Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Discovering Response-Eliciting Factors in Social Question Answering: A Reddit Inspired Study

Danish
Indian Institute of Science
danish037@gmail.com

Yogesh Dahiya
Indian Institute of Science
yogeshd2612@gmail.com

Partha Talukdar
Indian Institute of Science
ppt@cds.iisc.ac.in

Abstract

Questions form an integral part of our everyday communication, both offline and online. Getting responses to our questions from others is fundamental to satisfying our information need and in extending our knowledge boundaries. A question may be represented using various factors such as social, syntactic, semantic, etc. We hypothesize that these factors contribute with varying degrees towards getting responses from others for a given question. We perform a thorough empirical study to measure effects of these factors using a novel question and answer dataset from the website Reddit.com. We also use a sparse non-negative matrix factorization technique to automatically induce *interpretable* semantic factors from the question dataset. Such interpretable factor-based analysis overcomes limitations faced by prior related research. We also document various patterns on response prediction we observe during our analysis. For instance, we found that preference-probing questions are rarely answered by actors.

1 Introduction

Questions and the responses they elicit are a ubiquitous and fundamental part of our everyday communication. Through such Questions and Answers (QA), we quench our curiosities, clarify doubts, validate our ideas, and seek advice, among others. It has been established that questions form an integral part in our quest to extend our knowledge boundaries (Sammut and Banerji 1986). It has also been observed that useful responses correspond to *good* questions (Agichtein et al. 2008). This raises the following challenge: *what factors constitute a good question which is more likely to elicit a response*?

Importance of asking right questions in specific settings has been previously explored, e.g., in classroom (King 1994), and in corporate environment (Ross 2009). However, most of these studies either had no empirical evaluation at all or otherwise consisted of very small samples.

Along with the growth of the World Wide Web, many large online QA sites, such as Yahoo Answers, Stack Overflow, Quora, etc., have been successful in connecting responders to inquirers who post questions on these sites. Such online QA forums may be categorized as **Single Inquirer Multiple Responders (SIMR)**, where a question from a single user may be responded to by multiple other responders. Prior research has used datasets from these sites to predict whether a question is of high quality (Ravi et al. 2014; Li et al. 2012); or guess if a given question will be answered (Yang et al. 2011; Dror, Maarek, and Szpektor 2013); or estimate the number of answers a question will receive (Dror, Maarek, and Szpektor 2013); or select the best response to a given question (Adamic et al. 2008). However, analyzing factors of a question which are likely to elicit a response has been outside the scope of such prior work. In other words, we wish to bring out the commonalities among answered questions and understand reasons that prevent attaining answers.

To address these shortcomings, in this paper we present an empirical analysis to determine factors of a question which are more likely to elicit a response. We make use of the IAmA subreddit of the popular Internet website Reddit.com. In each discussion thread of this online forum, a celebrity answers questions submitted by anonymous users. Thus, dataset from this subreddit may be categorized as **Multiple Inquirers Single Responder (MISR)**. Such MISR datasets provide an ideal starting point to identify response-eliciting factors of a question, as the undesirable confounds produced due to the presence of multiple responders in SIMR datasets are not present in such MISR datasets. We further argue why such a dataset is better suited for our task in Section 3.

We make the following contributions:

- We address the important problem of automatically identifying response-eliciting factors of a question. We explore effectiveness of various factors, viz., orthographic, temporal, syntactic and also semantics of the question. To the best of our knowledge, this is the first thorough analysis of its kind.
- We make use of a novel dataset, questions and responses from the IAmA subreddit of reddit.com. This MISR dataset provides additional benefits compared to SIMR datasets which have been explored in previous related research.
- We provide a sparse, non-negative matrix factorization-based framework to automatically induce *interpretable* semantic factors of a question collection. Through extensive experiments on real datasets, we demonstrate that such a factorization-based technique results in signifi-

cantly more interpretable factors compared to standard topic modeling techniques, such as Latent Dirichlet Allocation (LDA).

- All the code and data used in the paper is now available at https://github.com/malllabiisc/reddit-icwsm16.

2 Related Work

Studies on questioning techniques date back to Socrates (Paul and Elder 2007; Carey and Mullan 2004), who encouraged a systematic, disciplined, and deep questioning of fundamental concepts, theories, issues and problems. Socratic questioning is widely adopted in education and psychotherapy. Under the Socratic Questioning scheme (Paul and Elder 2006), questions are grouped as follows — i) Clarifying questions: ones seeking further explanation, ii) Challenging the assumptions: questions that challenge the constraints, iii) Argument based questions: ones that reason behind the underlying theory or seek evidence, iv) Alternate viewpoints: questions that analyze the given scenario with an altogether different perspective, v) Implication and Consequence based questions.

Since Socrates, many different taxonomies have been discussed. Bloom's revised taxonomy given by (Anderson, Krathwohl, and Bloom 2001) is based upon dividing questions into levels such that the amount of mental activity required to respond increases after each level. Their categories are — remembering, understanding, applying, analyzing, evaluating and creating. Grouping of questions into Factual, Procedural, Opinion-oriented, Task-oriented and Advice related categories is presented in (Nam, Ackerman, and Adamic 2009).

Role of Socratic techniques in thinking, teaching and learning has also been explored (Elder and Paul 1998). Hypothetical questions too have been studied independently and have been found to foster creativity (Newman 2000). While there has been considerable thought given over such demarcations and question formulation techniques, they neither study question effectiveness with respect to response rate nor they are supported by any large datasets as most of the experiments were performed in a typical classroom sized setting.

