The Workshops of the Tenth International AAAI Conference on Web and Social Media
Social Media in the Newroom: Technical Report WS-16-19

What Questions Do Journalists Ask on Twitter?

Maram Hasanain,[1] Mossaab Bagdouri,[2] Tamer Elsayed,[1] Douglas W. Oard[3]
maram.hasanain@qu.edu.qa, mossaab@umd.edu, telsayed@qu.edu.qa, oard@umd.edu
[1]Computer Science and Engineering Department, Qatar University, Doha, Qatar
[2]Department of Computer Science, University of Maryland, College Park, MD, USA
[3]iSchool and UMIACS, University of Maryland, College Park, MD, USA

Abstract

Social media platforms are a major source of information for both the general public and for journalists. Journalists use Twitter and other social media services to gather story ideas, to find eyewitnesses, and for a wide range of other purposes. One way in which journalists use Twitter is to ask questions. This paper reports on an empirical investigation of questions asked by Arab journalists on Twitter. The analysis begins with the development of an ontology of question types, proceeds to human annotation of training and test data, and concludes by reporting the level of accuracy that can be achieved with automated classification techniques. The results show good classifier effectiveness for high prevalence question types, but that obtaining sufficient training data for lower prevalence question types can be challenging.

Introduction

Starting from 2011, the Arab Spring triggered a stream of events that changed and continue to affect the face of global politics. Along with this flood of events, Twitter is increasingly used as a global discussion and news reporting medium through which users express their thoughts and share and request information and news about those ongoing concerns. Journalists in particular have caught-up with Twitter as an invaluable source of information and are increasingly using it over time (Bruns, Highfield, and Burgess 2013). For example, after tracking tweets posted about the Egyptian and Tunisian revolutions, Lotan et al. (2011) found that journalists constituted about 14% of the users posting those tweets. Moreover, news agencies are continuously encouraging their journalists to use Twitter as part of their job (Parmelee 2013). The distinct nature of tasks journalists perform as part of their work indicates that the nature of tweets they post might be different from the general public (Bagdouri 2016).

Many users post questions on Twitter seeking answers from their network. Several studies focused on analyzing questions posted on Twitter (Morris, Teevan, and Panovich 2010; Efron and Winget 2010; Liu and Jansen 2012). However, most of them analyzed tweets posted by the general public; investigating questions of journalists specifically is rarely conducted. Looking at studies analyzing journalists' tweets (not particularly questions), some actually highlighted that journalists post questions on Twitter to collect information and opinion for journalistic reporting (Brautovi, Milanovi-Litre, and John 2013; Noguera-Vivo 2013; Parmelee 2013; Vis 2013; Revers 2014).

With the huge stream of tweets posted daily, journalists have a strong need for automatic systems that help them fulfill their information seeking tasks in an efficient and effective manner (Schifferes et al. 2014). Providing journalists with Twitter-based question-answering systems, for example, can help them acquire instantaneous answers and reactions to their questions. A first step in building such systems calls for an analysis and understanding of their questions posted through Twitter. Moreover, distinguishing between different types of questions journalists post is even more rarely done, yet it can help build "smarter" question-answering systems. For example, systems can choose to use different sources to answer different question types.

In this work, we present the first focused study aiming at analyzing types of questions journalists post on Twitter. As a case study, we focus our analysis on *Arabic* tweets.[1] We first collect tweets posted by Arab journalists, from which we automatically identify tweets with questions. Following a systematic analysis of extracted questions, we develop a taxonomy of seven categories based on the goal of posting the question. We recruit annotators to label question tweets following that taxonomy, and use these labels to train and test a question-type classifier for journalists questions. Our experiments show that the classifier is effective overall, yet classification of low prevalence question types can be improved with more labeled examples.

Research Questions We aim to answer the following research questions: (1) What types of questions do journalists ask on Twitter?, and (2) Can we effectively classify journalists' questions by type?

Contributions Our contributions in this work are three-fold:

- We introduce the first taxonomy of types of questions posted by journalists on Twitter.

[1] The average daily number of tweets in the Arab region increased from 1.2M in 2011 (Mourtada et al. 2011) to 17M in 2014 (Mourtada, Salem, and Al-Shaer 2014).

- We collect and release[2] manually-labeled data for two tasks on journalists' tweets: question tweet identification and question type classification.
- We implement and test an effective question type classifier.

