Analyzing the Linguistic Structure of Question Texts to Characterize Answerability in Quora

Suman Kalyan Maity, Aman Kharb, and Animesh Mukherjee

***Abstract*— Quora is one of the most popular community question & answer (Q&A) sites of recent times. However, with increasing question posts over time and the posts covering a wide range of topics (unlike focused Q&A sites like Stack Overflow), not all of them are getting answered. Measuring *answerability* (i.e., whether a question shall get answered or not) involves collecting expensive human judgment data that can differentiate the characteristics of an answered question from an unanswered (aka *open*) one. Factors to judge if a question would remain open include its subjectivity, openendedness, vagueness, ambiguity, and so on. It is difficult to collect such judgments for thousands of questions, requiring automatic framework to deal the issue of answerability of questions. In this paper, we quantify: 1) *user-level* and 2) *question-level linguistic activities*—that can *nicely correspond to many of the judgment factors noted earlier*, can be *easily measured for each question post* and that *appropriately discriminates an answered question from an unanswered one*. Our central finding is that *the way users use language while writing the question text can be a very effective means to characterize answerability.* This characterization further helps us to predict early if a question remaining unanswered for a specific time period t will eventually be answered or not and achieve an accuracy of 76.26% ($t = 1$ month) and 68.33% ($t = 3$ months). Notably, features representing the *language use patterns* of the users are most discriminative and alone account for an accuracy of 74.18%. We also compare our method with some of the similar works [1], [2] achieving a maximum improvement of ∼39% in terms of accuracy.**

***Index Terms*— Answerability, open questions, Quora, social question & answer (Q&A).**

I. Introduction

The wise man doesn't give the right answers, he poses the right questions.
—Claude Levi-Strauss

FROM a group of small users at the time of its inception in 2009, Quora has evolved in the last few years into one of the largest community driven Q&A sites with diverse user communities. A significantly large proportion of visits to the site is made by users from U.S. (36.7%)[1] followed by users from India (20.5%), U.K. (5%) as of July, 2018. With the help of efficient content moderation/review policies and active in-house review team, efficient Quora bots, this site has emerged into one of the largest and reliable sources of question & answer (Q&A) on the Internet. On Quora, users can post questions, follow questions, share questions, tag them with relevant topics, follow topics, follow users apart from answering, commenting, upvoting/downvoting, and so on. The integrated social structure at the backbone of it and the topical organization of its rich content have made Quora unique with respect to other Q&A sites such as Stack Overflow and Yahoo! Answers, and these are some of the prime reasons behind its popularity in recent times. Quality question posting and getting them answered are the key objectives of any Q&A site. In this paper, we focus on the answerability of questions on Quora, i.e., whether a posted question shall get eventually answered.

In order to facilitate the process of increasing answerable questions, Quora has a topical organization of the list of unanswered questions, referred to as "open questions." These open questions need to be studied separately to understand the reason behind their not being answered or to be precise, are there any characteristic differences between "open" questions and the answered ones. For example, the question "What are the most promising advances in the treatment of traumatic brain injuries?" was posted on Quora on June 23, 2011 and got its first answer after almost 2 years on April 22, 2013. The reason that this question remained open so long might be the hardness of answering it and the lack of visibility and experts in the domain. Therefore, it is important to identify the open questions and take measures based on the types—poor quality questions can be removed from Quora and the good quality questions can be promoted so that they get more visibility and are eventually routed to topical experts for better answers.

A. Motivation

Characterization of the questions based on question quality requires expert human interventions often judging if a question would remain open based on factors like if it is subjective, controversial, openended, vague/imprecise, ill-formed, off-topic, ambiguous, uninteresting, and so on. Collecting judgment data for thousands of question posts are a very expensive process. Therefore, such an experiment can be done only for a small

Manuscript received January 27, 2018; revised June 28, 2018; accepted July 16, 2018. Date of publication August 23, 2018; date of current version September 11, 2018. The work of S. K. Maity was supported in part by the Microsoft Research India Ph.D. Fellowship and in part by the IBM Ph.D. Fellowship during the research. *(Corresponding author: Suman Kalyan Maity.)*

S. K. Maity was with the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India. He is now with the Kellogg School of Management, Northwestern University, Evanston, IL 60208 USA (e-mail: suman.maity@kellogg.northwestern.edu).

A. Kharb was with the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India. He is now with Goldman Sachs, Bengaluru 560071, India.

A. Mukherjee is with the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India.

Digital Object Identifier 10.1109/TCSS.2018.2859964

[1] http://www.alexa.com/siteinfo/quora.com

2329-924X © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

TABLE I

EXAMPLES OF OPEN QUESTIONS WITH RESPECT TO VARIOUS LINGUISTIC ACTIVITIES

Open questions	Linguistic activities	Characteristics
Why Is Facebook And The Ad Agencies That Make Money Off Facebook Blaming Their Very Clients For The Poor Results Experienced On Facebook's Platform? Is That Just A Subterfuge?	high POS tag diversity, lengthy	too controversial, infuses debates/discussions
How does Max Weinberg feel to be the only person to be both on the last episode of Late Night with David Letterman (as the drummer for guest Bruce Springsteen) and the first episode of Late Night with Conan O'Brien (as the house band leader)?	high POS tag diversity, lengthy	ill-formed, not specific, vague, too many queries jumbled up
If a warehouse of physical goods is seized in the US because of illegal activity by the owner and a few customers using it, are the authorities required to return items that are ”innocent” and were collateral damage once the investigation is complete?	high POS tag diversity, high ROUGE-LCS score, lengthy	vague, ill-explained, requires experts to answer
How can Matthew Reilly write such astounding action books? How does he prepare himself while writing a new novel?	low ROUGE-LCS score	Very opinionated, difficult to answer
Spoiler Alert What is your interpretation of the Ichi the Killer's ending.	no topic edits happened	open-ended, directed towards specific set of audiences
1) How expensive it is to get into Big Data Analytics area with simple service offerings? 2) What is the most simple and popular service provided by companies? I would appreciate an early response on the above or pointers to knowledge sources.	lengthy, high POS tag diversity	too many questions, vague/imprecise
#Klout is being used as a type of door policy and a way to discriminate between social-media active consumers, but does it actually work?	no question text edits	too technical and difficult to answer
what are the ways to get a job in it industry as soon as possible?	large no. of text edits	very broad, open-ended, vague and ill-defined

set of questions and it would be practically impossible to scale it up for the entire collection of posts on the Q&A sites. In this paper, we show that appropriate quantification of various *linguistic* activities at the user and the question level can naturally correspond to many of the judgment factors mentioned earlier (see Table I for a collection of examples). These quantities encoding such linguistic activities can be easily measured for each question post and thus helps us to have an alternative mechanism to characterize the answerability on the Q&A sites.

Linguistic activities of the users mean the choice of words, the structure of sentences one adopts to formulate a question, writing answers, and so on. Our hypothesis is that these activities are crucial for popularity of questions/answers/topics, and so on. In our study, we shall focus on these micro-level language usage patterns and study how they can help us to characterize answerability of questions. Quora provides such micro-level language use data in the form of revision history of questions, user logs containing editing history of the users. We collect and analyze large-scale linguistic activities on Quora to characterize the answerability of questions. Although the case study is on Quora, the methods developed are generic and can be easily extended to any other community Q&A sites.

B. Research Objectives and Contributions

In this paper, we analyze a massive data spanning over a period of 4 years consisting of thousands of questions and answers and make the following contributions.

1) We identify and investigate two major kinds of linguistic activities on Quora: user level [e.g., basic activities like posting a question/answer/comment as well as linguistic styles that involves word/char usage, and part-of-speech (POS) tag usage] and question level [e.g., content, topic associations, and edits for a question). Remarkably, many of these activities are found to have a natural correspondence to the qualities that human judges would consider while deciding if a question would remain unanswered (see Table I for a set of motivating examples).
2) We perform an extensive measurement study to show that answerability can be indeed characterized based on the above-mentioned linguistic activities.
3) A central finding is that the language use patterns of the users is one of the most effective mechanisms to characterize answerability.
4) Our characterization further helps us to predict whether a given question remaining unanswered for a specific time period (1 month and 3 months) will remain “open” or not. We achieve **76.26**% accuracy for 1 month and **68.33**% accuracy for 3 months. Note that a question is said to remain open if it has got no answer till the point of investigation of the data. We observe that the features formulated based on *language use patterns* of the users are the most discriminative ones. We, by far outperform two state-of-the-art techniques [1], [2], and achieve a maximum accuracy improvement of ∼39%.