Prior research on Community Question Answering (CQA) has addressed issues as diverse as predicting whether a particular answer will be chosen by the inquirer or not (Adamic et al. 2008), predicting answer quality (Shah and Pomerantz 2010; Jeon et al. 2006) and quantity (Dror, Maarek, and Szpektor 2013), understanding question type (Allamanis and Sutton 2013) and quality (Ravi et al. 2014; Yang et al. 2011; Li et al. 2012), analyzing inquirer's satisfaction (Liu, Bian, and Agichtein 2008) and responders' motivation, finding similar questions (Li and Manandhar 2011) and expert potential responders (Li and King 2010). Although the aforementioned studies are helpful, we are more curious about the factors of a question that are more likely to generate a response, especially in the MISR setting where there is just one responder but multiple inquirers.

The task of predicting question quality is studied in (Ravi et al. 2014) where latent topics were found to be useful estimators for the task. Our work, although similar in spirit, goes one step beyond to identify latent topics that are not only effective but also *interpretable* so that one can understand the qualities that response-eliciting questions share along with issues that are common across questions that fail to generate a response. Another method to predict question quality is proposed in (Li et al. 2012), where the task of identifying salient features of quality of a question is left as part of future work. This is precisely the problem we address in this paper.

Variety of interesting questions have been studied using the Reddit conversation network ranging from understanding how people react to online discussions (Jaech et al. 2015), modeling the most reportable events in stories (Ouyang and McKeown 2015), deciphering persuasion strategies (Tan et al. 2016), understanding factors underlying successful favor requests (Althoff, Danescu-Niculescu-Mizil, and Jurafsky 2014) to analysing domestic abuse (Ray and Homan 2015).

Domain	Questions Asked	Questions Replied	Response Rate
Actor	58859	3060	5.19
Author	21295	3752	17.61
Politician	13866	1914	13.8
Director	24196	3295	13.61
Total	118216	12021	10.16

Table 1: Reddit IAmA datasets from four domains used in the experiments in this paper. See Section 3 for further details

3 Dataset

Reddit is the 26th most popular website, with about 231 million unique monthy visitors.[1][2] It also comprises of over 9,000 subreddits which are sub-forums within reddit focussed towards specific topics. Subreddits span diverse categories like News, Sports, Machine Learning etc. Reddit is also a home of subreddits like: ELIF (Explain like I'm five), TIL (Today I learnt), AMA(Ask Me Anything) etc.

Various celebrities and noteworthy personalities have used reddit as a means to interact with Internet users, such conversations fall under the Ask-Me-Anything and its variant subreddits. IAmA, AMA and casualama are three of the most popular Ask-Me-Anything variants. IAmA is reserved for distinguished personalities, with an exception for people who have a truly interesting and unique event to take questions about. The other two AMA's are open to a more wider audience for sharing their life events and allowing other reddit users to ask questions related to those events.

IAmA's is one of the most popular subreddits that has featured notable politicians, actors, directors, authors, businessmen, athletes and musicians. IAmA posts gain a lot of attention, and thousands of questions are asked in each IAmA

[1] https://www.similarweb.com/website/reddit.com

[2] https://www.reddit.com/about/

post. But owing to time constraints, not all questions are answered. This gives us a good ground to understand and analyze what gets answered and what not.

In particular, we study four popular categories of celebrities — actors, authors, directors and politicians. In each category, we analyse the top 50 upvoted posts, which aggregate over 110,000 questions, with an average reply rate of 10.16%. Since some questions arrive after the celebrity has moved out of the conversation, we ignore all the questions after the last successfully answered question. Reddit allows for threaded conversations, where users can comment over other comments. But to avoid any bias from the discourse of the comments in such threads, we ignore questions in deep threaded conversations and constrain ourselves to questions posted at the topmost level only. Since some comments also get posted at the topmost level, we only consider comments that have a question mark in them.

Unlike other analysis on community QA including Yahoo! Answers and StackOverflow, the Reddit dataset offers following unique advantages.

- In the reddit dataset, the responder in each IAmA is a *single* notable personality with average reply rate of around 10.16%. This gives us good ground to understand the issues with unanswered questions. On the contrary, 99.4% of questions studied in (Ravi et al. 2014) and 95% of questions in (Shah and Pomerantz 2010) received at least one answer, often involving multiple responders.
- The notion of quality is natural and intuitive in our dataset, where a single responder handpicks a few questions that he/she wishes to answer. Whereas, in (Shah and Pomerantz 2010), authors had to rely on domain experts and mechanical turks for quality evaluations. For the StackOverflow dataset in (Ravi et al. 2014), authors came up with a quality measure estimated by the ratio of score and views. A StackOverflow question might be of poor quality, yet inquiring about a common bug might mislead the quality estimation. Hence, we believe, our IAmA based dataset gives us a better setting to explore factors involved in question quality.
- Our dataset can be thought of as a *Multiple Inquirers Single Responder* forum, which helps us gain more control over the responses, as opposed to the Single Inquirer Multiple Responders datasets such Yahoo QA, StakckOveflow etc.

Table 1 presents statistics of various IAmA datasets we considered as a part of our study.