We next review some related studies before discussing our approach in analyzing journalists' questions.

Related Work

Our paper is related to studies that investigate the activities of journalists on Twitter, and those that focus on question detection and classification in Twitter.

Journalists and Twitter

Over the past few years, journalists have been increasingly using Twitter to support their work activities (Lasorsa, Lewis, and Holton 2012; Hermida 2013). In a recent work, Parmelee (2013) interviewed 11 professional journalists. Some of them found Twitter to be ideal for finding and following breaking news, crowdsourcing and polling, keeping track of opinions of active players and experts on some event, and for finding sources of information; all can be used in writing news stories.

Though these interviews did not focus on questions that journalists post through Twitter, some journalists briefly mentioned having posted questions to their followers to collect answers serving journalistic reporting. We find a similar observation in a work by Revers (2014), where he analyzed 4.5K tweets posted by 25 US-based reporters around the time of passing a new law to New York Senate. He reported that 0.5% of the tweets aimed for crowdsourcing, including posting questions to followers. A recent work (Noguera-Vivo 2013) analyzed tweets of Spanish journalists and found that, out of 1.1K tweets, 5.3% explicitly ask for information. A very related study reported that out of 7.6K tweets posted by Croatian journalists, 1% of them were seeking information by posting questions to followers (Brautovi, Milanovi-Litre, and John 2013). Vis (2013) analyzed tweets authored by two journalists around the 2011 UK summer riots. Results showed that a large percentage of the tweets were questions or requests of information about the event, and requests for verification of related news.

All the discussed studies did not focus on journalists questions specifically, and types of questions were not clearly distinguished from each other. Moreover, those studies depended on manual analysis and coding of tweets (including questions), while we propose an automatic method for differentiating types of journalists questions.

Questions in Twitter

Several systems were developed to detect and classify questions in Twitter.

Question Detection With the growing interest in analyzing questions posted to Twitter, question detection approaches constitute a vital first step for tweets filtering, with a dominance of rule-based filters. Efron and Winget (2010) developed a set of rules that syntactically describe a question tweet, including whether it contains a question mark. Li et al. (2011) used the same rules, in addition to matching the tweets with the 5WH1 words. Zhao and Mei (2013) used only the question mark as a matching rule. Hasanain et al. (2014) used a rule-based filter for Arabic tweets, matching them with a question mark or a set of question phrases. A different approach was proposed by Li et al. (2011) and uses binary classification to detect question tweets. Though more sophisticated, this method showed inferior quality to that achieved by the simple rule-based filter. Liu and Jansen (2015) created a question tweet dataset by crawling tweets that appeared in a Twitter-based QA website. This approach, of course, depends on the availability of such websites.

We therefore follow the trend of using rule-base filters for question detection, and adopt the same approach proposed in (Hasanain, Elsayed, and Magdy 2014), as it was specifically designed for Arabic tweets.

Question Classification Existing studies on question tweets mainly focused on those posted by the general public. Few examples targeted community-specific questions (Efron and Winget 2010), but usually worked on a small scale set of tweets. Several taxonomies were built to distinguish between different question types, we present some of these taxonomies next.

Efron and Winget (2010) proposed an 8-type taxonomy, covering both rhetorical and information-seeking questions. Their work focused only on analyzing the outcome of manually-labeled question tweets using this taxonomy. Li et al. (2011) proposed a 2-way question classification, differentiating between answer-seeking and non-answer-seeking questions. They also built a binary classifier to assign question tweets to these two types, achieving an accuracy of 77.5%. Similarly, Zhao and Mei (2013) distinguished information-seeking from non-information-seeking questions. Their binary classifier achieved an accuracy of 85.6% on their question tweet dataset. In a very similar problem that focuses on Arabic question tweets, a binary classifier achieved an F_1 score of 0.716 (Hasanain, Elsayed, and Magdy 2014). Liu and Jansen proposed a 2-type taxonomy specifically for information-seeking question tweets in which questions can be subjective or objective. Using an automatic classifier, they achieved an accuracy of 81.65%. We discuss how our proposed taxonomy is different in the following section.