We believe that our contribution[2] is significant because such an early characterization of questions based on language usage patterns may enable the Quora moderators to take necessary steps to reframe the language choice for a question (through appropriate question *promotions* and *edits*) and, thereby, reduce the volume of open questions.

II. RELATED WORK

The research works done in Q&A sites can be broadly classified into three types—user ranking, content quality, and recommendation as discussed in the following.

A. User Ranking

Adamic *et al.* [4] analyze the forum categories in Yahoo! Answers and cluster them according to content characteristics and user interaction pattern. They combine both user attributes and answer characteristics to predict, within a given category, whether a particular answer will be chosen as the best answer by the asker. Li and King [5] propose a mechanism to route questions to a set of answerers based on their past answering performance and users' availability of answering in a time range on Yahoo! Answers. Pal *et al.* [6] show the influence of experts on communities and their evolution dynamics. There have been studies [7], [8] which use the interaction network of users to rank them.

B. Content Quality

Jeon *et al.* [9] propose a framework for predicting answer quality using nontextual features on Naver, a Korean

[2] This is an extension of our work [3] reporting a much more detailed analysis emphasizing various editing activities that distinguish open questions from the closed ones.

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

Q&A sites. Agichtein *et al.* [10] exploit community feedback to identify high quality content on Yahoo! Answers. Shah and Pomerantz [11] use textual features to predict answer quality on Yahoo! Answers. Harper *et al.* [12] investigate the predictors of answer quality through a comparative, controlled field study of user responses. Apart from answer quality, there have been some works on question quality as well [13]–[15]. Liu *et al.* [16] study asker's satisfaction in community-based question answering (CQA). Li *et al.* [13] study question quality in Yahoo! Answers and propose a mutual reinforcement label propagation approach based on question and answer features. Harper *et al.* [17] proposes a framework for classifying factual and conversational questions. Shtok *et al.* [14] reuse the knowledge of the past resolved questions to answer new unresolved similar questions. Correa and Sureka [15] analyze the characteristics and predict "closed" questions on Stack Overflow. Bhat *et al.* [18] propose a framework for predicting the response time (getting the first answer) for a newly posted question. Another study by Correa and Sureka [19] analyze and predict deleted or poor quality questions on Stack Overflow. They found that the properties of deleted questions are different in both topic and content from others and they are poor in quality and off-topic in nature.

Asaduzzaman *et al.* [20] study the problem of how long questions remain unanswered on Stack Overflow based on several heuristics derived from the question such as title length, body, tag similarity, and so on. Dror *et al.* [1] propose a prediction model on how many answers a question shall receive on Yahoo! Answers. They used similar features such as length of title, body, and sentiment in the post to predict answerability. Yang *et al.* [2] analyze and predict unanswered questions on Yahoo Answers. They used several heuristic features such as question length, asker's history, question subjectivity, and question posting time to predict whether the question will be answered or not.

C. Recommendation

There are several works on recommendations in community question answering platforms, e.g., question recommendation, answerer recommendation, and tag recommendation. Wu *et al.* [21] propose an incremental question recommendation framework based on probabilistic latent semantic analysis (PLSA) that considers both users' short- and long-term interests as well as user feedback. Guo *et al.* [22] develop a generative model based on latent Dirichlet allocation (LDA) combining topic-level information about questions and users with word-level information. Qu *et al.* [23] adopt the PLSA model based on user-word aspect for question recommendation. Dror *et al.* [24] represent both the user and question as vectors consisting of multichannel features and cast question recommendation as a classification task. Xu *et al.* [25] propose a dual role model to capture the dual roles played by an user (as an asker and an answerer). Pedro and Karatzoglo [26] propose a supervised probabilistic topic model that extends the LDA model to account for the authorship of Q&A as well as for community feedback. Liu *et al.* [27] propose a group-based recommendation framework to recommend related Q&A documents for knowledge communities of Q&A websites.

Apart from question recommendation, there are also some works that deal with answerer recommendation. Zhou *et al.* [28] present several approaches with language models to represent the expertise of users based on their previous Q&A activities, and then route new questions to appropriate users. Yan and Zhou [29] use inherent semantic relations among asker-question-answerer simultaneously and perform answerer recommendation based on tensor factorization. Kim and Kim [30] propose a bag-of-words-based candidate answer recommendation for questions. Liu *et al.* [31] provide experts by investigating the similarities of question contents, expert profiles, and recommended experts as potential question answerers. Zhang *et al.* [32] combine keyword similarities of questions and users, differences in expertise levels, posting time, and replies to the question to find expert answerer. Pal *et al.* [33] proposed a probabilistic model to identify experts and potential experts according to the question selection preferences of users. Liu *et al.* [34] propose a social context-dependent diverse answerer recommendation framework utilizing various social information.

Tag recommendation is also an important problem in context of CQA. Stanley and Byrne [35] propose a Bayesian probabilistic framework for tag prediction. Nie *et al.* [36] leverage similar questions to suggest tags for new questions. Wu *et al.* [37] not only exploit question similarity but also leverage tag similarity and tag importance using supervised random walk framework.

1) Our Proposal: Our work is different from most of the above-mentioned works in various ways. While answerability and its prediction schemes have already been studied for certain CQAs [1], [2], linguistic styles and editing patterns of users have remained hitherto unexplored. *We performed an extensive study on how linguistic activities of users such as choice of words, structure of question, and editing behavior on a CQA platform help in characterizing answerability of a question.* These novel features prove to be most effective in obtaining better accuracy compared to existing frameworks.

III. Data set Description

We obtained our Quora data set [38] through web-based crawls between June 2014 and August 2014. This crawling exercise has resulted in the accumulation of a massive Q&A data sets spanning over a period of over 4 years starting from January 2010 to May 2014. We followed crawler etiquettes defined in Quora's robots.txt. We used FireWatir, an open-source Ruby library, to control a PhantomJS (Headless Webkit) browser object simulating clicks and scrolls to load the full page. We initiated crawling with 100 questions randomly selected from different topics so that different genre of questions can be covered as stated in [39]. The crawling of the questions follow a breadth first search pattern through the related question links. We obtained 822 040 unique questions across 80 253 different topics with a total of 1 833 125 answers to these questions. For each question, we separately crawl their revision logs that contain different types of edit information for the question and the activity log of the question asker.

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

TABLE II
EXAMPLE OPEN QUESTIONS

Has multimedia become a spammy word?
How are image attachments expected to be used?
How broken is the Facebook bug resolution process?
How big is the US market for parents traveling with children?
How can I contact Oprah to introduce her to my pet memorials?
How will WikiLeaks change how governments treat the communication of and definition of sensitive and secret information?
What data illustrates the concentration of attention across web properties?

TABLE III
PROPORTION OF OPEN QUESTIONS POSTED IN VARIOUS YEARS

Year	% of open questions	Year	% of open questions
2010	2%	2013	10%
2011	13%	2014	35%
2012	8%	2015	32%

IV. ANSWERABILITY OF QUESTIONS

The quality of the question text can play a key role in its answerability it has been also outlined in many previous works. In this section, we show through human judgment experiments that there are indeed differences in question quality between answered and unanswered questions [13]–[15]. To motivate this experiment, in Table II, we note some examples of questions that got posted in 2010 and are still open. Among these long-term open questions, some require domain experts for an answer, some are vague while some are very hard to answer. In the rest of this section, we shall first present some basic properties of open questions followed by a detailed description of the experiments based on human assessment.

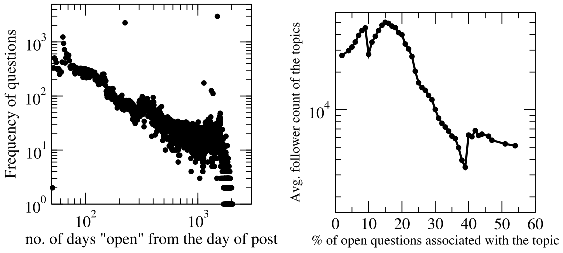
A. How Long Does a Question Remain "Open?"

In this section, we investigate the time duration for which a question remains open. In Table III, we show the proportion of open questions posted in various years. We observe that a significant proportion of questions got posted long back in 2010 and 2011.

To understand better the longevity of open questions, we obtain the time interval between the time of posting of the question and the time of crawl of the question from the set of open questions. We then plot the distribution of this time interval in Fig. 1(a). The figure shows a power-law behavior with a heavy tail. In addition to a heavy tail, there are several peaks around 1500 days. Therefore, though most of the open questions stay "open" few days, there are quite a significant number of questions remaining open long enough requiring special treatment (e.g., disposal or promoting to a domain expert).

