topics (i.e. in 8 or more topics). The sample serves as surrogate surveys simulating a situation when the users judge the quality of self-help experience on 3-point scale (e.g. -1, 0 or 1). The number of users in the sample was 79.

The principal component analysis of 79 3-point scale vote vectors has been carried out with R package "psych" [14]. Two components have been retained and then rotated with Promax rotation algorithm. The loadings for the components are shown in Fig. 4 and separated by red and cyan color between the first and second components, respectively. The correlation between components was 0.41. The first observation from the results is that loadings are not correlated with tax score. For example, topics 2 and 9 representing two extreme tax and product related topics shown in Fig. 2 now belong to the same component and are relatively close to each other. The second observation is that topic 23 stays apart from the rest of the topics and is negatively correlated with the second principal component. Recall that topic 23 is related to the status of the tax refund and hence cannot be 100% attributed to TurboTax service quality.

The high degree of correlation between votes recorded in different topics provides a framework for the concept of service quality markers in TurboTax. Specifically, the service quality models can be built from the vote statistics using factor or principal component analysis, provided that sufficient quality data samples are recorded in TurboTax self-help.

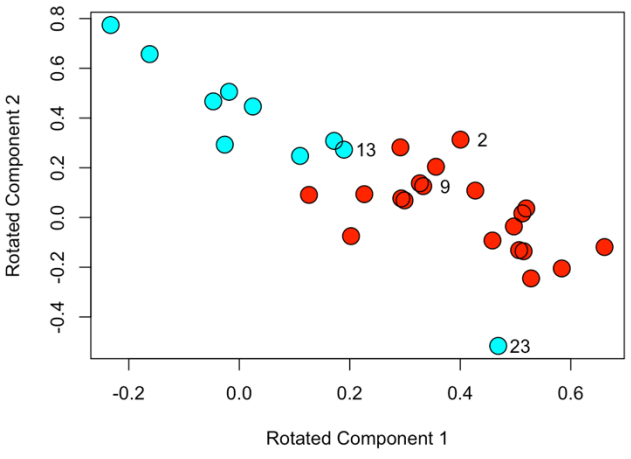

Figure 4: Rotated principal components of vote vectors. The topics labels are listed in Appendix A.

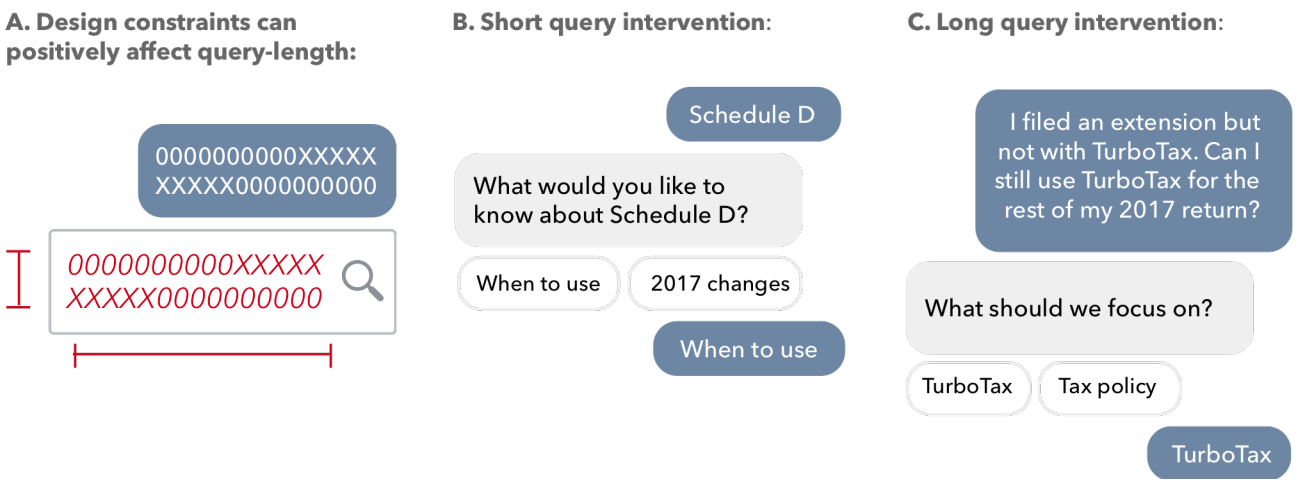

Figure 5: Redesigned query optimization experience. A) An input box optimized for 30 characters positively affects desired query-length. B) Short queries trigger elaboration and assume tax related intent. C) Long queries trigger disambiguation and assume product related intent.