Question Taxonomy

Several taxonomies have been suggested for questions asked on Twitter. Some of them consider the topic of the tweet (Forte et al. 2014; Liu and Jansen 2012; Paul, Hong, and Chi 2011), such as Technology and Sports, which is not our focus in this work. Some approaches for creating a taxonomy of question types directly survey users of Twitter and other

[2] To comply with Twitter's terms of service, we only share the tweet IDs along with labels through this link: http://qufaculty.qu.edu.qa/telsayed/datasets/

social media platforms about the questions they ask (Forte et al. 2014; Morris, Teevan, and Panovich 2010). While this helps gain some insights about the types of questions that people *think* they ask, these types might be different from what they *actually* post. In some studies, researchers performed their analysis directly on the tweets content over a short period of time (a week or less) (Efron and Winget 2010; Paul, Hong, and Chi 2011). Consequently, the number of tweets that could be annotated was limited.

Our approach for creating the taxonomy might appear similar to the latter studies. Nevertheless, designing our taxonomy is merely a first phase of a series of stages leading to the development of an automated system that detects the question types. This has an impact on the design decisions made during this process. For instance, we need a fairly large set of annotations (i.e., on the order of thousands, not just few hundreds) to train and test our classifier. As crowdsourcing is an appropriate choice to gather some of these assessments at a relatively low cost, we had to go through several iterations of refining and redefining the question categories. Indeed, not only the definitions have to be clear and agreed on among the authors, but they also need to be well articulated, in writing, to assessors who are often not willing to carefully read long instructions, and are eager to finish the task as fast as they can. In addition, a taxonomy that is developed for the general public (Morris, Teevan, and Panovich 2010; Paul, Hong, and Chi 2011) or for a particular population (e.g., teens (Forte et al. 2014) and information retrieval researchers (Efron and Winget 2010)) might not reflect some special characteristics of our group of interest (i.e., journalists). Finally, cultural and linguistic differences between a language emerging in Twitter (in our case, Arabic) and the dominant language (on Twitter, English) are themselves worth to be studied.

Creating the Taxonomy While creating the taxonomy, our goal was to identify common types of questions (based on their intent) that journalists use in their tweets, can be automatically detected by a classifier, contain a real information need, and will be useful as a feature for other subsequent stages that aim to satisfy this information need. For example, if the question types classifier accurately detects that a tweet is seeking opinions, we could send that tweet to some system that creates a poll, gathers votes from other Twitter users, and produces a histogram over the most prominent opinions. Similarly, if that classifier predicts that the journalist wants to verify some breaking news, we can route her tweet to some potential eye witnesses. Therefore, we want the taxonomy to be general enough and not too fine-grained that it may not be useful to guide such potential automatic systems.

Given a large dataset of Arabic tweets posted by journalists, we applied a question detection rule-based filter (Hasanain, Elsayed, and Magdy 2014), and then randomly sampled 90 tweets from the detected questions to be used in the taxonomy creation. Two judges (the two lead authors of this study) worked independently on tagging each tweet by the question type considering question intent and its expected answer (if any). Each judge also provided some definitions to the suggested types. Judges then discussed the types and definitions to reach a mutual understanding and converge to a final taxonomy of seven categories. Moreover, they added an eighth class ("other") to cover any type that does not fit in any of the proposed seven types or for cases were annotators did not understand the tweet. We present the taxonomy next, and show one example from each category in Table 1, with both the original Arabic tweet and its English translation.

1. **Find Fact**: expecting new facts or details about already-known facts or stories as a response, even if they concern a personal matter.
2. **Find Information Source**: requesting a source for specific information (e.g., an eyewitness or a document).
3. **Confirm Fact**: asking for confirmation of a fact or a piece of news; the journalist is usually aware of it but needs a verification.
4. **Find Opinion**: asking for opinions on a topic, in addition to polling and recommendation/advice requests.
5. **Clarify Opinion**: a special case of requesting opinions where the journalist asks for a clarification of another person's opinion. The original opinion usually appears in the same conversation.
6. **Enrich Argument**: pointing to a flaw in the logic of the argument of another person, whether this person is involved in the same Twitter conversation or not. These questions can come in different forms including irony, sarcasm, and joking.
7. **Disseminate**: pointing to some resources including: websites, articles, videos, etc.; or advertising for other things in general (e.g., a product, a TV show).
8. **Other**: for difficult to understand questions or those that do not fit in any of the above types.