B. Open Questions and Their Topics

In this section, we note the topics associated with the open questions. In Fig. 1(b), we show how the percentages of open questions is related to the follower count of the different question topics. We observe that highly followed topics have less percentages of open questions whereas the least followed topics have higher percentages of open questions. In Table IV, we show some example topics associated with higher fraction of open questions. We observe that a significant proportion of these topics technology related including some Quora-specific topics. This observation indicates that Quora (as opposed to forums like Stack Overflow) is not a suitable platform for asking technology related questions. Furthermore, if we rank the topics based on decreasing fraction of open questions then we observe through manual judgment that among the top 20, 50, and 100 topics 45%, 40%, and 48%, respectively, are technology related.



(a) (b)

Fig. 1. (a) Distribution of time duration of open questions (in days). (b) Percentages of open questions with respect to followers of the topics.

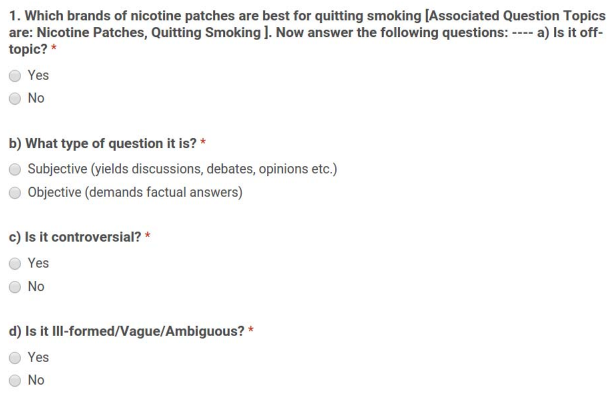
TABLE IV
TOP TOPICS ACCORDING TO FRACTION OF OPEN QUESTIONS (TOTAL NUMBER OF QUESTIONS IN THE TOPICS SHOULD BE AT LEAST 10)

Topics	Fraction of open questions
Charlie Cheever Status Change at Quora (September 2012)	0.81
Yogurt	0.8
College Hacks	0.79
Ron Conway & SV Angel	0.58
Fashion Industry	0.54
Electronic Health and Medical Records	0.53
Interest Graph	0.53
Mozilla CEO Woes (March 2014)	0.53
Seamless (startup)	0.5
Google Street View	0.5
Technology Startups	0.48
Information Technology	0.44
Android Application Development	0.436
Quora Bugs	0.434
Recruitment Stories	0.42

1) Human Assessment of the Questions: In this section, we perform an expensive (and therefore small scale) human assessment experiment to understand if it is actually possible to identify characteristic differences between the question quality of open and answered questions. Note that this experiment is only to establish the proof-of-concept that the question quality is indeed and indicator of the extent of answerability. To this purpose, we conduct an online survey[3] among 25 agreed participants (students, researchers, professors, and technical persons) regularly using Quora with ages ranging between 22 and 34 years. We choose 100 questions; 50 questions each randomly selected from the set of open and answered questions. Each participant is given a set of 12 Quora questions. They are asked to assess the Quora questions on

[3] http://tinyurl.com/hjwuexn

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.



1. Which brands of nicotine patches are best for quitting smoking [Associated Question Topics are: Nicotine Patches, Quitting Smoking]. Now answer the following questions: ---- a) Is it off-topic? *

○ Yes
○ No

b) What type of question it is? *

○ Subjective (yields discussions, debates, opinions etc.)
○ Objective (demands factual answers)

c) Is it controversial? *

○ Yes
○ No

d) Is it Ill-formed/Vague/Ambiguous? *

○ Yes
○ No

Fig. 2. Snapshot of the online survey.

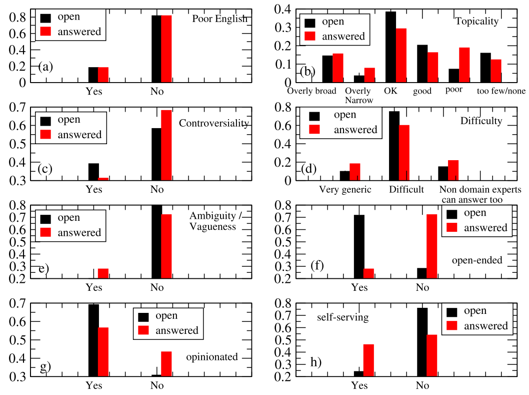


Fig. 3. Human assessment summary. (a) Poor English. (b) Topicality. (c) Controversiality. (d) Difficulty. (e) Ambiguity/vagueness. (f) Openendedness. (g) Opinionated. (h) Self-serving for open questions versus answered questions.

eight different points. These features are as follows—whether the question is written in bad English, number of the topics that are associated with it, controversiality of the question, hardness/difficulty in answering the question, ambiguity/vagueness of the question, openendedness of the question, how opinionated the question is, and how self-serving the question is.

In Fig. 3, we present a summary of the results. Each figure represents the distribution of responses for each of the above-mentioned feature types. Fig. 3(a) shows the extent to which a question is in written in poor English. We observe that this feature cannot discriminate the open questions from the answered questions. In Fig. 3(b), we show the topicality of the questions. We observe that few/no topics are associated with open questions more than the answered ones. From Fig. 3(c), we observe that open questions are more controversial than the answered ones. They are also more difficult to answer than answered questions since they require domain expertise [see Fig. 3(d)]. It is quite evident from Fig. 3(f) that the open questions are more openended in nature than the answered questions. Also they require more opinionated view points than the answered questions [see Fig. 3(g)]. Finally, answered questions see to be more self-serving than open questions [see Fig. 3(h)]. Therefore, indeed, there are certain characteristic differences existing between the quality of open and answered questions. However, scaling up this characterization study for all the questions in our data set is not only highly expensive but also practically unfeasible. Therefore, in the rest of this paper, in order to characterize answerability, we shall use various linguistic activities on Quora as a suitable proxy for expensive question quality assessment.

V. Linguistic Activities on Quora

In this section, we identify various linguistic activities on Quora and propose quantifications of the language usage patterns in this Q&A sites. In particular, we show that there exist significant differences in the linguistic structure of the open and the answered questions. Note that most of the measures that we define are simple, intuitive and can be easily obtained automatically from the data (without manual intervention). Therefore the framework is practical, inexpensive, and highly scalable. In particular, we divide linguistic activities into two broad categories: 1) User-level linguistic activities and 2) Question-level linguistic activities. User-level activities mostly correspond to the activities of the question asker while the question-level activities correspond to the activities of the users other than the question asker. We then attempt to understand the link between these activities and answerability of a question.

A. User-Level Linguistic Activities

User-level linguistic activities refer to various user activities on Quora. These include linguistic style adoption/portrayal while posting/answering questions, editing various aspects of a question, and so on. We quantify the extent of these activities and study the differences they possess for the askers of various questions. In specific, we consider all questions in our data set and analyze the activities of users who posted a question that is open and compare it to users who posted a question that has been answered.

1) Linguistic Styles of the Asker: Content of a question text is important to attract people and make them engage more toward it. The linguistic structure (i.e., the usage of POS tags, the use of out-of-vocabulary (OOV) words, and character usage) one adopts are the key factors for answerability of questions. In this section, we shall discuss the linguistic structure that often represents the writing style of a question asker.

In Fig. 4(a), we observe that askers of open questions generally use more number of words compared to answered questions. To understand the nature of words (standard English words or chatlike words frequently used in social media) used in the text, we compare the words with GNU Aspell dictionary[4] to see whether they are present in the dictionary or not. We observe that both open questions and answered questions follow similar distribution [see Fig. 4(b)]. POS tags are indicators of grammatical aspects of texts. To observe how the POS tags are distributed in the question texts, we define a diversity metric. We use the standard Carnegie Mellon University POS tagger [40] for identifying the POS tags of the constituent words in the question. We define the POS tag

[4] http://aspell.net/

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

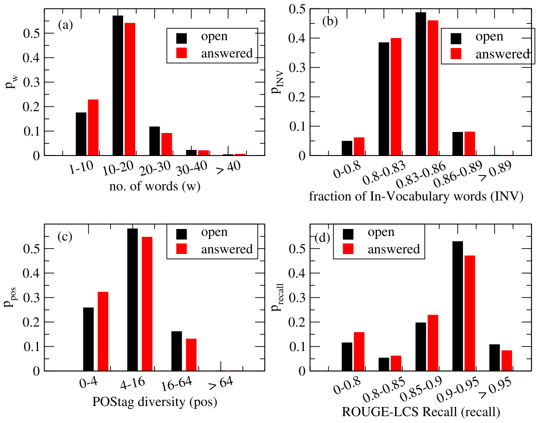


Fig. 4. Comparison of distribution of (a) number of words in the question, (b) fraction of in-vocabulary words, (c) POSDiv, and (d) ROUGE-LCS recall for open questions versus answered question.