4 Success Factors of Questions

In this section, we study various factors of questions that can result in healthy response rates. The factors we consider range from orthographic, temporal, social, and syntactical, to semantic aspects.

4.1 Orthographic Factors

Length: Do short questions win over their longer variants, as the responder may not be interested in comprehending and then answering long questions? Or, are longer questions better as they offer more context? Are shorter and crisper questions more direct and focused and have a better chance at getting answered? We analyze the impact of length on response rate to answer the aforementioned question.

4.2 Temporal Factors

Time of Question: Does the timing of asking a question play any role in determining the response rate? We hypothesize that questions that are asked early on have far less competing questions and hence should have better chances of eliciting response.

We capture temporal information in two ways: (1) we note the fraction of *questions asked* in the IAmA before a given question is posted as an estimate of the time of question; (2) we use the fraction of *time elapsed* in the IAmA as another indicator of the time of the question. In most cases, we see that the time features complement each other.

4.3 Social Factors

Politeness: Are polite questions more likely to generate a response? Or, is it the case that the default level of politeness expressed in the IAmA dataset already sufficient, and hence any additional politeness in the question is unlikely to positively affect response rate?

Politeness has been actively explored in the recent past in a variety of others research settings (Tsang 2006; Bartlett and DeSteno 2006). We employ the model introduced by Danescu-Niculescu-Mizil et al. (2013) to measure politeness level of questions. This model bases its politeness score on the occurrences of greetings, apologies and hedges in the question.

4.4 Syntactic Factors

Syntactic: We ask whether questions that are simply formulated have better chances of getting answered? Syntactic features, such as parse tree depth, verb phrase depth, and their ratios etc., have been used in past research (Klein and Manning 2003) as proxies for sentence complexity. In fact, such features have also been recently used to study syntactic complexity of reddit comments (Ouyang and McKeown 2015). After generating constituent parse trees from the Stanford Corenlp package (Manning et al. 2014), we employ 16 such features to capture the essence of syntactic complexity in a given question.

We look at a few simple and a few complex sentences from the IAmA by President Barack Obama in Table 2 and demonstrate how the features capture the varied levels of complexity. Since there can be various sentences and subquestions in a given question, we calculate the average, maximum and minimum values of parse tree depths and verb phrase depths. It is because of such statistical aggregation techniques that we end up with 16 syntax features, but the basis of these features rest upon — parse tree of the sentence, verb phrase subtree and their ratios.

4.5 Forum Factors

Redundancy: Is a question which is very similar to already asked (or answered) questions in a given IAmA forum less

Sentence	Depth of Sentence Parse Tree	#Verb Phrases	Max Verb Phrase Depth	Verb Phrase Depth / Sentence Depth
Who's your favourite Basketball player?	2	0	0	0.0
What's the recipe for the White House's beer?	6	0	0	0.0
Mr. President - What issues, if any, do you agree with Mitt Romney that are not commonly endorsed by the majority of the Democratic Party?	11	4	9	0.81

Table 2: A few example sentences from President Obama's IAmA and their corresponding syntax features. See Section 4.4 for details

likely to get a response? We think that is indeed the case and include factors in our analysis to account for question redundancy. For instance, a few redundant questions asked to a popular chef are listed below.
– *What's your favorite Middle Eastern Dish?*
– *What's your favourite dish to prepare?*
– *What's your favourite French meal?*
As the first few questions were not answered in the series of the above mentioned questions, it is nearly certain that the responder is not interested in any such questions. By accounting for redundancy we hope to tackle similar and frequent scenarios.

We estimated the redundancy score of a given question as the maximum similarity score achieved with any of the other questions previously asked in the same IAmA.

Relevance: For each IAmA, the responder usually posts a description to set the tone of the IAmA. We ask whether questions which are more aligned to the posted description more likely to receive a response? The posted descriptions usually carry information about the celebrity responder's current affiliation and engagements, and hence the hypothesis is that questions which are in line with such descriptions should outweigh other questions. In other words, relevant questions should attract more responses from the responder.

For both the relevance and redundancy factors, we came up with our own novel extension of Jaccard Similarity to account for sentence similarities. For two given sets A and B, the Jaccard Similarity is given by

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

For our case, let A and B be sets of words corresponding to the two questions to be compared. Strictly, $A \cap B$ would translate to the count of the words matched across the sets of A and B. But consider the following two sentences:
– *How far is your workplace from your house?*
– *How far is your office from your home?*

With the strict definition, we would not be able to capture that the two sentences are completely similar, for all practical purposes. Hence, we consider the Glove embeddings (Pennington, Socher, and Manning 2014), and synset hierarchies to extend the scope of our matching. Two words are considered same, if (1) the two words are synonyms to each other and (2) if one word lies in top-K nearest neighbours of the other word in Glove embedding space. We found 20 as a reasonable choice for k in our setting.

This technique helps us to capture similarity of pairs like <home, house> and <office, workplace> and hence helps us better estimate the similarity of two sentences.

4.6 Semantic Factors

The factors described so far consider various aspects of the questions being analyzed. However, none of them explicitly look at the semantic content of the question and perform analysis based on the semantic type of the question. For example, given questions of the following form posed to actors, *'what is your favorite movie?'*, *'what is your favorite book?'*, etc. we would like to automatically group all such preference-probing questions into one category and then determine the response rate for such types of questions from actor responders. However, such categorization of questions is not readily available as we only have the list of questions, and no additional annotation on top of them.