4 OPTIMIZING SERVICE QUALITY SIGNALS

The concluding task of this study is to redesign feedback collection experience to make it better aligned with the concept of service quality markers outlined in the previous session. This task can be completed by reducing the biases in recorded feedbacks and by increasing vote-through rate. The triangular pattern shown in Fig. 2 in blue can be exploited to make feedback collection more specific for the problem the user experienced in TurboTax before entering a search query and hence more accurate feedback attribution. Second, an intelligent user interface can be created by integrating feedback collection experience, predicted user intents (e.g. topic labels) and an estimate for the probability of vote direction predicted in real time. The interface can provide personalized tips for soliciting more meaningful feedbacks similar to the user experience included in AnswerXchange Question Optimizer described in [10]. As mentioned earlier, the binary votes can be enhanced by text feedbacks providing more detail on the specific problem encountered inside TurboTax or TurboTax self-help. Finally, the findings reported earlier in this study can be used to improve the overall self-help experience by delivering more relevant search results and hence by increasing user engagement with self-help. Better self-help experience may also result results in a better vote-though rate.

Shown in Fig. 5 is the redesigned query optimization experience. The query formulation experience is now integrated with the search relevancy data reported in Fig. 3. Specifically, the web interface shown in panel A encourages the user to submit a query that is about 30 characters long based on the premise that the better formulated query will deliver the most relevant search results. An indirect benefit of the longer search query is that the precision of the probabilistic topic model increases with query length resulting in more accuracy detecting the problem the user encountered in TurboTax. The query formulating experiences illustrated in panels B and C trigger data driven query disambiguation and therefore, increase the accuracy of the query attribution to the tax or product related intent.

Shown in Fig. 6 is the redesigned voting optimization experience. The feedback collection now assimilates the real-time information about user intent to make user experience task specific. The assimilated data includes probabilistic topic of the search query and clicked document, tax score and probability of the expected direction of the vote. The real-time data and predictive model outputs trigger task specific feedback collection experiences. For example, ambiguous topic triggers soliciting binary votes, while tax or product specific topics may trigger soliciting categorical or numeric inputs, respectively.

Panel A: If topic is tax/objective related, interface will: (1) Show: objective content-related social-proof (e.g. This is "rated accurate"), (2) Rate: Focus on rating the content, ask content-related follow-ups such as accuracy, sources etc., and (3) Contribute: Encourage contributors to focus on objectivity when authoring question or answer.

Panel B: If topic is ambiguous, interface will: (1) Show: social-proof warning that content is controversial or may conflate topics, (2) Rate: clarify attribution to product or tax, fork experience appropriately, and (3) Contribute: Encourage contributors to clarify attribution and lean into objectivity when authoring question or answer.

Panel C: If topic is product/subjective, interface will: (1) Show: subjective product-related social-proof (e.g. "improvement requests"), (2) Rate: Focus on rating the product, not the content via "help us improve" prompts. Ask product-specific "why" questions. Votes will not count towards content, and (3) Contribute: encourage contributors to categorize for benefit of product improvement, link to relevant product experience.