diversity (POSDiv) of a question q_i as follows:

$$\mathrm{POSDiv}(q_i) = -\sum_{j \in \mathrm{pos}_{\mathrm{set}}} p_j \times \log(p_j)$$

where p_j is the probability of the jth POS in the set of POS tags. Fig. 4(c) shows that the answered questions have lower POSDiv compared to open questions. Question texts undergo several edits so that their readability and the engagement toward them are enhanced. It is interesting to identify how far such edits can make the question different from the original version of it. To capture this phenomena, we have adopted ROUGE-load current substation (LCS) recall [41] from the domain of text summarization. Higher the recall value, lesser are the changes in the question text. From Fig. 4(d), we observe that open questions tend to have higher recall compared to the answered ones which suggests that they have not gone through much of text editing thus allowing for almost no scope of readability enhancement.

a) Psycholinguistic analysis: The way an individual talks or writes, give us clue to his/her linguistic, emotional, and cognitive states. A question asker's linguistic, emotional, cognitive states are also revealed through the language he/she use in the question text. In order to capture such psycholinguistic aspects of the asker, we use a text analysis application called linguistic inquiry and word count (LIWC) [42] that analyzes various emotional, cognitive, and structural components present in individuals' written texts. LIWC takes a text document as input and outputs a score for the input for each of the LIWC categories such as linguistic (POSs of the words and function words) and psychological categories (social, anger, positive emotion, negative emotion, and sadness) based on the writing style and psychometric properties of the document. In Table V, we perform a comparative analysis of the asker's psycholinguistic state while asking an open question and an answered question.

TABLE V

LIWC ANALYSIS FOR OPEN AND ANSWERED QUESTIONS. ***,**,* MEANS P-VALUE OF SIGNIFICANCE AS <0.01, <0.05, <0.1, RESPECTIVELY

LIWC category	Avg. LIWC score for open questions	Avg. LIWC score for answered questions	Significance Level
Linguistic processes			
Function words	53.41	50.685	***
Pronouns	12.465	8.903	***
Personal pronouns	2.053	3.25	***
1st person singular	0.547	1.25	***
1st person plural	0.226	0.31	***
2nd person	0.889	0.999	***
3rd person singular	0.087	0.151	***
3rd person plural	0.301	0.538	***
Impersonal pronoun	10.411	5.652	***
Articles	9.512	6.894	***
Adverbs	1.881	4.742	***
Conjunctions	2.956	5.554	***
Negation	0.212	0.563	***
Psychological processes			
Social process	5.129	5.779	***
Friends	0.0862	0.114	***
Humans	0.656	0.76	***
Positive Emotion	4.66	3.24	***
Negative Emotion	0.684	0.837	***
Anxiety	0.084	0.108	***
Anger	0.217	0.248	***
Sadness	0.12	0.171	***
Cognitive Processes	10.086	15.388	***
Cause	1.881	5.11	***
Tentative	1.998	3.436	***
Biological Processes	1.148	1.111	**
Body	0.267	0.262	***
Health	0.483	0.477	***
Sexual	0.081	0.078	***

Askers of open questions use more function words, impersonal pronouns, articles on an average whereas asker of answered questions use more personal pronouns, conjunctions and adverbs to describe their questions. Essentially, open questions lack content words compared to answered questions which, in turn, affect the readability of the question. As far as the psychological aspects are concerned, answered question askers tend to use more social, family, human related words on average compared to an open question asker. The open question askers express more positive emotions whereas the answered question asker tend to express more negative emotions in their texts. Also, answered question askers are more emotionally involved and their questions reveal higher usage of anger, sadness, and anxiety related words compared to that of open questions. Open questions, on the other hand, contains more sexual, body, and health related words which might be reasons why they do not attract answers.

2) Other Editing Activities: In Quora, each user has their logs that specify what kind of editing activities in terms of adding an answer, removing inappropriate answer, adding comments/topic, and editing question text is being performed by the user. We perform a comparative study of these edit-level activities among the users who posted a question that is open and the users who posted a question that got answered. In Fig. 5, we show distributions of various such activities. The figures show that there exist prominent differences among users asking open questions and users asking questions that get answered.

Fig. 5(a) shows the distribution of number of answers posted by these users. Askers of open questions usually post lesser number of answers. Since they post low number of answers, their answer removal rate is also low compared to askers of answered questions [see Fig. 5(b)]. However, the askers of open questions usually post larger number of comments compared to the askers of answered questions [see Fig. 5(c)]. In terms of number of question text edits made on the Q&A sites [see Fig. 5(d)], open question askers are found to perform less text edits compared to the answered question askers on average as well as in different ranges. Editing topics to make the question posts more appropriately topically organized is important for receiving answers. We observe that the open

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

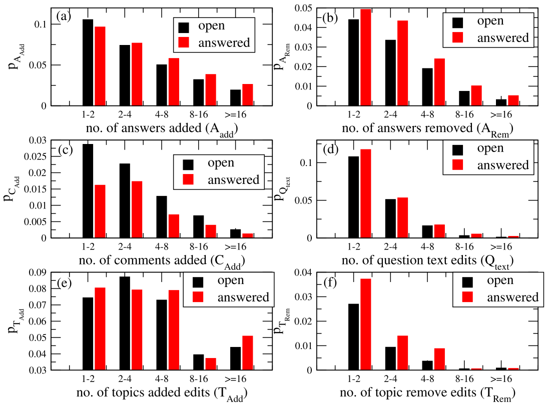


Fig. 5. Distribution of user-level basic editing activities in logarithmic bins for users based on their questions' answerability.

TABLE VI

PROPORTION OF EDITS FOR ANSWERED QUESTIONS AND OPEN QUESTIONS

Edit category	Answered questions	Open questions
Topic added edits	63%	59%
Context topic edits	13%	20%
Question text edits	9%	10%
Question details edits	9%	7%
Topic removed edits	6%	4%

question askers tend to do less number of topic addition edit where there are significant difference in the high zone of distribution [see Fig. 5(e)]. Also, the askers of the answered questions do much more topic removals compared to the askers of open questions [see Fig. 5(f)] on average as well as in different ranges.

B. Question-Level Linguistic Activities

In this section, we shall discuss various activities particular to a question on Quora. We measure activities to a question in terms of the pattern of edits the question underwent. Once a question gets posted on Quora, users can edit its contents, delete topics, add topics, and promote it. There are dedicated Quora moderation and Quora review teams that approve/moderate and also perform such kind of activities themselves. In Quora revision/edit logs for a question, all such edit activities from various users get recorded. These edit logs save various kind of edits. The most relevant edits for our purposes are the following—*context topic edits*, *question details edits*, *question text edits*, *topic added edits*, and *topic removed edits*.

In Table VI, we show the proportion of various kinds of these edits for open questions versus answered questions. We observe that there is a clear distinction in the distribution of various kinds of edits for open questions compared to the answered questions. Though the proportion of topic added edits and topic removed edits are comparable, the proportion of context topic edits for answered questions is ∼1.5 times the proportion of context topic edits in open questions. We next discuss each of these edits in further details.

1) Context Topic Edits: Context topic of a question is the primary topic associated with the question which is needed to remove ambiguities. For example, the question "why was Michael Vick picked for the Pro Bowl?" would be ambiguous due to lack of information regarding the event. Therefore, with context topic "National Football League (NFL) Season 2012–13" the revised question would be "NFL Season 2012–13: Why was Michael Vick picked for the Pro Bowl?."[5] In Fig. 6(a) and (c), we show the distribution of the context topic edits made by the Quora team as well as the other users. We find that context topic edits by Quora team are more for open questions compared to answered questions; in contrast, context topic edits by other (general) Quora users are less for open questions. We also show the temporal context topic edit profiles by Quora team and by other users [see Fig. 6(b) and (d)]. Though there exist differences in the profiles, it does not strongly discriminate the open and the answered questions.

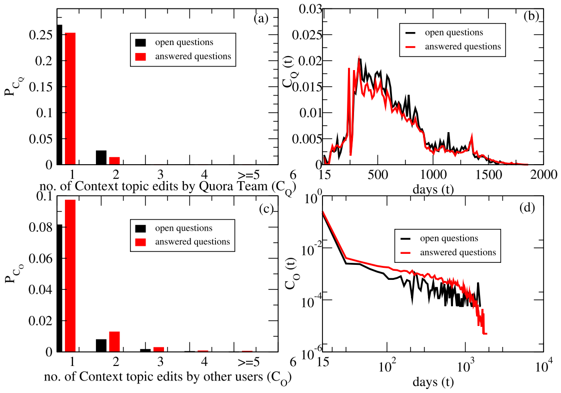


Fig. 6. Distribution of context topic edits by (a) Quora team and (c) other users. The fractions on the *y*-axis are normalized values over all questions in the respective category. Temporal profile of context topic edits by (b) Quora team and (d) other users.