Not all of these question types are equally important. Since our main focus is to detect (and perhaps later to answer) questions with real information needs, the categories that seem to be the most interesting are *Find Fact*, *Find Information Source*, *Confirm Fact*, *Find Opinion* and *Clarify Opinion*.

Question Classification

Given a set of journalists tweets, we work on answering our research questions by following a 2-stage approach: 1) question tweet detection, and 2) question type-classification. We describe each stage next.

Question Tweet Detection To automatically extract question tweets from a set of journalists tweets, we used a rule-based filter since such filters have proven to be reasonably-effective in detecting question tweets (Efron and Winget 2010; Li et al. 2011; Hasanain, Elsayed, and Magdy 2014). We use an existing filter specifically designed for Arabic tweets (Hasanain, Elsayed, and Magdy 2014). It labels a tweet as a question tweet if it contains a question mark or it has one of commonly-used dialectal and modern standard

Table 1: Example question tweets of the final types

Find fact	Find information source	Confirm fact
@DrBasselSaleh وين الانفجار ؟؟ @DrBasselSaleh Where did the explosion happen??	@hmmzayed هل لدى حضرتك رابط لمقال أو دراسة؟ @hmmzayed Do you have a link to an article or a study about that?	هي الجزيرة قالت فعلا اني اتقبض عليا ؟! Did AlJazeera really announce that I got arrested?
Find opinion	**Clarify opinion**	**Enrich argument**
ما رأيكم في هاتف غالاكسي S3 الجديد وهل بالفعل هو أفضل من ايفون 4S شاركونا بآرائكم؟؟ What do you think about the Galaxy S3 smartphone? is it better than iPhone 4S? please share your opinions??	@Mera_Ibrahim @AhmedElderiny1 @CegerxwinHassan @TheMiinz @Mtolba كيف العكس؟ وضح أكثر لو سمحت؟ @CegerxwinHassan @AhmedElderiny1 @Mera_Ibrahim @Mtolba @TheMiinz what do you mean by the “opposite will happen”? can you please be more clear?	@hadithii يعني نظام صدام لم يكن نظاماً شمولياً ديكتاتورياً؟ كان نظاماً ديمقراطياً .. مثلاً؟ :))))) عن جد ضحكتني ... من قلبي ! @hadithii So Saddam’s regime wasn’t dictatorship? it was democratic then, right? :))))) I find this really funny!
Disseminate		
#سوريا: هل كانت الثورة ضرورة؟ http://t.co/uwBPDMPu3H #Syria: Was the revolution a necessity? http://t.co/uwBPDMPu3H		

Arabic question phrases. The filter had a precision of about 79% when tested on a set of general public tweets.

Question-Type Classification In order to design and evaluate a classifier that automatically identifies the type of a question, we develop a set of features that have the potential of capturing certain characteristics of the question types. We organize these features in four conceptual families:

1. **Lexical Features:** Some words and expressions can be more likely associated with some categories, than others. For instance, the Arabic word meaning *in your opinion*, is perhaps a useful feature to detect the category *Find Opinion*, rather than *Find Fact*. For this, we generated the unigrams and bigrams of the content of the tweets after removing user mentions, URLs and punctuation, and processing the remaining terms with the Arabic light stemmer (Larkey, Ballesteros, and Connell 2007) implemented in Lucene 5.3.1.[3]
2. **Tweet Metadata Features:** In addition to the content, tweets contain some metadata that can be useful to classify some categories. For instance, the presence of a URL might indicate that a journalist wants to share a news story, and that the corresponding question type is *disseminate*. A reply can also indicate that the journalist is having a conversation to *enrich an argument*. Other features we consider are the length of the content of the tweet in terms of words and characters, and the number of hashtags, mentions, images and videos.
3. **User Metadata Features:** Some users tend to use Twitter in a manner different from others. Similarly, some journalists can be associated with some particular categories more than others. Hence, we include some user-specific features, such as the user ID, the ratio of followers over friends and the indication whether the user has a “verified” sign in her profile.
4. **Conversational Features:** Sometimes the question comes within a conversation. Using this conversation as a context might inform us about the category of the question. However, we limit the conversation to the “parents” of the tweet, instead of its “children.” That is, we only look at the series of tweets for which the current one is a reply. Including features of the replies to the current tweet would be unfair, as in a real-time scenario, we would not have access to such information. The features we use in this family are the duration of the conversation (in logarithmic seconds), the domination of the journalist in the conversation (i.e., the ratio of the count of her tweets over the number of tweets in the conversation), and the number of interrogative tweets in the conversation.