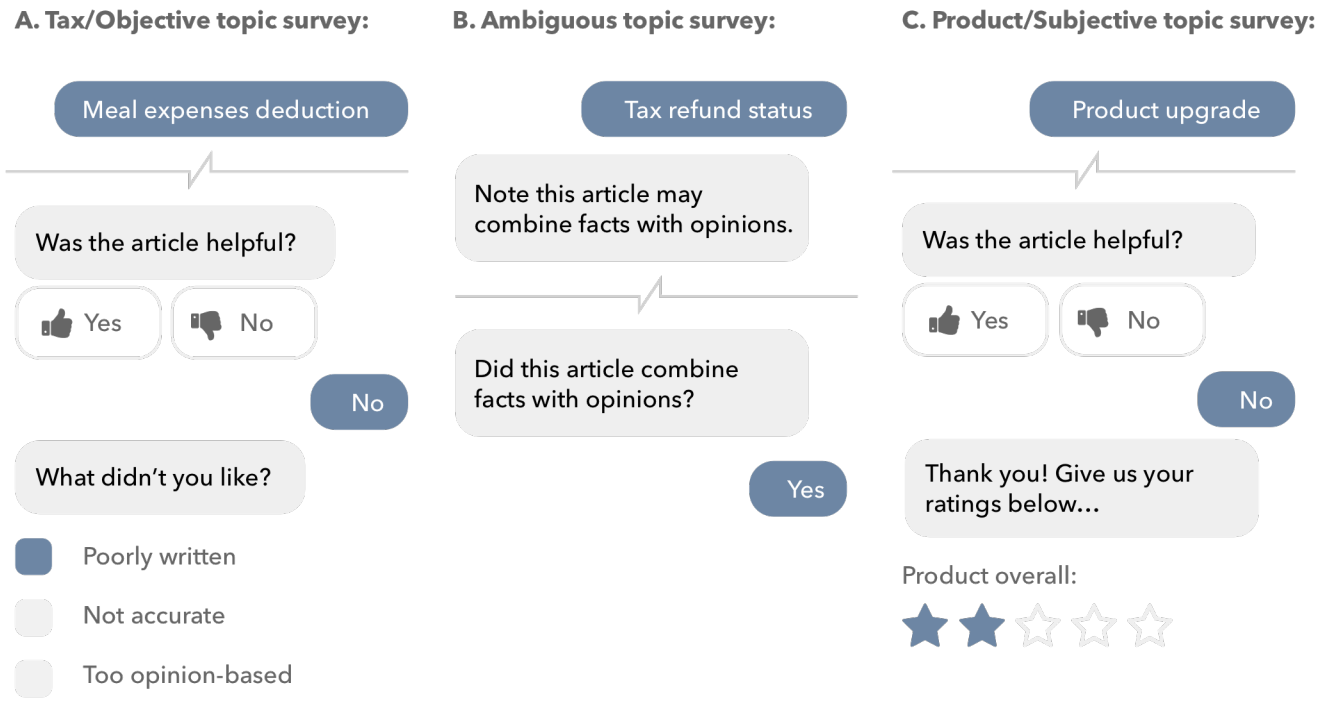

Figure 6: Redesigned voting optimization experience. A) Tax/objective topics tend to receive upvotes, but down votes invite an opportunity to improve article quality with a survey. B) Ambiguous topics present an opportunity to warn and disambiguate. C) Product/subjective topics tend to receive down votes. Down-voting invites an opportunity to focus on product improvement feedback.

5 CONCLUSION

The main goal of this study was to explore the avenues of using votes in TurboTax self-help for measuring service quality of tax preparation and tax filing experience in TurboTax. One of the conclusions was that user votes can be attributed to the search query probabilistic topics and then used as proxy metrics of user satisfaction with the overall TurboTax experience. The time elapsed from encountering the problem in TurboTax, to solving this problem with self-help and to leaving feedback is a matter of minutes making the feedback attribution to the specific part of TurboTax experience more reliable. We also concluded that user votes on self-help experience can serve as TurboTax service quality markers and as inputs to service quality models. Besides measuring service quality, the proposed approach offers several indirect benefits. First, the ability to attribute incoming feedback to the specific elements of TurboTax or self-help experience provides an opportunity to intervene before the feedback has been submitted to the system. Second, the attributes of service quality models can be modified inside the product as part of self-help experience resulting in an improved perception of the overall service quality.

The data analysis presented in this study provides a framework for personalized feedback collection. Specifically, the model trained to predict the incoming vote direction in real time can be integrated with TurboTax self-help in order to offer a customized user experience based on probability of the user to vote the self-help experience up or down. The voting experience can be further optimized for the most opinionated users whose predicted vote directions fall in the top or bottom 5% segments.

The approaches described in this study can be applicable beyond tax preparation experience offered by TurboTax or other providers and beyond the self-help options combining search and social question answering. In addition to tax preparation software, Intuit also offers personal money managing (e.g. Mint) and small business (e.g. QuickBooks) financial solutions. Both Mint and QuickBooks are integrated with self-help and so the service quality concepts outlined in this study can be extended to those products as well. Next, several CUI [7] based solutions are currently being deployed to several TurboTax products providing extra opportunities for measuring service quality. Specifically, CUI can combine both chat bot experience (e.g. bill pay as part of QuickBooks Assistant) and search triggered by user utterances (e.g. fallback interception in CUI). Since user engagement with CUI typically takes longer period of time compared to a shorter and often one-query interaction with self-help search experience, one can anticipate that CUI will have a higher vote-through rate and hence potentially more important role in measuring service quality in financial applications.

