Identifying Informational vs. Conversational Questions on Community Question Answering Archives

Ido Guy, Victor Makarenkov, Niva Hazon, Lior Rokach, and Bracha Shapira
Ben-Gurion University of the Negev, Beer-Sheva, Israel
idoguy@acm.org,{makarenk,nivah}@post.bgu.ac.il,{liorrk,bshapira}@bgu.ac.il

ABSTRACT

Questions on community question answering websites usually reflect one of two intents: learning information or starting a conversation. In this paper, we revisit this fundamental classification task of informational versus conversational questions, which was originally introduced and studied in 2009. We use a substantially larger dataset of archived questions from Yahoo Answers, which includes the question's title, description, answers, and votes. We replicate the original experiments over this dataset, point out the common and different from the original results, and present a broad set of characteristics that distinguish the two question types. We also develop new classifiers that make use of additional data types, advanced machine learning, and a large dataset of unlabeled data, which achieve enhanced performance.

ACM Reference Format:
Ido Guy, Victor Makarenkov, Niva Hazon, Lior Rokach, and Bracha Shapira. 2018. Identifying Informational vs. Conversational Questions on Community Question Answering Archives. In *Proceedings of 11th ACM International Conf. on Web Search and Data Mining (WSDM 2018).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3159652.3159733

1 INTRODUCTION

Community question answering (CQA) websites have experienced a great deal of success for over a decade. Questions on CQA sites span many domains and involve a variety of user needs, which generally map into two principal types, defined by Harper et al. [20]. *Informational* questions are asked with the intent of getting information that the asker hopes to learn or use via fact- or advice-oriented answers. *Conversational* questions are asked with the intent of stimulating discussion and may be aimed at getting opinions or reflect an act of self-expression [20]. While studies examined other aspects of user intent on CQA websites, such as the subjectivity orientation [3, 12, 28] or the desire to socialize [12, 35], we focus on the informational-conversational paradigm, which has been widely referenced and used since originally introduced (e.g., [32, 34, 38]).

While the original paper focused on classifying questions at posting time, in order to improve question routing and automated tagging in real time, we consider CQA archives [54], allowing us to examine additional information that accumulates after the question has been posted, such as answers and votes. Many of the uses of CQA archives suit one of the two question types and can therefore benefit from automated classification. Supporting question retrieval [25] and serving CQA-intent queries on Web search [48], perhaps the two most common practices of CQA archives, fit the informational type. On the other hand, with the recent advancements in social media mining, conversational questions on the archive can be used to support a variety of applications such as opinion mining [40], controversy detection [14], and stance identification [49], and automated debating [16].

Harper et al. [20] examined three CQA websites: Yahoo Answers, Answerbag and Ask Metfilter. The latter was found to have only 5% conversational questions, while the first two were more balanced. Answerbag, however, has decreased in popularity and eventually shut down in 2015. In this work, we opted to focus on Yahoo Answers (YA), one of the largest CQA websites. In 2008, it was reported to account for 74% of CQA traffic [20], and while somewhat decreasing in popularity since, it still enjoys tens of thousands of questions posted every day, which allows to inspect it along a decade of existence. Aside from examining a long time period, we also experiment with substantially larger data: Harper et al.'s proprietary dataset included only 151 labeled questions from YA, while we build a dataset of over 4000 labeled questions, which we publicly release. To the best of our knowledge, it is the first public CQA dataset that labels questions as informational or conversational. Additionally, we examine the use of millions of unlabeled YA questions to further enhance performance. We believe that the larger data, the use of additional textual fields, and the advancement in machine learning methods, all warrant the revisit of the informational/conversational classification challenge.

Our experiments follow the "replicate & extend" approach [53] and include three main parts. In the first, we replicate the classifiers built by Harper et al. [20], which include category-based, text-based, and social network-based classifiers, as well as as an ensemble of all three. We report both extended descriptive results and classifier performance results over our own dataset, which, in part, substantially differ from those reported by Harper et al. [20]. In the second part, we extend Harper's classifiers by using the metadata and text in additional fields: the description, answers, and votes. We enhance the text-based classifier and add a fourth classifier, which is based on different types of metadata, to the ensemble. Overall, these extensions increase AUC by 3.6%. In the third part, we experiment with recurrent neural networks that make use of unlabeled data in two ways: for pre-training word embeddings and for semi-supervised learning using label propagation. This approach yields a particularly high AUC, at 8.7% over the baseline, even when based solely on question titles and when not used as part of an ensemble.

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

ACM ISBN 978-1-4503-5581-0/18/02…$15.00
https://doi.org/10.1145/3159652.3159733

2 RELATED WORK

Within CQA research, our work falls under the broad category of content modeling [45], which includes areas such as question quality estimation (e.g., [27]), answer quality ranking (e.g., [47]), question topic classification [10], and question type classification [31]. Some of the CQA research along the years has focused on specific types of questions such as factoids (e.g., [5, 18]), advice-seeking [7], how-to questions [46, 51], why questions [39], and opinion questions [30]. Some of these types have stronger association with the informational class (e.g., factoids, how-to), while others are more connected to the conversational class (e.g., opinion, why). A good summary of CQA research can be found in the recent survey by Srba and Bielikova [45].

More specifically, our work focuses on user intent classification on CQA. User intent has been extensively studied in Web information access, most prominently in Web search. The seminal work by Broder [9] distinguished between three main types of Web search queries: navigational, informational, and transactional. At a high level, all three map to the informational class in CQA, as Web search does not involve any aspect of conversation among users. Later works refined Broder's taxonomy and developed automatic classifiers to distinguish between the types (e.g., [8, 23, 55]). On the other hand, on social media, user intent typically revolves around sharing, interacting, conversing, and socializing [15, 24, 29]. To a certain extent, CQA websites combine both of these worlds, as they involve explicit user input in the form of a question, but also enable some level of user interaction.

Our work is based on the definitions of informational and conversational questions by Harper et al. [20]. Several studies examined user intent on question answering systems from other angles. Prominently, the distinction between objective and subjective questions was explored in several works [3, 12, 28]. While this distinction bears some similarity to the informational and conversational division, it is not the same: the subjective class includes, by definition [12], general advice, which is normally considered informational by Harper's definitions. For example, the question *"I am a Bangladeshi National girl and I came to USA on B1/B2 visa and now I would like to take admission pls adv?"*, given as an example for a subjective intent [12], is informational. Indeed, as reported over a YA dataset [28], only 34% of the questions were labeled as objective, while in Harper et al. [20] and our work, 61.2% and 55.6%, respectively, were labeled informational.