Data and Human Annotations

To construct our dataset of question tweets posted by journalists, we first acquire a list of Twitter accounts of 389 Arab journalists (Bagdouri and Oard 2015). We use the Twitter API to crawl their available tweets, keeping only those that are identified by Twitter to be both Arabic, and not retweets (as these would contain content that was not originally authored by journalists). We apply the rule-based question filter to this dataset of 465,599 tweets, extracting 49,119 (10.6%) potential question tweets from 363 (93.3%) Arab journalists.

Question Tweets Annotation

To verify the performance of the automatic rule-based filter, we collect human annotations for potential question tweets by crowdsourcing through CrowdFlower.[4]

We randomly sample 10K tweets from the potential question tweets set, ensuring each journalist is represented by at least five tweets. Three Arabic speaking annotators were asked to label each tweet, judging whether it contains at least one question. To keep the annotators alert and maintain a good annotation quality, we insert gold tweets in the task. Annotators were required to pass a qualifying quiz over the gold tweets before doing any labeling, and to maintain an accuracy above 70% throughout the task. Using Fleiss' Kappa (κ) (Fleiss 1971) to measure inter-rater agreement for 3 annotators, the agreement was 0.473. This translates to moderate agreement based on the widely-used interpretation of

[3] http://lucene.apache.org

[4] http://crowdflower.com

the κ statistic (Sim and Wright 2005).

For each tweet, the crowdsourcing platform combines the labels from different annotators and produces a single label with a confidence level that measures the annotators agreement weighted by their accuracy over the gold tweets.[5] We set a confidence level threshold of 0.5, and find that all tweets meet at least this level. Out of all labeled tweets, 8.6K tweets are labeled as true question tweets. The filter shows a precision of 86% for the tweets, which is larger than that reported over a general dataset (Hasanain, Elsayed, and Magdy 2014).

Question-Type Annotation

To validate the proposed taxonomy and develop a question-type classifier, we create a dataset of journalists question tweets annotated by question type. In this task, annotators were asked to classify a true question tweet into one of the eight types of our taxonomy. We also provided them the links to the tweets to allow reading them in full context through Twitter's website. We ran several pilot studies to iteratively enhance the task and to guide our design decisions. In the following section, we explain a set of pilot studies we went though before collecting the final set of annotations.

Pilot Runs We used CrowdFlower to collect question type annotations. With the 8.5K true question tweets gathered during the first stage, we conducted several CrowdFlower tasks, ultimately amounting to 1.3K tweets in our final CrowdFlower task. The inter-rater agreement proved to be disappointingly low ($\kappa = 0.19$), corresponding to only slight agreement (Sim and Wright 2005). Only one-third of 87 annotators successfully passed a qualifying quiz, and several of those were later eliminated for failing to maintain a good accuracy over the gold tweets.

We observed that with such a large number of classes, the annotation guidelines were longer than what the assessors were willing to read. This obviously affected their understanding of the task, and as a result, their performance as well. Additionally, the annotators were required to label very short text snippets (i.e., tweets)—many written in dialectal Arabic—making the task even harder and more time consuming due to the lack of context of a tweet and a potential dialect barrier. Thus, we decided to recruit in-house annotators instead, which allowed us to offer them a more comprehensive training, and to have more control on the quality of the labels (e.g., by continuous encouragement and prompt feedback on their labels). Additionally, we improved type definitions and instructions, making them clearer and more concise.

In addition to the previously-discussed conclusions, we observed that the tweets usually lack context as they are very short in length, and many of them cover news-related events that not all annotators might be familiar with. This added more difficulties to the task; labeling a tweet with a single type can be genuinely difficult for any annotator no matter how well-trained she is. Thus, we improved the design of the task by allowing annotators to choose a maximum of two types for a tweet (we call them type x and type y).[6] They were also asked to express the level of confidence, on a 5-point scale, for which of the two types they think fits the question better. Choosing 1 on the scale means that they are almost completely confident that the question is of type x, while 5 means they are almost completely confident that the question is of type y.