A SEARCH QUERY PROBABILISTIC TOPICS

Topic ID	Topic Label	Query Examples	Tax Score	Upvote Rate
1	Forms and schedules	i need help filling out a form 8962	0.5559	0.5971
2	Personal deductions	meal expenses	0.8459	0.6865
3	Health insurance	paid health insurance premiums	0.7665	0.6315
4	Filing status	disabled spouse	0.6453	0.5681
5	Wages	w2 wages	0.6835	0.4836
6	Tax payments	estimated tax paid	0.5091	0.5603
7	Amending return	amend 2013 tax return	0.4913	0.4795
8	Small business	add employee to business	0.7561	0.5587
9	Change product	i accidentally upgraded	0.1892	0.3900
10	Tax residence	living outside the us	0.7690	0.5495
11	Dependents	dependent makes more than 4000	0.8353	0.6839
12	Tax credits	adoption tax credit	0.7853	0.6716
13	Return rejected	why my return rejected	0.4016	0.3591
14	Start over	start a new return	0.4087	0.4741
15	How to enter	where do i enter 1099q	0.6230	0.5689
16	Print and file	did my 2015 taxes get e-filed	0.3365	0.3875
17	Fees	where do i enter brokerage fees	0.5209	0.6204
18	Real estate	home adjusted basis	0.6724	0.6008
19	Mortgages and loans	i received a 1098 for a business loan	0.8415	0.6831
20	Investments	deducting 3000.00 stock loss	0.7584	0.5918
21	AGI	adjusted gross income	0.5915	0.4310
22	Last year return	loading last years files	0.4983	0.4302
23	Refund status	where's my refund	0.6311	0.3892
24	State taxes	charge for state	0.4910	0.4628
25	Cars and trucks	excise tax on car	0.7486	0.6266
26	IRA	sep ira	0.8019	0.6908
27	Bank information	change bank info	0.3336	0.4873
28	Free version	i thought this was free	0.3207	0.4795
29	Reporting income	farm income error	0.7417	0.5904
30	Moving expenses	military move	0.7845	0.5626

ACKNOWLEDGMENTS

We thank anonymous reviewers for valuable comments.

REFERENCES

[1] David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM* 55, 4: 77 - 84.
[2] Andrei Broder. 2002. A taxonomy of web search. In *ACM SIGIR*, 36, 3-10.
[3] Christian Grönroos. 2001. The perceived service quality concept - a mistake?, *Managing Service Quality* 11, 3: 150-152.
[4] Joseph F. Hair, Ronald L. Tatham, Rolph E. Anderson, and William Black. 2006. Multivariate data analysis, 1–758.
[5] Riadh Ladhari. 2008. Alternative measures of service quality: a review. *Managing Service Quality*, 18, 65-86.
[6] Yandong Liu, Jiang Bian, Eugene Agichtein. 2008. Predicting information seeker satisfaction in community quest ion answering. In *Proceedings of SIGIR*, 483 – 490.
[7] Michael McTear, Zoraida Callejas, and David Griol. 2016. The Conversational Interface Talking to Smart Devices. Springer, 1-422.
[8] Bojuwon Mustapha and Siti Normala Bt. Sheikh Obid. 2014. Tax Service Quality: The Mediating Effect of Perceived Ease of Use of the Online Tax System. In *GCBSS-2014*, https://doi.org/10.1016/j.sbspro.2015.01.328.
[9] Igor A. Podgorny, Matthew Cannon, and Todd Goodyear. 2015. Pro-active detection of content quality in TurboTax AnswerXchange. In *Proc. of ACM Conference Companion on CSCW*, 143-146.
[10] Igor A. Podgorny, Chris Gielow, Matthew Cannon, and Todd Goodyear. 2015. Real time detection and intervention of poorly phrased questions. In *CHI'15 Extended Abstracts*, 2205-2210.
[11] Igor A. Podgorny and Chris Gielow. 2018. Semi-automated prevention and curation of duplicate content in social support systems. In *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics, ESIDA '17.*
[12] A. Parasuraman, Valarie A. Zeithaml, and Arvind Malhotra. 2005. E-S-QUAL A Multiple-Item Scale for Assessing Electronic Service Quality. *Journal of Service Research*, 7 (3), 213-33.
[13] Jeffrey A. Pomerantz. 2005. A linguistic analysis of question taxonomies. *J. Am. Soc. Inf. Sci. Technol.*, 56, 7: 715 - 728.
[14] William Revelle. 2017. psych: Procedures for Personality and Psychological Research. Northwestern University, Evanston, https://cran.r-project.org/web/packages=psych. R package version 1.7.8.
[15] Daniel E. Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, 13-19. https://doi.org/10.1145/988672.988675.
[16] Ivan Srba and Mária Bieliková. 2016. A Comprehensive Survey and Classification of Approaches for Community Question Answering. In *TWEB*, 10 (3), 18:1-18:63
[17] Yla R. Tausczik and James W. Pennebaker, 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Language and Social Psychology*, 29, 1: 24 – 54.
[18] V. A. Zeithaml, A. Parasuraman, and A. Malhotra. 2002. Service quality delivery through web sites: a critical review of extant knowledge. *Journal of the Academy of Marketing Science*, 30 (4), 362 – 375.