Mendes Rodrigues and Milic-Frayling [35] experimented with both YA and MSN QnA, a CQA website that was closed in 2009. They defined "social" intent for questions that are posted in order to informally engage and interact with other community users, as typically occurs in chatrooms. Social intent implies conversational intent, but the latter is far more comprehensive, also reflecting other needs, such as polling or discussing a topic, without socializing. Indeed, social intent was more common on MSN QnA, which enabled more intense user interaction through flexible comments and tagging, rendering a rich thread structure. A classifier deemed 6.5% of the questions on MSN QnA as social, while on YA this type's occurrence was "not sufficient to train a classifier" [35].

3 DATASET

Our dataset includes questions from Yahoo Answers, a large and diverse CQA website, which has been active for over a decade [2]. Questions on YA are spread across more than 1600 categories, in a taxonomy of up to 3 levels, with 26 categories at the top level [2]. Similarly to Harper et al. [20], we refer to the most general type of nodes in the hierarchy as "Top Level Categories" (TLCs) and to the most specific type of categories as "Low Level Categories" (LLCs).

The questions for our dataset were selected uniformly at random from the entire collection of YA non-deleted English questions, posted in the years 2006-2016. For our experiments, each question included, in addition to its title and timestamp, its description (81.5% had non-empty description), its TLC and LLC, and its list of answers with the answer's text, timestamp, and number of upvotes and downvotes per answer. In addition, each answer included a flag indicating whether it was selected as the "best answer" [2].

Questions in our dataset are manually labeled as either 'informational' or 'conversational'. In principle, as defined by Harper et al. [20], informational questions seek for facts or advice, while conversational questions pursue opinions, polling, or self-expression. As input for the labeling process, annotators were given the question's title, description, LLC, and TLC, similarly to Harper et al. [20].

At the first stage of the labeling process, two of the authors went through an iterative process to form more specific guidelines, annotating three batches of 100 questions each and carefully reviewing the disagreements. The level of agreement grew from 75% for the first batch to 93% for the third batch. The two authors then labeled a set of 1088 additional questions with a level of agreement of 93.75%, to which we refer as the *author dataset*. For our experiments, we used the 1020 questions for which there was an agreement between both annotators, similarly to Harper et al. [20]. As a few examples, the question "*Anyone knows how to get a sharpie stain off jeans?*" was labeled informational; the question "*Who thinks fur is murder?*" was labeled conversational; and for the question "*Why is Youtube so slow today?*", there was no agreement between the two annotators.

To increase the size of our data, we asked students of the "Introduction to Information Retrieval" course to label additional questions from our random sample. Each student labeled a batch of 105 questions. The batch included a benchmark of 5 questions for which the label was rather obvious. These 5 questions were randomly shuffled together with 100 "genuine" questions. Volunteering students were graded based on their success in labeling the 5 benchmark questions, while their grade accounted for a small portion (2%) of the course's grade. The students received detailed written guidelines, based on the insights gained by the two authors during their own annotation process, with definitions and examples for both types of questions. Each question was assigned to two volunteering students. Overall, 83 students completed their annotations, with 76 (91.6%) answering correctly all 5 benchmark questions. We discarded the input from the other 7 students. The level of agreement between the student annotators was 84.7% (Harper et al. [20] reported agreement of 87.1% between their annotators), yielding a total of 2996 questions whose label was agreed between the two annotators, which make up the *student dataset*.

The portion of questions labeled as informational is very similar between the author and the student datasets, at 55.9% and 55.4%, respectively. Harper et al. [20] reported a somewhat higher portion at 61.2%. The gap can be explained, to some extent, by the mild decrease in the portion of informational questions over the years, as we will later show. The *combined dataset*[1], which includes the questions from both the author and student datasets (4016 in total), has 55.6% of the questions labeled as informational.

In addition to the labeled dataset, we created a large unlabeled dataset by sampling 20 million questions uniformly at random from the entire YA archive of non-deleted English questions in 2006-2016. We elaborate on the use of this unlabeled dataset in Section 6.

4 BASELINES

In this section, we fully replicate the experiments conducted by Harper et al. [20] (henceforth referred to as "Harper" in short) with our own dataset. Our experiments consistently showed the best performance over the combined dataset, with no principal differences between the author dataset and the student dataset. We therefore focus on the combined dataset when reporting the results, and henceforth refer to it as *the experimental dataset* or merely *our dataset*. Similarly to Harper, we used the Weka workbench [19] to build all our classifiers and 5-fold cross validation to evaluate performance. As noted in Section 1, Harper experimented with 3 CQA sites. Their models were trained and evaluated separately for each CQA site. Since we do not have access to the original dataset, we present the results they reported for their YA dataset, which consisted of 151 questions.

We use the same three principal metrics as Harper for measuring classifiers' performance: (i) sensitivity – the proportion of conversational questions that are correctly classified; (ii) specificity – the proportion of informational questions that are correctly classified; and (iii) AUC - area under the ROC curve.

4.1 Category-based Classifier

The first classifier examined by Harper deems a question as informational or conversational based on its category. Table 1 presents the sensitivity, specificity, and AUC results for different variants we experimented with, and those reported by Harper et al. [20]. Like Harper, we used a Bayesian network for all variants. The first variant included TLCs only and achieved a slightly higher AUC than Harper, with higher sensitivity and lower specificity. Using LLCs only achieved further improvement in AUC, with both specificity and sensitivity higher than Harper. Using the combination of TLCs and LLCs, as done by Harper, improved the AUC very slightly compared to LLCs only. Finally, working with a dataset of a similar size to Harper (151 questions), to which we refer as the *151-Dataset*, yielded a slightly lower AUC than they reported.

Overall, we see that with our own dataset, an enhanced performance is obtained by the category-based classifier as compared to Harper. We conjecture that the main reason is that LLCs become more effective for classification when the dataset is larger, while with a smaller dataset they might be prone to sparsity.

Table 2 presents the distribution of TLCs in our dataset by their portion of informational (or conversational) questions. Harper reported these portions only for the 3 most common TLCs, due to data sparsity. By contrast to our findings, they reported Entertainment & Music to be substantially more conversational (20 of 26 questions, i.e., 76.9%), while similarly to us they found health (16 of 22, 72.7%), and Science & Mathematics (10 of 12, 83.3%) to be more informational.

Table 1: Performance of the category-based classifier.