2) Topic Added and Removed Edits: Topics play a vital role in organizing content in Quora. A question appropriately distinguished by topics has higher chances of getting answers than a question that is off topic. From the Quora revision logs for questions, we find out phrases that match "Topic added to question by." We then count the numbers of such cases that are performed by users other than the askers. In Fig. 7(a) and (c), we show the distribution of the topic added edits made by the Quora team and by the other users. We observe that the number of topic added edits by Quora team is significantly higher in case of open questions compared to that of answered questions. For topic added edits made by the other users, the trend is almost reverse. Like "topic added to question by," there are "topic removed from question by" phrases in the revision log of a question which we term as topic removed edits. In Fig. 7(e), we show the distribution of topic removed edits which clearly portrays that the number of topic removal activities performed by the Quora team are far more in answered questions compared to open questions. Also, for other users the topic removal activities are more for answered questions compared to that of open questions [see Fig. 7(g)].

a) Temporal profiles: We also show the temporal profile of the topic added edits by Quora team and by other users

[5] http://tinyurl.com/mzy23lb

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

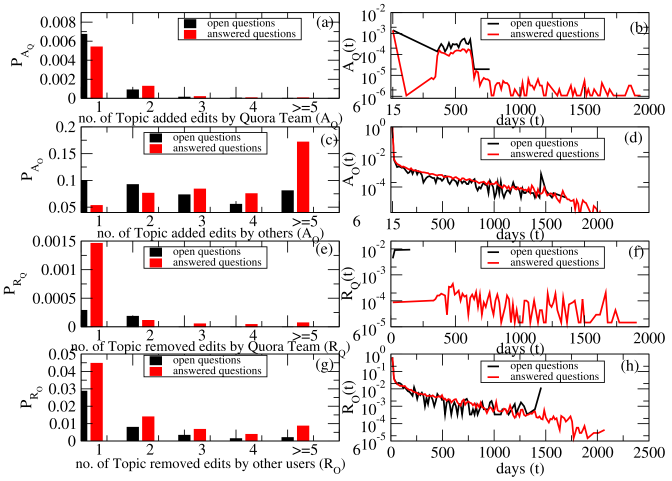


Fig. 7. Distribution of topic added edits by (a) Quora team and (c) other users. Distribution of topic removed edits by (e) Quora team and (f) other users. All the fractions on the *y*-axis are normalized over all questions in their respective categories. Temporal profile of topic added edits by (b) Quora team and (d) other users. Temporal profile of topic removed edits by (f) Quora team and (h) other users.

[see Fig. 7(b) and (d)]. Topic additions are mostly done at the early stage and are prevalent throughout time. There are very few topic removals by Quora team on open questions compared to the steady profile for answered questions [see Fig. 7(f)]. In contrast, the temporal profile for topic removal by other users for open as well as answered questions are quite similar [see Fig. 7 (h)].

3) Question Details and Text Edits: In addition to topics, questions' details and texts are also editable in Quora. The distribution of question detail edits by Quora team shows these activities for open questions are more compared to answered questions [see Fig. 8(a)]. In contrast, the question details edits made by ordinary users are more for answered questions compared to the open questions [see Fig. 8(c)]. The edits in question texts include typos, grammatical errors, and so on. In Fig. 8(e), we observe that the question text edits performed by the Quora team is higher for open questions. On the other hand, question text edits made by other users are significantly higher in case of answered questions compared to that of open questions [Fig. 8(g)].

a) Temporal profiles: The temporal profiles of question details edits by Quora team for open and answered questions are markedly different with most of the edit activities for open questions happening early in time; those for the answered questions happens throughout the lifespan of the question [see Fig. 8(b)]. The temporal profile for question details edit by other users show similar trend for both answered and open questions [see Fig. 8(d)]. Fig. 8(f) and (h) show the temporal variation of question text edits over time by Quora team and by other users, respectively.

b) General users versus Quora team: In summary, we present various activity features both at the user level and at the question level. We observe that there exist clear differences in the activity levels when studied in terms of answerability of questions focusing on open questions and answered questions. A critical observation is that while question text and details

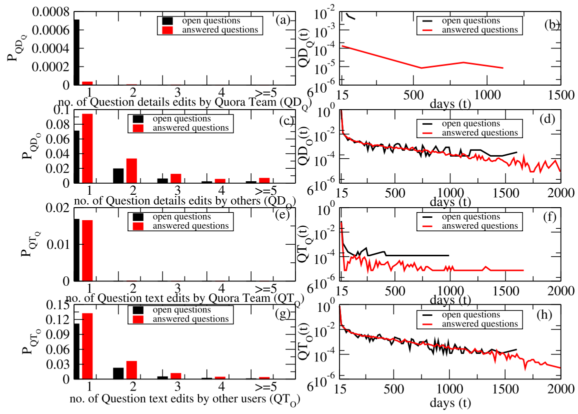


Fig. 8. Distribution of question details edits by (a) Quora team and (c) other users. Distribution of question text edits by (e) Quora team and (f) other users. All the fractions on the *y*-axis are normalized over all questions in their respective categories. Temporal profile of question details edits by (b) Quora team and (d) other users. Temporal profile of question text edits by (f) Quora team and (h) other users.

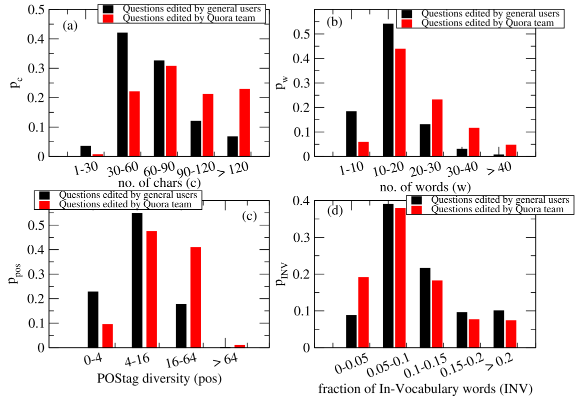


Fig. 9. Comparison of distribution of (a) number of characters in question texts, (b) number of words in the question, (c) POSDiv, and (d) fraction of INV words for questions that are mostly edited by the general Quora users and questions that are mostly edited by the Quora team.

edits for open questions are mostly done by the Quora team, answered question texts are primarily edited by general users. To investigate if there exist some content-level differences between questions edited by the Quora team and those by the general users, we further perform a comparative study of the language usage patterns. *Character usage*—in Fig. 9(a), we show the distribution of the number of characters in question texts. We observe that questions edited by the Quora team have higher fraction of lengthy questions compared to questions edited by general users while questions that are edited by general users are mostly shorter in length. *Word usage*—Fig. 9(b) reveals that general user edited questions have less number of words compared to questions edited by Quora team while there are higher fractions of Quora team edited questions with larger number of words. *POS patterns*—to observe how the POS tags are distributed in the question texts, we use the POSDiv metric defined earlier. Fig. 9(c) shows that the questions edited by Quora team have higher POSDiv compared to questions edited by general users. *Type of words*—to find out the nature of words used in the text, we compare the words with GNU Aspell dictionary[4]. We then

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

TABLE VII

LIST OF FEATURES USED IN THE PREDICTION FRAMEWORK

User-level linguistic styles:	**Question-level linguistic features**
Character length of a question	No. of context topic edits of the question
No. of words in a question	No. of question text edits of the question
Fraction of non-frequent words in a question	No. of question detail edits of the question
No. of function words in a question	No. of times new topics have been added to the question
Fraction of INV words	No. of times existing topics have been removed from the question
Presence of bigrams, trigrams, 4-grams	No. of times topics added by users other than the asker
Part-of-speech tag diversity	No. of topic edits done by the Quora review team
LDA topical diversity	No. of other kinds of edits done by the Quora review team
Psycholinguistic aspects of question texts (LIWC category scores)	Average time interval between edits
ROUGE-LCS recall of modified question text w.r.t original post	No. of question promotions
User-level editing activities:	No. of people to which it has been promoted
Total no. of answers added by the question asker	No. of topics associated with a question
Total no. of answers removed by the question asker	Average depth of the question topics in topic hierarchy
Total no. of question text edits made by the asker	Maximum depth of the question topics in topic hierarchy
Total no. of question details edits made by the asker	Variance of depth of the question topics in topic hierarchy
Total no. of comments made by the asker	Maximum no. of question topics in same levels in topic hierarchy
Total no. of topic added edits made by the asker	No. of connected components of the topic hierarchy graph the question topics belong to
Total no. of topic removed edits made by the asker	Difference in question topics at the time of question post and the topics associated with the question after time period t (observation period)

find out the fraction of in-vocabulary (INV) words. Fig. 9(d) shows that questions edited by Quora team have in general a lower fraction of INV words. Therefore, we observe significant differences existing between these two categories in terms of language structure.