Ideally, we would like to discover such categorical structure in the data automatically. Topic modeling techniques such as Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) may be employed to discover such latent structure in the question dataset. Given a set of questions, such techniques will induce topics as probability distribution over words. Previously, Sentence Topic Models (STM) (Ravi et al. 2014; Brody and Elhadad 2010) have been used to capture local aspects within documents and online reviews. The sentence model is generated by running LDA over sentences rather than documents. Ultimately, each question is going to be represented in terms of such induced topics. We note that interpretability, i.e., coherence among questions which share a given topic with high weights, is of paramount importance here as all subsequent response-rate analysis are going to be hinged on the label or meaning of each topic. Unfortunately, as we shall see in Section 5, topics induced by LDA and STM don't achieve the desired level of interpretability.

To overcome this limitation, we explore other latent factorization methods. Recently, Non-Negative Sparse Embedding (NNSE) (Murphy, Talukdar, and Mitchell 2012; Fyshe et al. 2015) has been proposed which tends to induce effective as well as interpretable factors. In order to apply NNSE to our question dataset, we first represent the data as a co-occurrence matrix X where rows correspond to questions

Tags	#Questions	#co-occurrence entries before	#co-occurrence entries after	Factor increase
Author	21295	349288	679476	1.94
Actor	58859	701702	1368703	1.95
Politics	13866	228820	438089	1.91
Director	24196	344176	658226	1.91

Table 3: Effect of extension using Wordnet synsets on the co-occurrence matrix. See Section 4.6

and columns correspond to words. Each question is additionally augmented with word sense-restricted synsets from Wordnet. The effect after the synset extension from Wordnet can be seen in Table 3, This extended co-occurrence matrix X is usually of very high dimension (e.g., 100k x 1m). We first reduce dimensionality of the matrix using sparse SVD. The number of dimensions in the SVD space is selected based on knee-plot analysis of eigenvalues obtained during SVD decomposition. The rank r approximation $X_{n\times r}$ obtained from SVD is then factorized into two matrices using NNSE, which minimize the following objective.

$$\arg\min_{A,D} \frac{1}{2}\sum_{i=1}^{n} \parallel X_{i,:} - A_{i,:} \times D \parallel^2$$

$$\begin{aligned} \text{such that} : D_{i,:}D_{i,:}^T \leq 1, \forall 1 \leq i \leq k \\ A_{i,j} \geq 0, 1 \leq i \leq n, 1 \leq j \leq k \\ \parallel A_{i:} \parallel_1 \leq \lambda_1, 1 \leq i \leq n \end{aligned}$$

where n is the number of questions, and k is the resulting number of latent factors induced by NNSE. We note that NNSE imposes non-negativity and sparsity penalty on the rows of matrix A. Though the objective represents a non-convex system, the loss function is convex when we solve for A with a fixed D (and vice versa). In such scenarios, Alternating Minimization has been established to converge to a local optima (Mairal et al. 2010; Murphy, Talukdar, and Mitchell 2012). The solution for A is found with LARS implementation (Efron et al. 2004) of LASSO regression with non-negativity constrains; and D is found via gradient descent methods. The SPAMS package may be used for this optimization (Bach et al. 2010). At the end of this process, $A_{i,j}$ represents the membership weight of question i belonging to latent factor j.

5 Experiments

In this section, we evaluate impact of various factors discussed in Section 4 on response rate of questions from different domains. In Section 5.1, we measure the effect of factors by the predictive power they possess. It might be worthwhile to note that our end goal is not to predict the answerability of questions, but to estimate the predictive strength of the factors. In Section 5.2, we visit the key question of understanding the underlying latent factors responsible for eliciting responses.

Domain	Temporal Factor Feature 1	Temporal Factor Feature 2	Redundancy
Author	89.57	83.24	56.92
Politician	84.29	115.91	63.06
Actor	189.81	218.37	141.23
Director	63.92	85.37	47.42

Table 4: Average Precision (AP) gains for temporal and redundancy factors over a random baseline. See Section 5.1 for details.

5.1 Do non-semantic factors influence response-rate?

Datasets: We experiment with four popular domains — actors, authors, director and politicians. These domains covered more than 110,000 questions, and only about 10% of them generated a response. Statistics of the IAmA datasets are presented in Table 1.

Metric & Classifier: In order to measure predictive power of a subset of factors, we train a L_2 and L_1 regularized (i.e., elastic net) classifier using only those subset of factors. Hyperparameters of the classifier is tuned using over a development set using grid search. We use area under the receiver operating characteristics curve (ROC AUC) of the classifier on held out test data as our metric. This metric essentially measures how well the classifier ranks a randomly chosen positive question over a randomly chosen negative question. Please note that the dataset is highly skewed with significantly more negative questions than positive ones. This measure provides a balanced metric while accounting for the skewed data.

Baselines: To evaluate the strength and decisiveness of our probable factors, we test our system against the random and bag-of-words (BoW) baselines. In the Random baseline, each question is randomly given one of the two labels — answered or not answered.

The bag of words model comprises of each and every word in the vocabulary as a feature, hence aggregating up to 13,704 features averaged across the four domains. Due to the large number of features, this Unigram model performs reasonably well (AUC 0.65), but it doesn't help us in answering our general question of — *Which factors help a question get answered?* – because the unigram features don't generalize to the factors that we are interested in evaluating.