	Sensitivity	Specificity	AUC
Harper et al. [20] (TLC+LLC)	0.660	0.820	0.810
Our dataset (TLC only)	0.671	0.792	0.823
Our dataset (LLC only)	0.695	0.861	0.847
Our Dataset (TLC+LLC)	0.732	0.827	**0.851**
151-Dataset (TLC+LLC)	0.768	0.732	0.798

Table 2: All 26 Top-level categories (TLCs) by the portion of informational (left, '%inf') or conversational (right, '%conv') questions, with the portion of questions of each TLC out of the total number of questions in our dataset ('%').

Top Informational TLCs			Top Conversational TLCs		
TLC	%	%inf	TLC	%	%conv
Computers & Internet	7.2%	92.7%	Family & Relationships	12.6%	95.8%
Science & Mathematics	7.1%	92.3%	Social Science	1.9%	80.0%
Home & Garden	1.3%	92.2%	News & Events	0.6%	73.1%
Yahoo! Products	2.0%	87.7%	Society & Culture	8.6%	72.8%
Cars & Transportation	2.0%	86.6%	Sports	3.2%	64.6%
Consumer Electronics	3.6%	83.9%	Politics & Government	5.4%	58.5%
Games & Recreation	3.1%	83.7%	Arts & Humanities	2.8%	55.3%
Environment	0.1%	75.0%	Beauty & Style	5.0%	50.5%
Business & Finance	2.7%	75.0%	Dining Out	0.1%	50.0%
Travel	2.1%	75.0%	Pregnancy & Parenting	4.1%	48.8%
Pets	2.8%	73.9%	Entertainment & Music	7.4%	48.3%
Health	7.8%	71.0%	Food & Drink	1.4%	36.4%
Education & Reference	4.6%	69.4%	Local Businesses	0.3%	30.8%

Table 3: Low-level categories (LLCs) with highest percentage of informational (left, '%inf') and conversational (right, '%conv') questions, with the portion of questions of each LLC out of the total number of questions in our dataset ('%').

Top Informational LLCs			Top Conversational LLCs		
LLC	%	%inf	LLC	%	%conv
Chemistry	1.2%	100.0%	Singles & Dating	8.7%	98.0%
Mathematics	2.6%	98.1%	Marriage & Divorce	0.8%	93.5%
Programming & Design	0.9%	97.1%	Psychology	0.7%	93.3%
Software	1.2%	95.8%	Friends	1.5%	91.9%
Maintenance & Repairs	0.9%	91.7%	Politics	2.1%	90.6%
Laptops & Notebooks	0.8%	90.9%	Baby Names	0.9%	89.2%
Video & Online Games	2.2%	86.7%	Religion & Spirituality	3.5%	81.7%
Homework Help	1.2%	85.7%	Other - Society & Culture	1.1%	79.1%
Cell Phones & Plans	1.1%	84.8%	Mental Health	0.8%	76.5%
Biology	1.0%	82.5%	Other - Beauty & Style	0.9%	75.0%

Table 3 presents the top 10 LLCs with the highest portion of informational questions and the top 10 LLCs with the highest portion of conversational questions (only LLCs that account for at least 0.5% of the questions were considered). The informational list includes different domains of science, computers/cellphones, and homework help. The conversational list is more diverse and includes the popular LLC Singles & Dating, with others ranging from psychology to politics and from baby names to mental health.

[1] http://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=82

Table 4: Performance of the text-based classifier.

	Sensitivity	Specificity	AUC
Harper et al. [20]	0.480	0.710	0.600
Our dataset	0.667	0.845	**0.756**
151-Dataset	0.710	0.732	0.733

4.2 Text-based Classifier

The text classifier used by Harper et al. [20] was based on the 500 most common unigrams or bigrams occurring in titles of each of the two question types (in a lower-case form). For classification, Weka's sequential minimum optimization (SMO) algorithm for training support vector machines (SVM), with linear kernel, was used. In this case, they used all three datasets in conjunction for training.

Table 4 presents the classification results as reported by Harper and as obtained by applying the exact same method over our dataset and over the 151-Dataset. A substantially higher performance was achieved on our dataset. This can be associated with the larger data, since all features are sparse. Yet, even over our 151-Dataset the results are still substantially higher than the baseline. We believe this attests to the importance of training the data on YA text only, rather than a mix of questions from three different CQA sites.

In terms of prominent unigrams/bigram features, Harper reported their information gain and their occurrence portions on both informational and conversational questions, across questions words and for the pronouns "you" and "I". They also reported the four additional unigrams/bigrams with the highest information gain for the informational and for the conversational class, respectively. They did not present a specific report for YA, but rather joined the statistics across their three datasets. In Table 5, we report the same statistics on our own dataset, alongside the reported statistics by Harper, when such exist.

Observing the question words, similarly to Harper, we found that "how" and "where" were substantially more common on informational questions, while "why" was more common on conversational. Indeed, why-questions are considered more open ended [21] and may thus more often reflect a need for discussion. In addition, we also identified some signal on "who" towards conversational questions, whereas Harper found it to be slightly more common (with no information gain) on informational questions. The most common question word, "what", did not have any information gain, and so did "when" and "which".

As for "you" versus "I", like Harper we also found a strong signal in "you" for conversational questions, while "I" was more common on informational questions, albeit with a relatively small information gain.

Inspecting the top informational tokens, only one of them, "can", was also among the top four reported by Harper. The other three Harper reported were "is there", "help", and "do I". The last two were also more common on the informational class in our dataset, with a small information gain (0.002 and 0.004, respectively), while the first was not found to have any information gain. The top conversational tokens include two of the top-4 found by Harper. The word "think" had the most information gain out of all tokens in our dataset, and was far more common on conversational questions. Harper also reported "would you" and "is your" among their top four conversational tokens. The former was also found to have a substantial information gain for the conversational class in our dataset (0.006), while the latter posed a more modest information gain for conversational questions (0.001).

Table 5: Portions of informational and conversational question titles containing the following types of tokens, as well as the token's information gain for the text-based classifier: (i) question words; (ii) personal pronouns; and other tokens with highest information gain for predicting the (iii) informational class and (iv) conversational class.