4) Summary of Results: In Table I, we show a collection of examples of open questions to illustrate that many of the above-mentioned quantities based on the linguistic activities described in this section naturally correspond to the factors that human judges consider responsible for a question remaining unanswered. This is one of the prime reasons why these quantities qualify as appropriate indicators of answerability. In summary, we observe that there are sharp differences in the language usage patterns between the open and the answered questions. This is one of the key findings of this paper, and we therefore utilize this observation in Section VI to early predict if a question would remain open.

VI. PREDICTION FEATURES

In this section, we describe the prediction framework in detail. Our goal is to predict whether a given question after a time period t will be answered or not. For the task of prediction, we learn two major types of features (a list of all features are presented in Table VII) inspired by the detailed analysis in Section V as follows.

1) *User-Level Linguistic Features:* these set of features include user-level (question asker) activities toward a question.
2) *Question-Level Linguistic Features:* these features are related to various level of activities (mostly editing) tied to a question.

A. User-Level Linguistic Activities of the Question Asker

As observed in Section V, there exist differences among the askers of open questions and answered questions. Therefore, we have used those activities as various features for the prediction model.

1) User-Level Linguistic Styles: The content and way of posing a question is important to attract answers. We have observed in the previous section that these linguistic as well as psycholinguistic aspects of the question asker are discriminatory factors between answered questions and open questions. For the prediction, we use the following features.

1) Character length of a question, number of words in a question, fraction of nonfrequent words in a question, and number of function words in a question.
2) INV and OOV words in question text—For each question text, we check whether a word appearing in the question text, is an INV word or OOV word by comparing with GNU Aspell dictionary. We then consider the fraction of INV words as a feature of our model.
3) Presence of n-grams of the question content in English texts—We search for 2, 3, 4 grams of the words from the question text in the corpus of 1 million contemporary American English words.[6] We use the presence of bigrams, trigrams, 4 gm each as features for the prediction model.
4) POSDiv of the words in the question. We also use the difference in POSDiv between the initial question text and the question text after time period t (observation period) as a feature to the model.
5) Distribution of LDA topics obtained from question texts—For topic discovery from the question corpus, we adopt LDA [43] model, a renowned generative probabilistic model for discovery of latent topics in a document. For a question q_i, we consider all the words in that question as a document for the LDA model. We set the number of topics as $K = 10, 20, 30$ and find out $p(\text{topic}_k|D_i)$ for a document D_i containing all the words of the ith question. Each of these $p(\text{topic}_k|D_i)$ for $k = 1 \ldots K$ act as a feature of the model.
6) LDA topical diversity—We also compute LDA topical diversity (TopicDiv) of a question (q_i) from the document–topic distributions obtained above as follows and use this metric as a feature

$$\text{TopicDiv}(q_i) = -\sum_{k=1}^{K} p(\text{topic}_k|D_i) \times \log p(\text{topic}_k|D_i).$$

[6] http://www.ngrams.info/samples_coca1.asp

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

7) Psycholinguistic aspects of question texts—We consider the LIWC scores from the different categories as features for the model.
8) ROUGE-LCS recall of the question text at the end of the observation period of the prediction with reference to the original question text posted by the asker.

2) User-Level Editing Activities:

1) Total number of: 1) answers added by the question asker; 2) answers removed by the question asker; 3) question text edits made by the asker; 4) question details edits made by the asker; 5) comments made by the asker; 6) topic added edits made by the asker; and 7) topic removed edits made by the asker.

B. Question-Level Linguistic Features

We have considered several question-level linguistic features which we have noted as follows.

1) *Question Edits:*
   a) Number of context topic edits of the question.
   b) Number of question text edits of the question.
   c) Number of question detail edits of the question.
   d) Number of times new topics have been added to the question.
   e) Number of times existing topics have been removed from the question.
   f) Number of times topics added by users other than the asker.
   g) Number of topic edits done by the Quora review team.
   h) Number of other kinds of edits done by the Quora review team.
   i) Average time interval between edits as a feature.
2) *Question Promotions:* A question can be promoted to various users for increased visibility. We use the number of question promotions as well as the number of people to which it has been promoted as features for our model.
3) *Features Based on Topic Hierarchy:* Question topics play an important role in organizing the question and better the organization a question has, better is its chance of exposure to the experts. In Quora, topics are hierarchically organized via parent-child relationship in the form of forests with a core tree. We have separately crawled the topic hierarchy of almost all the topics available in Quora. We devise various features related to this topical organization in Quora as follows.
   a) Number of topics associated with a question.
   b) Average depth, maximum depth, and variance of depth of the question topics in the topic hierarchy.
   c) Maximum number of question topics belonging to the same level in the topic hierarchy tree.
   d) Number of connected components of the topic hierarchy graph the question topics belong to.
   e) Difference in question topics at the time of question post and the topics associated with the question after time period t (observation period).

VII. Prediction Model

In this section, we shall discuss the prediction model. We perform our predictions at two time points—$t = 1$ month after a question is posted and $t = 3$ months after a question is posted. In other words, for the first (second) case any question that remains open at the end of 1 month (3 months) is labeled as "open" in the ground-truth data, else it is labeled "answered." Furthermore, in the first (second) case, all the features described in the previous section are calculated only using the 1 month (3 months) observation data. Restricting the computation of the features to the observation period only ensures that there is strictly no scope for data leakage.

In the prediction task, we remove all the questions posted by the anonymous users and only consider those questions which are posted by the nonanonymous users. This operation significantly reduces the size of the data set. For the first case, at the end of the 1 month observation period the total number of questions that got answered (after remaining open for 1 month) are ∼4000. We consider a matching set of another 4000 questions that remained open at the end of the 1 month thus making a balanced set of 8000 questions for prediction. Similarly, in the second case, the total number of questions that got answered (after remaining open for 3 months) are ∼2000. We consider a matching set of another 2000 questions that remained open at the end of the 3 months, thus making a balanced set of 4000 questions for prediction.

A. Choice of Classifiers

We have used several classifiers—support vector machine (SVM), logistic regression (LR), and random forest (RF) classifier available in Weka Toolkit [44]. We choose the three classifiers for their diversity since they are known to be able to solve a vast range of different types of classification problems. Each of these classifiers represent different schools of thoughts and have their own set of strengths and advantages.[7] LR, for instance, is a well-behaved classification algorithm that can be trained as long as one expects the features to be roughly linear and the problem to be linearly separable. It is also pretty robust to noise and one can avoid overfitting and even do feature selection by using L2 or L1 regularization. SVMs use a different loss functions (Hinge) from LR. They are also interpreted differently (maximum margin). However, in practice, an SVM with a linear kernel is not very different from a LR. Another advantage of using SVM is if the problem is found to be not linearly separable. In that scenario, one can use an SVM with a nonlinear kernel (e.g., radial basis function). Also, SVMs are suitable for high-dimensional feature space. SVMs have been reported to work better for text classification. Since, we have a large number of features (∼150) and various text-based features, we have used SVM. The advantage of RF algorithm is that it does not expect linear features or even features that interact linearly. The other main advantage is that because of how they are constructed (using bagging or boosting) these

[7] https://bit.ly/2LkuSf0

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII

PERFORMANCE OF VARIOUS METHODS FOR DIFFERENT TOPIC SELECTIONS FOR LDA FEATURE WITH NUMBER OF TOPICS ($K = 10, 20, 30$) USING SVM CLASSIFIER

Time-period	Method	K	Accuracy	Precision	Recall	F-Score	ROC Area
t = 1 month	Our Method	10	75.21%	0.752	0.752	0.752	0.752
		20	**76.26%**	0.763	0.763	0.763	0.762
		30	76.11%	0.761	0.761	0.761	0.761
	Yang et al. (2011)		57.4%	0.534	0.734	0.618	0.543
	Dror et al. (2013)		55%	0.543	0.73	0.624	0.554
t = 3 months	Our Method	10	64.3%	0.643	0.643	0.643	0.643
		20	**68.33%**	0.684	0.683	0.683	0.683
		30	66.8%	0.669	0.669	0.669	0.669
	Yang et al. (2011)		59.3%	0.587	0.64	0.613	0.592
	Dror et al. (2013)		59.8%	0.596	0.628	0.612	0.598