Experimental results comparing performance of the classifier with different features on multiple datasets are presented in Table 5. Based on this table, we discuss predictive capabilities of various factors below. Please refer to Section 4 for description of the factors and how we computed them.

Orthographic Factors From Table 5, we observe that the length of the questions (measured in terms of numbers of tokens in the question), the only orthographic factor feature we considered, plays practically no role in influencing response rate. This is evident from the fact that the classifier with length as the only feature achieves AUC of 0.51 on av-

Feature (Factor)	Actor	Author	Politician	Director	Average
Random Baseline	.50	. 50	.50	.50	0.50
Unigram Baseline	.68	.66	.64	.61	0.65
Length (Orthographic)	.48	.49	.54	.52	0.51
Syntactic	.53	.52	.53	.50	0.52
Syntactic + Length	.54	.52	.53	.49	0.52
Temporal	.66	.67	.67	.60	0.65
Redundancy (Forum)	.71	.65	.64	.62	0.66
Relevance (Forum)	.49	.51	.58	.51	0.52
Politeness (Social)	.48	.52	.54	.52	0.52
Temporal + Politeness + Relevance + Redundancy + Syntax	**.74**	**.70**	**.73**	**.64**	**0.70**

Table 5: ROC AUC values for a regularized logistic regression classifier using different features in various domains. For reference, performance of a random baseline is also shown. Apart from length, all other features improve performance over the random baseline. See Section 5.1. See Section 5 for details.

erage across all four domains compared to AUC of 0.5 of the random classifier.

Syntactic Factors From Table 5, we clearly see that syntax-based features add very little little predictive power to the classifier (0.52 vs 0.50 of random). Though our syntax features are rigorous enough to capture the nuances of complexity (e.g., see Table 2), but the responses to questions don't heavily depend on the complexity of the sentence. We observed that combining syntax with orthographic features also didn't increase predictive power.

Temporal Factors We find that temporal features play a significant role in the response rate. This is evident from Table 5 where the classifier with temporal factor features achieves a significantly higher AUC score of 0.65 compared to random 0.5. As we had hypothesized earlier, questions that are asked early tend to be replied more often than others.

In addition to classifier's AUC score, we measured effect of temporal factors using Average Precision (AP) as well. For questions in a given domain, AP is computed over two ordering of the questions in that domain: (1) ordering of all questions based on the value of the temporal factor features; and (2) randomly shuffled question sequences. Percentage AP gains of the feature-based ranking over the random ranking (AP averaged over thousand trials) are summarized in Table 4. From this, we observe a clear trend that temporal features significantly aid in response prediction, sometimes with gains as high as 218%. We think that the responder is initially exposed to far lesser number of questions compared to a situation in the middle or towards the end of the IAmA when the number of questions demanding his or her attention are huge.

Forum Factors

Redundancy Our dataset consists of prominent celebrities, and they gain undeniably high attention among Reddit users. Due to large participation, the number of similar questions is high, as many users wish to know similar facts, preferences, likings and happenings. Redundancy comes out as one of the most promising factors in understanding questions that get answered. Examples of a few redundant questions are shown in Section 4.5.

The original, and genuine questions, which are identified by our redundant factor feature, are heavily preferred over questions that are redundant and stale. This is established by the fact the classifier which accounts for redundancy achieves a significantly higher AUC score of 0.66 compared to the random baseline.

Relevance of the question, with the post description by the celebrity responder, shows only faint signals with the response rate. The description given by the celebrities is usually very short to capture the variety of questions. Hence we don't see any meaningful dependencies between relevance and response rate (0.52 AUC).

Politeness Politeness, a seemingly important cue for demystifying question qualities, surprisingly, didn't come out as a strong predictor of response rate. In Table 5 the classifier with politeness forum factor feature achieves an AUC score of only 0.52. We have observed that the Reddit culture is very informal, frank and open. Hence, making requests extra polite might not help while framing questions in such scenarios. Of all domains, politeness is most important in the case of prominent politicians.

5.2 Do Induced Semantic Factors Help Discover Response Trends?

So far, we have handcrafted the seemingly most important factors but we can never account for patterns other than what we are looking for. In any large dataset as ours, creating an exhaustive set that can capture all such factors is humanly impossible. Also for each factor, we need to train a system that can well detect and measure it in an unknown question. In such scenarios, the need to automatically discover latent dimensions is essential. As mentioned in Section 4.6, we use LDA, STM[3], and NNSE to induce semantic factors present in the question dataset. First, we shall present comparisons between interpretability of factors induced by these three methods. Subsequently, we shall measure the response predictive power of these induced semantic factors.

[3]We don't consider the Generalized Mallows Models (GMM) as it was not found to be effective in (Ravi et al. 2014).