	Our Dataset			Harper et al. [20]		
	% Inf.	% Conv.	Info Gain	% Inf.	% Conv.	Info Gain
how	19.6%	9.5%	0.014	29.8%	11.3%	0.029
where	5.0%	0.8%	0.012	15.5%	1.9%	0.033
why	3.8%	8.7%	0.007	5.6%	17.0%	0.012
who	2.0%	3.8%	0.002	13.9%	11.3%	0
what	16.9%	17.7%	0	30.2%	30.4%	0
when	3.9%	3.9%	0	17.0%	13.2%	0
which	1.9%	2.1%	0	n/a	n/a	n/a
you	8.1%	22.2%	0.024	25.8%	54.7%	0.05
I	31.3%	24.9%	0.003	68.6%	27.4%	0.124
can	14.6%	4.5%	0.02	35.1%	9.4%	0.069
what is	6.2%	1.7%	0.01	n/a	n/a	n/a
find	3.1%	0.6%	0.007	n/a	n/a	n/a
how do	6.3%	2.6%	0.006	n/a	n/a	n/a
think	0.1%	6.6%	0.031	n/a	n/a	n/a
you think	0.0%	4.8%	0.023	1.3%	8.2%	0.021
do you	2.6%	9.8%	0.017	4.9%	22.0%	0.047
what do	0.6%	4.5%	0.013	n/a	n/a	n/a

4.3 Social Network-based Classifier

The third classifier used by Harper et al. [20] was based on social network (SN) properties of the question asker. Three features were used to represent the asker's social network signature [52], i.e., a representation of their egocentric network based on previous question asking and answering on the site. Specifically, they constructed a directed weighted graph based on a YA dataset that spanned a period of 49 days, with over 1.5M users and 4.3M questions[2]. The graph's vertices represented users and directed edges represented the act of one user answering another user's question. The three features that were calculated based on this graph for each question asker were: (i) *NUM_NEIGHBORS*: the number of neighbors of the asker ; (ii) *PCT_ANSWERS*: the asker's number of outgoing edges divided by the asker's total number of edges ; and (iii) *CLUST_COEFFICIENT*: the clustering coefficient [50] of the asker's ego network, reflecting its connectivity (the portion of the asker's neighbor pairs who are connected themselves).

For our own SN classier, we computed the same three features based on the entire YA repository rather than a partial dataset of 49 days. Like Harper, we calculated each feature with respect to the timestamp of the respective question, ensuring that we have an accurate snapshot of interactions up to the particular time the question was asked. Results, presented in Table 6, indicate that our own classifier did not reach the performance reported by Harper. This is in spite of the fact it was trained on a much larger dataset

[2] The dataset was not publicly released.

Table 6: Performance of the SN-based classifier, with different classifiers and feature subsets. When not otherwise stated, Bayesian network was used for classification.

	Sensitivity	Specificity	AUC
Harper et al. [20] (all features)	0.710	0.870	**0.810**
SVM (linear kernel) (all features)	0.572	0.620	0.596
Random forest (all features)	0.491	0.710	0.635
C4.5 decision tree (all features)	0.339	0.848	0.633
Logistic regression (all features)	0.412	0.796	0.653
Bayesian network (all features)	0.621	0.631	0.659
PCT_ANSWERS +CLUST_COEFFICIENT	0.658	0.553	0.639
NUM_NEIGHBORS +PCT_ANSWERS	0.551	0.674	0.645
NUM_NEIGHBORS +CLUST_COEFFICIENT	0.620	0.626	0.657
151-Dataset (all features)	0.058	0.915	0.462

and that the features were calculated over the entire YA graph[3]. To further explore this, we experimented with several other common classifiers using Weka [19], yet as Table 6 shows, they all performed worse than the Bayesian network. Inspecting the performance using each subset of two out of the three features (ablation tests) indicated that the combination of *NUM_NEIGHBORS* and *CLUST_COEFFICIENT* reached almost the same performance as all three features, i.e., *PCT_ANSWERS* did not contribute much to the overall performance. Finally, running on top of the 151-Dataset demonstrated poor performance.

Table 7 shows the statistics of the three SN features, in a similar manner as presented by Harper. The trends on our data are similar to those reported by Harper: there is a statistically significant gap in favor of conversational question askers for all three features, i.e., they tend to have more neighbors, higher ratio of answers to questions, and tighter connectivity of their neighbors. As expected, the number of neighbors is higher on our dataset, since we consider a larger time period. Statistical significance is stronger on our data, as could be expected given its size.

Overall, this analysis indicates that on a descriptive level, we observe the same differences between informational and conversational questions, validating the finding that askers of conversational questions tend to have a larger, more tightly interconnected ego network [20]. Yet, trying to use these differences as discriminative features yields a substantially weaker signal than reported by Harper[4]. The immediate suspicion that arises is that the SN model developed by Harper suffered from an overfitting issue. The small size of their dataset may help explain it, although we could not replicate the results when using our own dataset of the same size. Harper did not report the number of users in their dataset, nor the method used for creating the dataset, but it could be that multiple questions on the training and test set had the same askers, which led to memorizing their SN features in the model.

Table 7: Average and standard deviation of the three SN features for informational vs. conversational questions, as well as the p value of a two-tailed unpaired t-test.

	Our Dataset			Harper et al. [20]		
	avg inf.	avg conv.	t-test	avg inf.	avg conf.	t-test
NUM_NEIGHBORS	400	1204	$p < 0.0001$	252	757	$p < 0.01$
PCT_ANSWERS	0.38	0.52	$p < 0.0001$	0.26	0.32	$p = 0.02$
CLUST_COEFFICIENT	0.015	0.028	$p < 0.0001$	0.06	0.15	$p < 0.01$

Table 8: Diversity between classifier pairs.

Classifier pair	Our dataset		Harper et al. [20]
	Yule's Q	Agreement	Yule's Q
Categories—Text	0.72	71.8%	0.31
Categories–SN	0.45	61.8%	0.72
Text – SN	0.3	57.8%	0.58

Table 9: Performance of the 3-way ensemble classifier.

	Sensitivity	Specificity	AUC
Harper et al. [20]	0.780	0.950	**0.950**
Our dataset	0.754	0.866	0.886

4.4 3-Way Ensemble

Under the premise that the features introduced by the three classifiers (categories, text, and SN) complement one another, at least to some extent, Harper examined an ensemble classifier (meta classifier) that uses the output scores (i.e., confidence scores that the question is conversational) of the three classifiers as its features to learn one ultimate classification. To measure the diversity among the three output scores, Harper used the Yule's Q metric, which ranges from −1 to 1 [26]. Classifiers that tend to categorize the same instances correctly pose more positive values, while classifiers that tend to categorize different instances incorrectly take more negative values. Table 8 presents the Yule's Q metric for each pair of classifiers; for our own data, we also present the mere level of *agreement* between each pair, i.e., the portion of instances for which the prediction was identical.