TABLE IX

TOP 20 PREDICTIVE FEATURES AND THEIR DISCRIMINATIVE POWER FOR $K = 20$ AND OBSERVATION TIME PERIOD $t = 1$ MONTH

χ^2 Value	Rank	Feature
1521.94647	1	Impersonal pronouns (LIWC)
955.16735	2	Pronouns (LIWC)
897.49848	3	Causation (LIWC)
836.10745	4	Adverb (LIWC)
727.80798	5	Cognitive Processes (LIWC)
559.56933	6	Conjunction (LIWC)
423.85513	7	Articles (LIWC)
244.35396	8	Tentative (LIWC)
203.5993	9	Positive Emotion (LIWC)
175.84601	10	Function words (LIWC)
169.76226	11	Affective processes (LIWC)
152.08032	12	Achievements (LIWC)
145.01633	13	Personal pronoun (LIWC)
134.30715	14	ROUGE-LCS recall
111.43357	15	1st person singular (LIWC)
105.94163	16	Max. topic in same level of the topic hierarchy
100.34834	17	No. of connected components in topic hierarchy
98.1692	18	Exclusive (LIWC)
89.06608	19	Topic edits
87.2797	20	Other topic added edit

algorithms handle very well high-dimensional spaces as well as large number of training examples.

VIII. RESULTS

In this section, we discuss the results obtained. We perform a tenfold cross validation with SVM classifier and achieve **76.26**% accuracy with high average precision and recall rates for $t = 1$ month and **68.33**% for $t = 3$ months (see Table VIII for details). LR and RF classifiers yield very similar classification performance (at $t = 1$ month, accuracy of **75.11**% and **74.42**% for LR and RF, respectively) although SVM performs best among them. We consider linear kernel and the cost parameter (C) as 1. SVM results in a slightly better performance due to its inherent ability to deal with large number of features, many of which are textual, as is the case for our work. Consequently, we report the performance of the SVM classifier in detail. We observe that the number of topics (K) of LDA does not have a significant effect on the classification results. For $K = 20$, we achieve the best accuracy, average precision, recall, and the area under the receiver operating characteristic (ROC) curve. Note that as time progresses, it becomes increasingly difficult to distinguish between an open question and a question that will get answered. However, thanks to the rich set of features, even with 3 months observation period, we are able to obtain a decent prediction accuracy.

There are a very few early works [1], [2] regarding answerability of the questions. We adopt these works in the context of Quora as baseline methods and compare our methodology with them. Our method outperforms the above-mentioned baselines significantly. We achieve 32.8% $((76.26 - 57.4/57.4) \times 100\%)$ improvement in terms of accuracy over Yang *et al.*'s method [2] and 38.65% $((76.26 - 55/55) \times 100\%)$ improvement over Dror *et al.*'s method [1] for $t = 1$ month. Note that our method achieves similar improvement over other metrics such as precision, ROC area (see Table VIII for further details). Also our method performs best for prediction on shorter time periods than the baselines. The reason for this improved accuracy lies in the selection of the features.

A. Reason for Superior Performance

As we shall see later in this section that the linguistic features—specifically, the LIWC features are the most prominent feature types and have significant discriminatory power; the previous models fail to take into account this important family of features.

1) Feature Importance: To understand the importance of the features, we perform ablation experiments through removal of various feature groups to understand how each of these feature groups affect the classification and whether any feature group is masked by a stronger signal produced by the other feature group. We observe that user-level linguistic features are the most discriminative ones achieving an accuracy of **74.18%** whereas question-level linguistic features contributes to **59.07%** accuracy for $K = 20$ and time period of observation $t = 1$ month. As we shall see in the following, this can be attributed to the features based on the linguistic activity of the users that have the highest discriminative power.

B. Discriminative Features

In order to determine the discriminative power of each feature, we compute the chi-square (χ^2) value and the information gain. Table IX shows the order of all features based on the χ^2 value, where larger the value, higher is the discriminative power. The ranks of the features are very similar when ranked by information gain (Kullback–Leibler divergence). The most prominent ones among the user-level linguistic activities are the LIWC features. Among question-level features, the most discriminative features are the topic hierarchy features and topical edit features.

IX. DISCUSSION

There has been tremendous progress of web search over the past two decades. In spite of this success, many users' requirement still remains unanswered. Query assistance in form of query completion, related or similar queries, cannot always deal with satisfying complex, heterogeneous needs. There exists subjectivity in many query responses. Apart from subjectivity, there are questions, whose answers keep on changing based on the current context, and hence demand fresh and new outlook at it. CQA sites, such as Quora, Yahoo! Answers, Stack Overflow, have been serving the purpose to answer these different needs, e.g., opinionated, seeking

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

recommendations, openendedness. Questions can be framed in different styles, yet every asker expects to receive the correct and precise answers. However, not all questions get answers. In this paper, we tried to quantify various linguistic styles of askers, their editing activities to understand the answerability of question and have shown that indeed these linguistic features act as important discriminators separating the answered from the open questions.

A. Importance of This Paper

This paper is important in various ways because it unfolds the linguistic aspect of answerability and its importance in receiving answers. Unlike the previous studies by Dror *et al*. [1] and Yang *et al*. [2] which considers various heuristic features relating to mostly content of the question, we have looked beyond the content, emphasizing on the asker's writing styles, and editing activities as indicators of answerability. If an asker in acute need of answer, do not receive any answers, due to ill-framing or poor asking styles, it might create negative impact on the asker's mind and the asker might eventually leave the platform. This paper has potential to early detect such questions and recommend corrections which might lead to answers from the community. Our system can also help promoting these lowly visible questions to the more informed answerers, experts so that they receive adequate number of responses, thereby, improving the answer rate.

B. Generalizability

Though the entire study has been performed on a Quora data set, it can be easily adopted to other CQA sites such as Yahoo! Answers, Stack Overflow as well. The linguistic styles of the askers, the psycholinguistic aspect of the questions is available for other websites such as Yahoo! Answers, Stack Overflow, and so on. Topic, which is a unique entity to Quora, is similar to tag in other CQA platforms. Like Quora, Stack Overflow also contains editing histories of the questions. Thus, a summary of similar features based on linguistic styles of users can be easily engineered for all these other platforms to reap a similar benefit.

X. Conclusion

In this paper, we investigate various linguistic activities and observe how such activities affect answerability of questions in Quora by analyzing a large data set spanning over a period of 4 years. One of the primary lessons is that the language usage patterns correspond to quality factors that human judges would consider to decide if a question would remain unanswered. Based on these linguistic activities, we can efficiently discriminate the open and the answered questions. The characterization helps us to design a prediction model that can automatically identify whether a question remaining "open" for a specific time period t, will be answered or not. Our proposed prediction framework achieves an accuracy of **76.26%** and **68.33%** with high precision and recall for observation time period of 1 month and 3 months, respectively, outperforming the baseline methods convincingly. We observe that the user-level *linguistic* features are most discriminative compared to others.

One immediate future direction could be to automatically generate answers for the open questions by analyzing various other social media such as Twitter, Reddit, and so on. This cross-platform knowledge accumulation is an interesting future work.

References

[1] G. Dror, Y. Maarek, and I. Szpektor, "Will my question be answered? Predicting 'question answerability' in community question-answering sites," in *Proc. ECML-PKDD*, 2013, pp. 499–514.

[2] L. Yang *et al.*, "Analyzing and predicting not-answered questions in community-based question answering services," in *Proc. AAAI*, 2011, pp. 1273–1278.

[3] S. K. Maity, A. Kharb, and A. Mukherjee, "Language use matters: Analysis of the linguistic structure of question texts can characterize answerability in Quora," in *Proc. ICWSM*, 2017.

[4] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, "Knowledge sharing and Yahoo answers: Everyone knows something," in *Proc. WWW*, 2008, pp. 665–674

[5] B. Li and I. King, "Routing questions to appropriate answerers in community question answering services," in *Proc. CIKM*, 2010, pp. 1585–1588.

[6] A. Pal, S. Chang, and J. A. Konstan, "Evolution of experts in question answering communities," in *Proc. ICWSM*, 2012.

[7] P. Jurczyk and E. Agichtein, "Discovering authorities in question answer communities by using link analysis," in *Proc. CIKM*, 2007, pp. 919–922.

[8] K. Lerman and A. Galstyan, "analysis of social voting patterns on Digg," in *Proc. WOSN*, 2008, pp. 7–12.

[9] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, "A framework to predict the quality of answers with non-textual features," in *Proc. SIGIR*, 2006, pp. 228–235.