Method	Top Two Questions in the Latent Factor
LDA	*– Do you think that if you lived in an urban environment when these stories came to you, you might have* **written about** *rats or pigeons?* *– To what extent should historical analysis of religious figures impact the way people practice* **faith***? Or do you feel that the events of history are independent from the values of modern religion?*
	– Oddly **wanting your book** *in my collection. A long shot but... Can I please have a* **signed copy** *too? d: ps:* **sending internet hugs** *from uk!* *– I have a history final today and your* **crash course video on civil rights** *was enormously helpful. thank you. So my question is: where do you get all of your information for crash course videos?*
	– How did everyone else in the fox news studio **treat** *you? Were they hostile, friendly, indifferent, etc?* *– I* **love your writings***, I have read fight club, survivor, and damned and now can not wait to read doomed. My question, do you ever read Neil Gaiman?*
STM (Ravi et al. 2014)	*– Do you have any* **humiliating tales** *of my uncle Mike Loughery with whom you worked back in the day. I'd like to humiliate him with them* *– Can you at all comment as to whether or not the Yuuzhan Vong will appear in the new* **Star Wars**
	– What does your **book** *have to say of Pontious Pilate...* *– Have you ever actually* **read the book** *to your daughter in an attempt to get her to fall asleep*
	– How many **rejection slips** *did you get before you got published* *– Did it make you happy knowing you* **deprived me** *and countless others* **sleep** *for weeks*
NNSE	*– I will ask this - Is there any* **advice** *that you wish someone wouldve given your parents that would have smoothed out some of those painful but lets be honest here funny experiences for you* *– I'm only 21 and havent really* **published anything** *in places people have actually heard of. I've been submitting to some lit journals thoughheres hoping writing is what I want to do for a living...*
	– thinking of updating the site, I preorderd book the first I heard about it it says it will be delivered on oct 29. I'm excited. **I' m excited for you** *I'm excited by so many things ...* *– Hi Robert,* **love your work***. Thank you I'm curious how you feel about the modern binge style of consuming tv shows and comics. I really love walking dead in both forms but I...*
	– ohn **I'm a big fan** *first off I'm a teenage guy whos read most of your books and while they do involve love I wouldnt say your books are love stories ...* *– Hi john whats your favourite question to be asked, Yes, I'm canadian and is there anything that youre hoping* **people would ask big fan of your work** *both in print and on youtube*

Table 6: Three randomly selected latent factors induced by LDA, STM and NNSE, and top ranking questions in each such factor. The main perceived theme of each question is highlighted in bold manually. We find that the factors induced by NNSE are usually much more interpretable compared to LDA and STM. Because of this interpretability, we use semantic factors induced by NNSE for all experiments in the paper. See Section 5.2 for details.

LDA vs STM vs NNSE: We reiterate that finding latent factors that are interpretable is not just a luxury but a bare necessity in our setting as we need to understand what kind of latent semantic factors play a role in maximizing response rate. For this, we compared the latent factors induced by LDA, STM and NNSE, examples of which are in Table 6. In this table, three randomly selected latent factors induced each by LDA, STM and NNSE are shown. Also, for each latent factor, top two most active questions in that dimension are shown. For easy reference, the main theme of each question is manually marked in bold. From this table, we observe that NNSE is able to produce much more interpretable latent semantic factors compared to LDA and STM. Such lack of interpretability in LDA topics was also observed in another prior work (Althoff, Danescu-Niculescu-Mizil, and Jurafsky 2014). Given the interpretability advantage with NNSE, we use the latent factors induced by this method in subsequent analysis.

Having successfully induced interpretable semantic factors using NNSE which have good number of questions attached to them, we analyzed the dimensions of questions with extremely high and extremely low reply rates. Please note that such latent factors are induced separately for each domain. Experimental results comparing NNSE latent factors in three domains, overall response rate in the domain, response rate over questions in the factor, and examples of top questions in each such factor are shown in Table 7. Based on this table, we list below a few trends. We point out that this analysis and trend recognition would have been impossible without the ability to automatically induce interpretable semantic factors.

Actors We found that adulation techniques worked well in eliciting a response for actors: 15.88% response rate in Actor latent factor 524 in Table 7 compared to domain response rate of 5.19%. Based on the top questions in this factor, we can easily identify that this is a fan-related factor. Authors seem to reply more if the inquirer describes himself as a huge fan or if he expresses some liking for their movies and role. We also learnt that actors weren't very comfortable when it came to questions diving into their non-camera life (Actor factor 880). Also many actors were evasive when asked about their favorite actors, movies, meals etc (Author factor 852).

Politicians We observe that Politicians were prompt in clarifying all fund related issues pertaining to their campaigns (Politician factor 927 in Table 7). Whereas not many politicians seemed to be happy in taking questions on wage rise and the job situations in the country (Politician factor 304).

Author We observe that many users inquired authors about how they can pursue a career in writing, even more asked about writing advices. We found that such questions were generously replied: 36.53% response rate in factor 742 of the Author domain, compared to domain response rate of 17.62%. Also, authors answered a lot of questions that