Our results in terms of Yule's Q are quite different than Harper's. This might be attributed to the differences in performance for the text and SN classifiers. As opposed to Harper, we found a rather high agreement between the category and text classifiers. We believe this makes sense, since categories are often characterized by their own language [11]. On the other hand, we found a lower agreement between the SN classifier and the other two classifiers, likely due to the lower performance of the SN classifier.

Table 9 presents the results of the "*3-way ensemble*" classifier. Like Harper, we implemented it using JMP's neural network algorithm. It can be seen that the performance they reported could not be attained in our own settings, most probably due to the lower performance of the SN classifier. From this point onward, we refer to the 3-way ensemble result achieved on our data as our baseline, since it achieved the highest performance among all the methods reported by Harper when ran on our dataset.

5 IMPROVING THE BASELINE

In this section, we evaluate two complementary methods for improving the baseline reported in the previous section. First, we use

[3] We also experimented with calculating the SN features based on 49 days, but the results were lower.

[4] Our experiments over the author and the student datasets separately yielded very similar performance and descriptive results. For example, the AUC for the SN-based Bayesian network classier was 0.651 and 0.643 for the two datasets, respectively

Table 10: Performance of n-gram SVM with linear kernel, using different combinations of the question's textual fields. Unless otherwise noted, unigram, bigram, and trigram features were used.

	Sensitivity	Specificity	AUC
Title	0.718	0.805	0.761
Description	0.600	0.826	0.713
Answers	0.533	0.716	0.625
Best answer	0.560	0.792	0.676
Title+description	0.730	0.838	0.784
Title+answers	0.657	0.786	0.721
Title+best answer	0.747	0.821	0.784
Title+description+answers	0.691	0.830	0.761
Title+description+best answer	0.743	0.844	**0.794**
Title+description+best answer (unigrams only)	0.719	0.808	0.764
Title+description+best answer (unigrams + bigrams)	0.742	0.819	0.780

lexical features from fields beyond the title, i.e., the description and the answers, assessing if their language can also help distinguish between the two classes. Second, we experiment with a set of metadata features, mostly based on description, answers, and votes, aiming to find more discriminative characteristics between the classes. For these experiments, we used the scikit-learn library [41] to run machine learning algorithms and the NLTK suite [6] for natural language processing. We continued using our combined dataset for experimentation and 5-fold cross validation for evaluation.

5.1 N-gram Classifier

Our first extension focused on the text-based classifier. As we observed good results for the title-based text classifier (Section 4.2), we set out to examine if other textual fields, specifically, the question's description and answers, can further assist in the classification task. We experimented with unigram, bigram, and trigram features, pruned based on an occurrence threshold of 2. As a classifier, we continued to use SVM with linear kernel.

The first four rows of Table 10 indicate that the title is the most useful field for classification, achieving the best AUC (similar to the baseline, up 0.6%), followed by the description, and finally the answers. Using the text of the best answer only, even though it is not selected for nearly 15% of the questions, was substantially more productive than using the text of all the answers. Apparently, all answers' text introduces a level of diversity and noise that masks the class of the question, while the best answer is more indicative.

We also experimented with combining the text from two or three fields. It can be seen that the combinations of title and description, as well as title and best answer, achieved a higher AUC compared to the tile alone (+3.0% for both). By contrast, combining title with all answers degraded the AUC (−5.3%). Using the threesome of title, description, and best answer further improved the performance (+1.3%). Following, using the n-gram classifier based on title, description, and best answer, instead of the original text-based classier, as part of the 3-way ensemble (Section 4.4), yielded an increase in performance: AUC of 0.904, up 2.0% compared to the ensemble baseline (Table 9). Throughout our experimentation, we observed that using unigrams, bigrams, and trigrams achieved higher performance than using unigrams and bigrams, which in turn was higher than unigrams only. The bottom section of Table 10 demonstrates this for the combination of title, description, and best answer.

Table 11: Performance of different n-gram classifiers based on the combination of title, description, and best answer on their own and as part of a 3-way ensemble also including category and SN classifiers.

	n-gram classifier only			3-way ensemble		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
Random forest	0.564	0.876	0.720	0.778	0.836	0.885
Gradient boosting	0.633	0.883	0.758	0.789	0.855	0.903
SVM (linear kernel)	0.743	0.844	0.794	0.796	0.861	0.904
Logistic regression	0.745	0.877	0.811	0.813	0.869	**0.914**

Thus far, we experimented with an SVM classifier, similarly to Harper. For the title, description, and best answer variant, we examined other classifiers, as reported in Table 11. It can be seen that logistic regression attained the best performance and surpassed that of SVM. Using the logistic regression version for the ensemble classifier further increased its performance to 0.914 (up 1.2%, to a total of +3.2% compared to the ensemble baseline). We also ran logistic regression with other field combinations as detailed in Table 10 and the respective results were similar to those reported in the table for SVM, with the combination of title, description, and best answer producing the best outcome.

Inspecting the n-grams with most information gain for fields other than the title (which is reported in detail in Section 4.2), we observed that for the informational class, descriptions included explicit calls for assistance, such as "does anyone know" and "please help", while best answers included verbs, such as "use", "click",or "check". For conversational questions, "think" had high information gain across all fields (titles, descriptions, and answers), while best answers included adverbs such as "really", "very", and "just".

5.2 Additional Features

As another step to improve the performance of our classifier, we set out to examine additional basic features that were not inspected by Harper. We generally refer to these features as "metadata" features.

Table 12 lists the key features and their statistics for informational versus conversational questions. Some characteristics, such as the length of the title or the portion of questions for which a best answer was selected, are similar for both classes, while other pose fundamental differences. For example, conversational questions more often include a description and when they do, it tends to be longer. Among the most distinctive features are the number of answers and number of votes (both upvotes and downvotes), which are significantly higher for conversational questions. Votes per answer are also higher for conversational questions. Interestingly, while answer length is slightly higher for informational questions, the best-answer length is somewhat higher for conversational questions. While URLs are more common on conversational titles or descriptions, they are substantially more common on informational best answers. Finally, while informational questions have substantially fewer answers and votes, the time span from question posting to the most recent answer/vote is comparable between the classes.

We also used the question's posting year as a feature. As Figure 1 indicates, the portion of informational questions slightly decreases over the years. One reason may be that Web search engines started

Table 12: Metadata feature statistics (average, standard deviation, median, maximum; or respective portions) for informational and conversational questions. Features with statistically significant average gaps, based on a two-tailed unpaired t-test, are marked with * ($p<0.01$) or ** ($p<0.001$).