[10] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. WSDM*, 2008, pp. 183–194.

[11] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community QA," in *Proc. SIGIR*, 2010, pp. 411–418.

[12] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan, "Predictors of answer quality in online Q&A sites," in *Proc. CHI*, 2008, pp. 865–874.

[13] B. Li, T. Jin, M. R. Lyu, I. King, and B. Mak, "Analyzing and predicting question quality in community question answering services," in *Proc. WWW Companion*, 2012, pp. 775–782

[14] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor, "Learning from the past: Answering new questions with past answers," in *Proc. WWW*, 2012, pp. 759–768

[15] D. Correa and A. Sureka, "Fit or unfit: Analysis and prediction of 'closed questions' on stack overflow." in *Proc. COSN*, 2013, pp. 201–212.

[16] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering," in *Proc. SIGIR*, 2008, pp. 483–490.

[17] F. M. Harper, D. Moy, and J. A. Konstan, "Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites," in *Proc. CHI*, 2009, pp. 759–768.

[18] V. Bhat, A. Gokhale, R. Jadhav, J. S. Pudipeddi, and L. Akoglu, "Min(e)d your tags: Analysis of question response time in stackoverflow," in *Proc. ASONAM*, 2014, pp. 328–335.

[19] D. Correa and A. Sureka, "Chaff from the wheat: Characterization and modeling of deleted questions on stack overflow," in *Proc. WWW*, 2014, pp. 631–642

[20] M. Asaduzzaman, A. S. Mashiyat, C. K. Roy, and K. A. Schneider, "Answering questions about unanswered questions of stack overflow," in *Proc. MSR*, 2013, pp. 97–100.

[21] H. Wu, Y. Wang, and X. Cheng, "Incremental probabilistic latent semantic analysis for automatic question recommendation," in *Proc. RecSys*, 2008, pp. 99–106.

[22] J. Guo, S. Xu, S. Bao, and Y. Yu, "Tapping on the potential of Q&A community by recommending answer providers," in *Proc. CIKM*, 2008, pp. 921–930.

[23] M. Qu *et al.*, "Probabilistic question recommendation for question answering communities," in *Proc. WWW*, 2009, pp. 1229–1230

[24] G. Dror, Y. Koren, Y. Maarek, and I. Szpektor, "I want to answer; who has a question?: Yahoo! answers recommender system," in *Proc. KDD*, 2011, pp. 1109–1117.

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.

[25] F. Xu, Z. Ji, and B. Wang, "Dual role model for question recommendation in community question answering," in *Proc. SIGIR*, 2012, pp. 771–780.

[26] J. San Pedro and A. Karatzoglou, "Question recommendation for collaborative question answering systems with rankslda," in *Proc. RecSys*, 2014, pp. 193–200.

[27] D.-R. Liu, Y.-H. Chen, and C.-K. Huang, "QA document recommendations for communities of question–answering websites," *Knowl.-Based Syst.*, vol. 57, pp. 146–160, Feb. 2014.

[28] Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao, "Routing questions to the right users in online communities," in *Proc. ICDE*, 2009, pp. 700–711.

[29] Z. Yan and J. Zhou, "A new approach to answerer recommendation in community question answering services," in *Proc. ECIR*, 2012, pp. 121–132.

[30] B. Kim and J. Kim, "Question-aware prediction with candidate answer recommendation for visual question answering," *Electron. Lett.*, vol. 53, no. 18, pp. 1244–1246, 2017.

[31] X. Liu, W. B. Croft, and M. Koll, "Finding experts in community-based question-answering services," in *Proc. CIKM*, 2005, pp. 315–316.

[32] J. Zhang, M. S. Ackerman, L. Adamic, and K. K. Nam, "QuME: A mechanism to support expertise finding in online help-seeking communities," in *Proc. UIST*, 2007, pp. 111–114.

[33] A. Pal, F. M. Harper, and J. A. Konstan, "Exploring question selection bias to identify experts and potential experts in community question answering," *ACM Trans. Inf. Syst.*, vol. 30, no. 2, p. 10, May 2012.

[34] Y. Liu, Z. Lin, X. Zheng, and D. Chen, "Incorporating social information to perform diverse replier recommendation in question and answer communities," *J. Inf. Sci.*, vol. 42, no. 4, pp. 449–464, 2016.

[35] C. Stanley and M. D. Byrne, "Predicting tags for stackoverflow posts," in *Proc. ICCM*, 2013, pp. 1–6.

[36] L. Nie, Y.-L. Zhao, X. Wang, J. Shen, and T.-S. Chua, "Learning to recommend descriptive tags for questions in social forums," *Trans. Inf. Syst.*, vol. 32, no. 1, 2014, Art. no. 5.

[37] Y. Wu, W. Wu, Z. Li, and M. Zhou, "Improving recommendation of tail tags for questions in community question answering," in *Proc. AAAI*, 2016, pp. 3066–3072.

[38] S. K. Maity, J. S. S. Sahni, and A. Mukherjee, "Analysis and prediction of question topic popularity in community Q&A sites: A case study of Quora," in *Proc. ICWSM*, 2015, pp. 238–247.

[39] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Wisdom in the social crowd: An analysis of Quora," in *Proc. WWW*, 2013, pp. 1341–1352.

[40] O. Owoputi, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith, "Improved part-of-speech tagging for online conversational text with word clusters," in *Proc. NAACL*, 2013, pp. 380–390.

[41] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL Workshop*, vol. 8, 2004.

[42] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic Inquiry and Word Count*. Mahwah, NJ, USA: Lawerence Erlbaum Associates, 2001.

[43] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[44] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.

Suman Kalyan Maity received the M.S. and Ph.D. degrees in Computer Science and Engineering from IIT Kharagpur, Kharagpur, India, in 2014 and 2018, respectively.

He is currently a Post-Doctoral Fellow with the Kellogg School of Management, Northwestern University, Evanston, IL, USA. His current research interests include data science, computational social science, machine learning, and computational linguistics.

Dr. Maity was a recipient of the Ph.D. Fellowship from Microsoft Research India, Bengaluru, India, from 2014 to 2016, to conduct his research. He was also a recipient of the IBM Ph.D. Research Fellowship from 2016 to 2017, the Best Paper Honorable Mention Award in the ACM CSCW'16, and the Best Paper Award in the ICTS4eHealth 2018.

Aman Kharb received the B.Tech. and M.Tech. degrees in Computer Science and Engineering from IIT Kharagpur, Kharagpur, India, in 2017.

He is currently an Analyst with Goldman Sachs, Bengaluru, India.

Animesh Mukherjee received the Ph.D. degree from the Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, India. His Ph.D. thesis focus on self-organization of human speech sound inventories.

He was a Post-Doctoral Researcher with the Complex Systems Lagrange Lab, ISI Foundation, Turin, Italy. He was an Assistant Professor with the Department of Computer Science and Engineering, IIT Kharagpur, where he is currently an Associate Professor. He has authored or co-authored in top conferences such as ACM SIGKDD, ACM CIKM, ACM CSCW, ICWSM, ACL, EMNLP, COLING, and ACM/IEEE JCDL and journals such as *Proceedings of the National Academy of Sciences of the United States of America*, *Scientific Reports*, *ACM Transactions on Knowledge Discovery from Data*, *ACM Communications of the ACM*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, *Physical Review*, *Europhysics Letters*. His current research interests include applying complex system approaches (mainly complex networks and agent-based simulations) to different problems in computer science including human language evolution and change, web social media, information retrieval, and natural language processing.

Dr. Mukherjee is the Technical Program Committee Chair of IEEE/ACM COMSNETS 2018 and the Workshop Chair of ACM COMPUTE 2017. He regularly serves on the program committee of various top conferences such as IJCAI, EMNLP, COLING, IEEE/ACM JCDL, IEEE GLOBECOM and high impact journals such as *Proceedings of the National Academy of Sciences of the United States of America*, *Scientific Reports*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Knowledge Based Systems*, *Advances in Complex Systems*. He also has a prolific collaboration record with collaborators from different parts of the world from both academia and industry. His work has received media attention from various agencies including MIT Tech Review, BBC, Guardian, and so on. He was a recipient of the Humboldt Fellowship for Experienced Researchers from 2018 to 2021, Simons Associate, ICTP, Italy from 2014 to 2019, the Gandhian Young Technological Innovation Award in 2017, the IBM Faculty Award 2015, INSA Medal Young Scientists 2014, and INAE Young Engineering Award 2012 among many others.

Authorized licensed use limited to: MIT Libraries. Downloaded on December 06,2021 at 02:54:43 UTC from IEEE Xplore. Restrictions apply.