Domain (Overall Response Rate)	Latent Factor # (Response Rate)	Sample Frequent n-grams of Questions in Latent Factor	Top Ranking Questions in Latent Factor
Actor (5.19%)	524 (15.88%)	*huge fan, loved movies, really love*	*– i m a huge fan of your cooking and have been watching you on television since i was a child so id love your input on a couple things ...* *– just like to start off by saying i love the show ... have you been involved with any popular shows or films?*
	297 (13.79%)	*story behind, behind the scenes*	*– i heard that you got a concussion and had to go the er while shooting one of the seasons whats the story behind that ...* *– hey arnold whats the story behind this picture*
	880 (0.0%)	*real life*	*– have you started saying bitch more in real life since the show started* *– do you say bitch as much in real life.*
	852 (1.09%)	*favorite actor, favorite actress, favorite play*	*– what role was your favorite to play and why* *– hey bryan just wanted to say youre an awesome actor and i was curious what your favorite breakfast cereal*
Politician (13.8%)	927 (31.5%)	*money, campaign, influence money, hard earned*	*– has your campaign accepted any money from corporate donors if so which ones and will their contributions affect your decisions* *– what about campaign money? are you running this campaign without anything to fund it?*
	567 (28.3%)	*issue, matter, think, social issues, net neutrality*	*– what in your opinion is the most pressing issue facing the uk at the present time?* *– ... david cameron himself wants to confront the european court of human rights so id like to know your take on this as well as the underlying issues ...*
	304 (2.43%)	*pay, wage, tax, job, minimum wage*	*– do you not worry that a ten pounds minimum wage would crush independent businesses and severely increase mass unemployment* *– what are your thoughts on the proposals on minimum wage to 8?*
	567 (3.57%)	*movie, film, estate*	*– what do you think about the movie lego?* *– have you seen the movie ...?*
Author (17.61%)	742 (36.53%)	*writing stories, advice, aspiring, approach writing*	*– i am a somewhat aspiring author i write a lot on writing prompts and people there have gotten to know me a bit i am currently working on a book based on a writing prompt ...* *– my question is how did your following on reddit help you get a publisher on board ...*
	136 (34.61%)	*idea, thought experiment, mind*	*– what made you come up with that idea and how do you come up with ideas in general for your stories?* *– what are your thoughts regarding the 2012 mayan prophecy?*
	4 (7.14%)	*inspired, inspires, work, write*	*– hi john im a great fan having read all of your books bar looking for alaska may i ask what inspired you to write paper towns?* *– ... i was wondering what inspired you to write what made you decide to write suspenseful novels ...*
	118 (7.31%)	*favorite, favourite, book, author, read*	*– who is your favorite author. do you ever read your own books if so which one is your personal favorite?* *– ... also out of curiosity who is your favorite author a very original question i know ?*

Table 7: Automatically induced latent semantic factors with highest and lowest response rates in multiple domains are shown. Base response rate for the domain, and the response rate for each factor is shown in brackets. Top ranking questions in each latent factor along with the most frequent n-grams in questions belonging to the particular latent factor are also shown. We point out the interpretable nature of each semantic factor (based on high-ranking questions associated with it), which allows us to draw sample conclusion as follows: while actors are unwilling to answer questions relating to their favorites or real life, authors are more willing to answer questions relating to supporting aspiring new authors. Ability to discover such insights using an automated process and a novel dataset is the main contribution of the paper. Please see Section 5.2 for details.

questioned about their ideas, thoughts and preferences (Author factor 136). However, they were a little less responsive when asked about inspiration (factor 4) or favorites (factor 118). This might be attributed to the fact that questions of these types are extremely frequently posed to authors, and due to the redundancy, they may answer only a few of them (please note that the response rate in these factors is not 0).

5.3 Generalizability to SIMR Datasets

In order to test the generalizability of the NNSE-based analysis method explored in the paper, we also applied it over a Yahoo! Q&A dataset which is available as a part of Yahoo! Research Alliance Webscope program[4]. Even though we are unable to include complete details of the experiment due to lack of space, we report that we do observe trends consistent with the ones reported in the experiments above. For instance, in the 'Authors & Books' category in the Yahoo! Q&A dataset, we found that questions related to *favorite book and author* received the most number of responses. Also, questions about Harry Potter books, and Da Vinci Code attracted numerous responses. Questions concerning favorite movies and Harry Potter movies elicited the largest number of replies in the 'Entertainment' category. This is a reverse trend compared to the observations in the IAmA dataset. We hypothesize that since Yahoo! Answers falls into a Single Inquirer Multiple Responders dataset, many responders shared their favorite books, authors and movies and thus the high response rate; whereas in a single responder setting (like IAmA) such questions remained predominantly unanswered. In the 'Politics' domain, questions about President

[4]https://webscope.sandbox.yahoo.com/

George Bush and Gay Marriage Rights were the most answered.

As the dataset consisted of questions from October, 2007, the factors succinctly described above are depictive of that period. For instance, the movie Harry Potter and the Order of the Phoenix was released around that time.

6 Conclusion

Question-Answering forms an integral part of our everyday communication. While some questions elicit a lot of responses, many others go unanswered. In this paper, we present a large-scale empirical analysis to identify interpretable factors underlying response-eliciting questions. To the best of our knowledge, this is the first such analysis of its kind. In particular, we focus on the Multiple Inquirers Single Responder (MISR) setting where there are multiple inquirers asking questions to a single responder, and where the responder has a choice to not answer any particular question. We used a novel dataset from the website Reddit.com, and considered several factors underlying questions, viz., orthographic, temporal, syntactic, and semantic. For semantic features, we used a sparse non-negative matrix factorization technique to automatically identify *interpretable* latent factors. Because of this automated analysis, we are able to observe a few interesting and non-trivial trends, while overcoming limitations faced by prior related research (Ravi et al. 2014). For instance, we observed that all the advice related questions were generously entertained by Authors, as long as they carried some context about their writing pursuits. Similarly, Actors were keen on making people aware about the behind-the-scene events, whenever asked. These trends are hard to capture otherwise, as designing a system to detect such particular cases requires training over large annotated corpus.