	Informational				Conversational			
	Avg	Std	Med	Max	Avg	Std	Med	Max
Title length (words)	12.0	5.8	11	62	11.7	5.5	10	56
Description length (words)**	46.3	55.3	32	605	83.5	100.1	52	1016
Number of answers**	3.6	3.5	3	43	6.8	6.8	5	85
Answer length (words)*	60.3	99.5	32	3141	55.8	88.2	30	2725
Best-answer length (words)*	64.7	93.3	33	946	69.2	106.1	35	1718
Total up-votes**	1.4	4.8	0	121	5.5	15.6	1	291
Total down-votes**	1.1	4.1	0	68	4.0	11.2	0	158
Up-votes per answer**	0.43	0.82	0	14	0.73	1.14	0.33	21
Down-votes per answer*	0.30	0.62	0	5.33	0.49	0.79	0	7
Time span to last answer (days)	412	951	0.52	3768	379	920	0.51	3796
Time span to last vote (days)	86	394	0	3478	80	347	0.44	3744
Description exists	76.8%				85.6%			
No answer	2.9%				1.8%			
No votes	56.3%				39.7%			
Best answer selected	85.0%				86.2%			
Title/desc. contain URL	3.2%				5.9%			
Best answer contains URL	11.7%				4.0%			

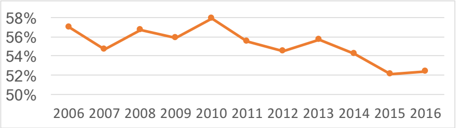

Figure 1: Percentage of informational questions by year.

Table 13: Performance of different metadata classifiers on their own and as part of a 4-way ensemble also including category, n-gram, and SN classifiers.

	metadata classifier only			4-way ensemble		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
SVM (linear kernel)	0.369	0.769	0.569	0.815	0.861	0.912
Random forest	0.475	0.794	0.634	0.818	0.863	0.913
Logistic regression	0.504	0.848	0.676	0.811	0.871	0.915
Gradient boosting	0.545	0.814	0.680	0.818	0.872	0.918

to provide better treatment to informational queries by including direct answers on their results page [4, 17].

We used the features above to build another classifier, the *metadata classifier*, for informational versus conversational questions. Results, depicted in Table 13, indicate that the best performance was achieved by gradient boosting (with logistic regression at close second). This classifier did not reach the high performance attained by the category and text classifiers, but surpassed the performance (as we could reproduce) of the SN classifier. Using it in a *4-way ensemble* with the n-gram, category, and SN classifiers, reached a slightly higher performance than the 3-way ensemble reported in Table 11, at 0.918 (+0.44%).

Inspecting the performance of the metadata classifier when excluding different feature families (ablation tests) indicated that the description and answers metadata features were the most important, as their removal led to a decrease of 5.7% and 5.3%, respectively, in AUC. On the other hand, title, votes, and year features did not pose any substantial contribution to performance.

6 LSTM NETWORK MODEL

Thus far, we relied on n-gram features for text-based classification. In this section, we describe our experimentation with recurrent neural networks (RNNs) with long short term memory (LSTM) [22] for the classification task. RNNs with LSTM can capture more complex word relations and have recently demonstrated high performance in a variety of text processing tasks [36]. In our experiments, we lower-cased and tokenized the text and then sequentially fed it, from left to right, into the model. Similarly to Section 5.1, we experimented with title-only, title+description, title+description+answers, and title+description+best answer as our text input. To avoid overfitting, we used three dropout strategies for regularization: between the embedding layer and the recurrent LSTM layer; between the hidden layers of the LSTM recurrent layers ; and between the LSTM and the output used for classification.

We implemented all LSTM models using Tensorflow [1] with the Keras API [13]. As in Section 5, we used scikit-learn [41] with 5-fold cross validation for evaluation and NLTK [6] for text processing. We performed extensive hyper-parameter tuning, which included the number of LSTM layers, the size of the LSTM memory cell, the embedding size, the number of iterations over the entire training set , the dropout rate for each of the three types mentioned above, and the batch size of the stochastic gradient descent used for optimization. We also experimented with a bi-directional RNN with LSTM, however this model did not yield any performance gain.

As deep learning techniques typically demonstrate high performance on large data, we set out to explore the use of the large unlabeled dataset of 20M YA questions, described in Section 3. We examined two ways for leveraging the unlabeled data: pre-trained word embedding and label propagation.

6.1 Pre-trained Word Embedding

The first layer of the neural network generates a latent representation of the text in a lower-dimensional space. Instead of letting the network learn this representation in training time, based on the labeled dataset, we can learn the word embeddings in a pre-processing step [33]. Since word embedding does not require human-annotated data, it can be learned over a much larger dataset. We trained skip-gram negative sampling (SGNS) [37] word embeddings over our unlabeled dataset using the Gensim library [44], experimenting with three variants of the text used for training: titles only; titles and descriptions; and titles, descriptions, and answers. In all cases, performance differences among the three were minor, with title-only achieving slightly higher performance than the others. We therefore report only title-based embeddings, which also require less storage and computation power. To further inspect the value of the unlabeled YA dataset, we also examined the use of the popular general-purpose publicly-available GloVE [42], pre-trained on the Wikipedia 2014 and Gigaoword 5 corpora.

Table 14 summarizes the results of the LSTM model, both on its own and when used as part of a 4-way ensemble with the category, SN, and metadata classifiers, as described in Section 5.2. Inspecting the results of the LSTM model alone, it can be seen that using GloVE

Table 14: Performance of LSTM model with combinations of the question's textual fields, without pre-trained word embeddings ("No Embed") and with pre-trained word embeddings using global vectors for word representation ("GloVe") and our unlabeled YA dataset ("YA").

		LSTM classifier only			4-way ensemble		
		Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
No Embed	Title	0.753	0.794	0.856	0.792	0.861	0.910
	Title+description	0.754	0.797	0.857	0.776	0.860	0.902
	Title+description+answers	0.470	0.609	0.562	0.741	0.849	0.871
	Title+description+best answer	0.730	0.770	0.829	0.765	0.864	0.894
GloVe	Title	0.780	0.807	0.885	0.800	0.857	0.913
	Title+description	0.780	0.821	0.880	0.795	0.856	0.909
	Title+description+answers	0.558	0.667	0.663	0.738	0.848	0.873
	Title+description+best answer	0.823	0.693	0.842	0.773	0.852	0.894
YA	Title	0.794	0.836	0.907	0.823	0.867	0.923
	Title+description	0.837	0.847	0.915	0.840	0.871	**0.926**
	Title+description+answers	0.695	0.629	0.713	0.745	0.847	0.878
	Title+description+best answer	0.866	0.793	0.906	0.817	0.855	0.918

pre-trained embeddings yielded a noticeable performance gain. Using our unlabeled YA dataset for embedding improved the results further. Across the board, title and title+description achieved the highest performance. As was the case with the n-gram models (Section 5.1), adding all answers' text to the input degraded performance, while in the case of LSTM, the best answer was again preferable to all answers, but did not improve over title or title+description. Particularly, the LSTM model based on title+description and YA pre-trained embeddings achieved the highest performance, which was substantially higher than the n-gram best classifier reported in Table 11 (+12.8% in AUC). Despite the large performance gain, using this classifier as part of a 4-way ensemble only slightly increased the AUC reported for the best performing 4-way ensemble in Table 13, bringing it to 0.926 (+0.87%).

6.2 Label Propagation

Our second attempt to leverage the unlabeled data was by using it for the complete model training, rather than for pre-trained embeddings only. Many different techniques for semi-supervised learning, which use a relatively small seed set of labeled examples and a larger set of unlabeled examples, have been proposed [43]. We opted to use *label propagation* (LP), which simply and iteratively increases the training set with examples from the unlabeled dataset, based on their predicted probability scores [56]. We started with the labeled dataset of 4016 questions and trained an initial model, using LSTM with pre-trained YA embeddings. We then selected 2000 questions uniformly at random from the unlabeled dataset and classified them using the trained model. In the next iteration, we used the questions that were classified with high confidence as part of the training set. After experimenting with various confidence thresholds, we opted to set it to 0.9, i.e., questions classified with a probability of 0.9 or higher were added to the training set in the next iteration, labeled with their predicted class. We repeated the process for 100 iterations, covering 1% of the unlabeled dataset, as performance started to converge. We experimented with title and title+description for the input text, as they yielded the best performance to begin with (Table 14).

Table 15: Performance of LSTM model with YA pre-trained embeddings after 100 iterations of label propagation.

	LSTM classifier only			4-way ensemble		
	Sensitivity	Specificity	AUC	Sensitivity	Specificity	AUC
Title	0.905	0.907	**0.963**	0.907	0.909	0.952
Title+description	0.933	0.822	0.947	0.921	0.849	0.936

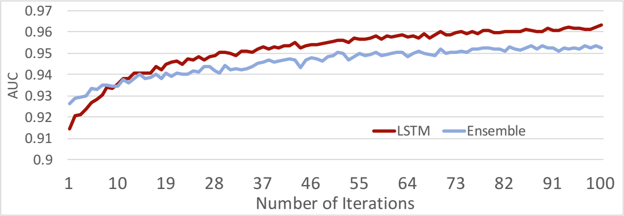

Figure 2: AUC performance of LSTM (with YA pre-trained embeddings) and 4-way ensemble by the number of label propagation iterations.

Results, presented in Table 15, show that the LP process produced a considerable performance gain. Using the results of the LP process as part of a 4-way ensemble did not yield a substantial improvement in performance and even led to a decline in AUC. Overall, running the LSTM model with LP based solely on titles yielded the best AUC, at +4.0% compared to the best performing model reported in Table 14. Figure 2 illustrates the increase in AUC with the number of LP iterations, both for the standalone LTSM model and the 4-way ensemble. After 10 iterations, the AUC of the LTSM model bypasses the ensemble.

7 CONCLUSION

We experimented with a variety of methods for improving the classification of questions on CQA archives as either informational or conversational. While supervised approaches achieved some performance gain relative to the baseline (up to 4.5% in AUC), semi-supervised learning using LP and a large unlabeled dataset yielded a more substantial increase (+8.7% compared to the baseline), up to an AUC of over 0.96, with both specificity and sensitivity above 90%. To the best of our knowledge, these are the highest reported results for intent classification on CQA websites. It should be noted, however, that this approach requires not only large volumes of unlabeled data, but also significant computation power[5].

The LP approach was found to work best when using only question titles as input text. Titles were also found most effective for pre-training the word embeddings. It appears that when large volumes of data are in hand, titles alone best reflect the differences between the two classes. Yet, when only relying on the labeled data, the text of the description, and in some cases the best answer, was found to be helpful for achieving better performance.

As part of our evaluation, we replicated the experiments reported by Harper et al. [20] over our larger YA dataset. For the category and text classifiers, we were able to reproduce the results and even achieve higher performance, as could be expected given the larger

[5] LP over 200K question titles ran for 48 hours on Nvidia GeForce GTX 1080 Ti GPU.

training data. For the SN classifier, however, despite observing similar descriptive differences for the features, we could not reproduce the reported performance. We therefore suggest to carefully refer to the results reported for this classifier in the original paper [20].

For future work, we intend to examine additional graph metrics that may enhance classification based on social network features. We also plan to examine additional semi-supervised learning approaches, such as co-training using text and network features.

REFERENCES

[1] Martín Abadi et al. 2016. TensorFlow: A System for Large-scale Machine Learning. In *Proc.of OSDI.* 265–283.

[2] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proc. of WWW.* 665–674.

[3] Naoyoshi Aikawa, Tetsuya Sakai, and Hayato Yamana. 2011. Community QA question classification: Is the asker looking for subjective answers or not? *IPSJ Online Transactions* 4 (2011), 160–168.

[4] Michael S. Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct Answers for Search Queries in the Long Tail. In *Proc. of CHI.* 237–246.

[5] Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media. In *Proc. of WWW.* 467–476.

[6] Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proc. of COLING-ACL.* 69–72.

[7] Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. 2016. Supporting human answers for advice-seeking questions in CQA sites. In *Proc. of ECIR.* 129–141.

[8] David J. Brenes, Daniel Gayo-Avello, and Kilian Pérez-González. 2009. Survey and Evaluation of Query Intent Detection Methods. In *Proc. of WSCD.* 1–7.

[9] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (2002), 3–10.

[10] Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. 2011. Large-scale Question Classification in cQA by Leveraging Wikipedia Semantic Knowledge. In *Proc. of CIKM.* 1321–1330.

[11] Wen Chan, Weidong Yang, Jinhui Tang, Jintao Du, Xiangdong Zhou, and Wei Wang. 2013. Community Question Topic Categorization via Hierarchical Kernelized Classification. In *Proc. of CIKM.* 959–968.