As part of future work, we hope to explore other factorization techniques, e.g., hierarchical latent factors, for even more effective and interpretable latent factors. Additionally, we hope to use the insights gained in this study to explore how an existing question may be rewritten to elicit response from voluntary responders.

7 Acknowledgments

This research is supported in parts by gifts from Google Research and Accenture Technology Labs. We thank the anonymous reviewers for their constructive comments and Tushar Nagarajan for useful discussions. We also acknowledge Uday Saini, Aakanksha Naik, Madhav Nimishakavi and Siddhant Tuli for carefully reading drafts of this paper.

References

Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, 665–674. ACM.

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 183–194. ACM.

Allamanis, M., and Sutton, C. 2013. Why, when, and what: analyzing stack overflow questions by topic, type, and code. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, 53–56. IEEE Press.

Althoff, T.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2014. How to ask for a favor: A case study on the success of altruistic requests. *ICWSM*.

Anderson, L. W.; Krathwohl, D. R.; and Bloom, B. S. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon.

Bach, F.; Mairal, J.; Ponce, J.; and Sapiro, G. 2010. Sparse coding and dictionary learning for image analysis. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*.

Bartlett, M. Y., and DeSteno, D. 2006. Gratitude and prosocial behavior helping when it costs you. *Psychological science* 17(4):319–325.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Brody, S., and Elhadad, N. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 804–812. Association for Computational Linguistics.

Carey, T. A., and Mullan, R. J. 2004. What is socratic questioning? *Psychotherapy: Theory, Research, Practice, Training* 41(3):217.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A computational approach to politeness with application to social factors. In *ACL*.

Dror, G.; Maarek, Y.; and Szpektor, I. 2013. Will my question be answered? predicting question answerability in community question-answering sites. In *Machine Learning and Knowledge Discovery in Databases*. Springer. 499–514.

Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R.; et al. 2004. Least angle regression. *The Annals of statistics* 32(2):407–499.

Elder, L., and Paul, R. 1998. The role of socratic questioning in thinking, teaching, and learning. *The Clearing House* 71(5):297–301.

Fyshe, A.; Wehbe, L.; Talukdar, P. P.; Murphy, B.; and Mitchell, T. M. 2015. A compositional and interpretable semantic space. *Proceedings of the NAACL-HLT, Denver, USA*.

Jaech, A.; Zayats, V.; Fang, H.; Ostendorf, M.; and Hajishirzi, H. 2015. Talking to the crowd: What do people react to in online discussions? *arXiv preprint arXiv:1507.02205*.

Jeon, J.; Croft, W. B.; Lee, J. H.; and Park, S. 2006. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 228–235. ACM.

King, A. 1994. Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American educational research journal* 31(2):338–368.

Klein, D., and Manning, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 423–430. Association for Computational Linguistics.

Li, B., and King, I. 2010. Routing questions to appropriate answerers in community question answering services. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 1585–1588. ACM.

Li, S., and Manandhar, S. 2011. Improving question recommendation by exploiting information need. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 1425–1434. Association for Computational Linguistics.

Li, B.; Jin, T.; Lyu, M. R.; King, I.; and Mak, B. 2012. Analyzing and predicting question quality in community question answering services. In *Proceedings of the 21st international conference companion on World Wide Web*, 775–782. ACM.

Liu, Y.; Bian, J.; and Agichtein, E. 2008. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 483–490. ACM.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2010. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research* 11:19–60.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Murphy, B.; Talukdar, P. P.; and Mitchell, T. M. 2012. Learning effective and interpretable semantic models using non-negative sparse embedding. In *COLING*, 1933–1950.

Nam, K. K.; Ackerman, M. S.; and Adamic, L. A. 2009. Questions in, knowledge in?: a study of naver's question answering community. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 779–788. ACM.

Newman, C. F. 2000. Hypotheticals in cognitive psychotherapy: Creative questions, novel answers, and therapeutic change. *Journal of Cognitive Psychotherapy* 14(2):135–147.

Ouyang, J., and McKeown, K. 2015. Modeling reportable events as turning points in narrative. In *EMNLP*.

Paul, R., and Elder, L. 2006. *Thinker's Guide to the Art of Socratic Questioning*. Foundation Critical Thinking.

Paul, R., and Elder, L. 2007. Critical thinking: The art of socratic questioning. *Journal of Developmental Education* 31(1):36.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* 12:1532–1543.

Ravi, S.; Pang, B.; Rastogi, V.; and Kumar, R. 2014. Great question! question quality in community q&a. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Ray, N. S. C. O. A., and Homan, P. C. M. 2015. An analysis of domestic abuse discourse on reddit. *anxiety* 4183:23300.

Ross, J. 2009. How to ask better questions. *Harvard Business Review Blogs. Viitattu* 8:2010.

Sammut, C., and Banerji, R. B. 1986. Learning concepts by asking questions. *Machine learning: An artificial intelligence approach* 2:167–192.

Shah, C., and Pomerantz, J. 2010. Evaluating and predicting answer quality in community qa. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 411–418. ACM.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *Proceedings of 25th International World Wide Web (WWW) Conference*.

Tsang, J.-A. 2006. Brief report gratitude and prosocial behaviour: An experimental test of gratitude. *Cognition & Emotion* 20(1):138–148.

Yang, L.; Bao, S.; Lin, Q.; Wu, X.; Han, D.; Su, Z.; and Yu, Y. 2011. Analyzing and predicting not-answered questions in community-based question answering services. In *AAAI*.