[12] Long Chen, Dell Zhang, and Levene Mark. 2012. Understanding user intent in community question answering. In *Proc. of WWW Companion.* 823–828.

[13] François Chollet. 2015. Keras. https://github.com/fchollet/keras. (2015).

[14] Shiri Dori-Hacohen and James Allan. 2015. Automated controversy detection on the web. In *Proc. of ECIR.* 423–434.

[15] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. 2007. The benefits of Facebook friends: Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication* 12, 4 (2007), 1143–1168.

[16] Iryna Gurevych, Eduard H Hovy, Noam Slonim, and Benno Stein. 2016. Debating Technologies. In *Dagstuhl Reports*, Vol. 5.

[17] Ido Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *Proc. of SIGIR.* 35–44.

[18] Ido Guy and Dan Pelleg. 2016. The Factoid Queries Collection. In *Proc. of SIGIR.* 717–720.

[19] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (2009), 10–18.

[20] F. Maxwell Harper, Daniel Moy, and Joseph A. Konstan. 2009. Facts or Friends?: Distinguishing Informational and Conversational Questions in Social Q&A Sites. In *Proc. of CHI.* 759–768.

[21] Jaakko Hintikka and Ilpo Halonen. 1995. Semantics and pragmatics for why-questions. *The Journal of Philosophy* 92, 12 (1995), 636–657.

[22] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780.

[23] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2007. Determining the User Intent of Web Search Engine Queries. In *Proc. of WWW.* 1149–1150.

[24] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proc. of WebKDD/SNA-KDD.* 56–65.

[25] Jiwoon Jeon, W Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *Proc. of CIKM.* 84–90.

[26] Ludmila I. Kuncheva and Christopher J. Whitaker. 2003. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning* 51, 2 (2003), 181–207.

[27] Baichuan Li, Tan Jin, Michael R. Lyu, Irwin King, and Barley Mak. 2012. Analyzing and Predicting Question Quality in Community Question Answering Services. In *Proc. of WWW Companion.* 775–782.

[28] Baoli Li, Yandong Liu, and Eugene Agichtein. 2008. CoCQA: Co-training over Questions and Answers with an Application to Predicting Question Subjectivity Orientation. In *Proc. EMNLP.* 937–946.

[29] Kuan-Yu Lin and Hsi-Peng Lu. 2011. Why people use social networking sites: An empirical study integrating network externalities and motivation theory. *Computers in human behavior* 27, 3 (2011), 1152–1161.

[30] Yandong Liu and Eugene Agichtein. 2008. On the Evolution of the Yahoo! Answers QA Community. In *Proc. of SIGIR.* 737–738.

[31] Yandong Liu, Nitya Narasimhan, Venu Vasudevan, and Eugene Agichtein. 2009. Is This Urgent?: Exploring Time-sensitive Information Needs in Collaborative Question Answering. In *Proc. of SIGIR.* 712–713.

[32] Zhe Liu and Bernard J. Jansen. 2015. A Taxonomy for Classifying Questions Asked in Social Question and Answering. In *Proc. of CHI EA '15.* 1947–1952.

[33] Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sören Auer. 2017. Neural Network-based Question Answering over Knowledge Graphs on Word and Character Level. In *Proc. of WWW.* 1211–1220.

[34] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design Lessons from the Fastest Q&A Site in the West. In *Proc. of CHI.* 2857–2866.

[35] Eduarda Mendes Rodrigues and Natasa Milic-Frayling. 2009. Socializing or Knowledge Sharing?: Characterizing Social Intent in Community Question Answering. In *Proc. of CIKM.* 1127–1136.

[36] Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural Variational Inference for Text Processing. In *Proc. of ICML.* 1727–1736.

[37] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint* abs/1301.3781 (2013).

[38] Liqiang Nie, Meng Wang, Yue Gao, Zheng-Jun Zha, and Tat-Seng Chua. 2013. Beyond text QA: multimedia answer generation by harvesting web information. *IEEE Transactions on Multimedia* 15, 2 (2013), 426–441.

[39] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Jun'ichi Kazama, and Yiou Wang. 2012. Why Question Answering Using Sentiment Analysis and Word Classes. In *Proc. EMNLP-CoNLL.* 368–378.

[40] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 1–135.

[41] Fabian Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12 (Nov. 2011), 2825–2830.

[42] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*, Vol. 14. 1532–1543.

[43] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. 2015. Semi-supervised Learning with Ladder Networks. In *Proc. of NIPS.* 3546–3554.

[44] Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proc. of LREC 2010 Workshop on New Challenges for NLP Frameworks.* 45–50.

[45] Ivan Srba and Maria Bielikova. 2016. A Comprehensive Survey and Classification of Approaches for Community Question Answering. *ACM Trans. Web* 10, 3 (Aug. 2016), 18:1–18:63.

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

[47] Hapnes Toba, Zhao-Yan Ming, Mirna Adriani, and Tat-Seng Chua. 2014. Discovering High Quality Answers in Community Question Answering Archives Using a Hierarchy of Classifiers. *Inf. Sci.* 261 (2014), 101–115.

[48] Gilad Tsur, Yuval Pinter, Idan Szpektor, and David Carmel. 2016. Identifying Web Queries with Question Intent. In *Proc. of WWW.* 783–793.

[49] Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proc. of NAACL-HLT.* 592–596.

[50] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 6684 (1998), 440–442.

[51] Ingmar Weber, Antti Ukkonen, and Aris Gionis. 2012. Answers, not links: extracting tips from yahoo! answers to address how-to web queries. In *Proc. of WSDM.* 613–622.

[52] Howard T Welser, Eric Gleave, Danyel Fisher, and Marc Smith. 2007. Visualizing the signatures of social roles in online discussion groups. *Journal of social structure* 8, 2 (2007), 1–32.

[53] Max L. L. Wilson, Paul Resnick, David Coyle, and Ed H. Chi. 2013. RepliCHI: The Workshop. In *Proc. of CHI EA.* 3159–3162.

[54] Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. 2008. Retrieval Models for Question and Answer Archives. In *Proc. of SIGIR.* 475–482.

[55] Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, Zengfu Wang, Tat-Seng Chua, and Xian-Sheng Hua. 2010. Visual Query Suggestion: Towards Capturing User Intent in Internet Image Search. *ACM Trans. Multimedia Comput. Commun. Appl.* 6, 3 (Aug. 2010), 13:1–13:19.

[56] Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. (2002).