We recruited 3 in-house annotators: one post-doc and 2 graduate students.[7] Before qualifying the annotators to work on the final task, we had one-to-one training sessions with each of them, in which we asked them to label 150 tweets and gave them our feedback on their performance. During these sessions, all annotators expressed that they found the instructions to be fairly clear, but the task to be very difficult; lack of context for a tweet is among the most difficult issues according to annotators.

Final Task To ensure annotators' time is not wasted on labeling tweets that do not contain questions, we chose to work with the true question tweets that received labels with the highest label confidence level in the first stage, resulting in 7.1K tweets. We randomly sampled 2.25K tweets out of this set and asked annotators to label them by question type following the same task design in our final pilot study discussed earlier. We did not use test questions in this task since we trust that our annotators are committed to accurately label the tweets.

As explained earlier, annotators were allowed to choose a maximum of two types (i.e., type x and type y) per question, along with choosing a value from 1 (almost completely confident in type x) to 5 (almost completely confident in type y) on a confidence scale. For calculating inter-annotator agreement, we assigned a single label for each of the 2.25K question tweets as follows. We refer to the confidence value indicated by the annotator as l. If $l = 3$, we randomly assign either of the two types given by the annotator as the question type, else if $l < 3$ (i.e., the annotator is more confident in type x than type y), the tweet is assigned type x. Otherwise, we select type y as the label. We apply this process to all tweets that were assigned two types by the annotator. Over 2.25K tweets, Fleiss' κ was 0.450 which is considered a moderate agreement. Majority agreement was achieved for 85% of the tweets, and 44% of the tweets had full agreement among all annotators.

For classifier training and testing, we used a slightly different (and more strict) technique to aggregate the type labels given to a question tweet. The question type for a tweet is selected to be one of the 8 types in our taxonomy such that this type gets a maximum cumulative score over all annotators. The score per annotator is computed as follows: a) if a single type is given to the tweet, that type gets a score of 1, b) if a tweet receives two type labels from the same annotator, the score for type x = (6-l)/6 and score for type y = 1-((6-l)/6) are computed. Accumulating the resulting scores

[5] http://success.crowdflower.com/hc/en-us/articles/201855939

[6] Annotators were not asked to select the two types in any specific selection order, they were shown the 8 types in a checklist and they can check two types at most.

[7] We thank Dr. Marwan Torki, Reem Suwaileh and Abeer Al-Marridi for their voluntary work on data annotation.

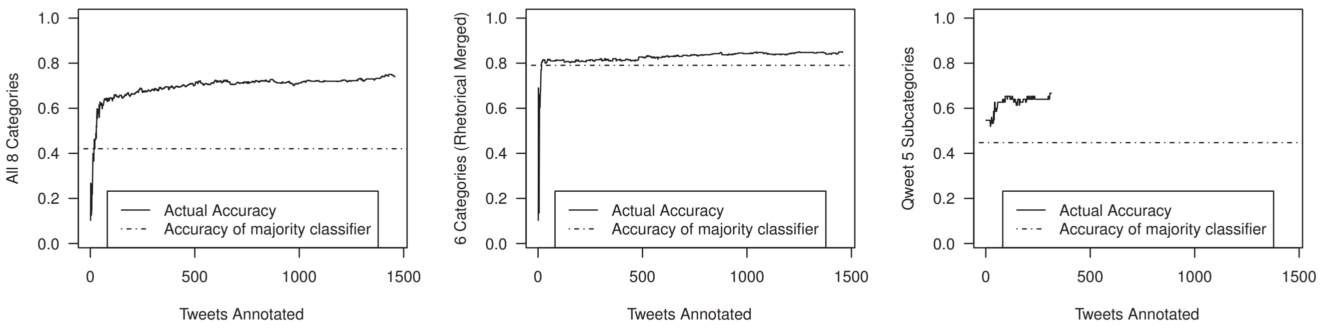

Figure 1: Average multi-class accuracy based on the number of categories considered.

over all annotators, we select the final type label to be the type with the maximum score. We further filter the labeled tweet set to include only tweets that received a type with a maximum score that is greater than 1.5 (out of a maximum possible value of 3). This resulted in 1,858 labeled question tweets that could be used for training and testing. Table 2 shows the question type distribution of this set of labels.

Table 2: Distribution of labeled tweets across types

Type	**% of total**
Enrich argument	42.0%
Disseminate	28.5%
Find fact	9.4%
Find opinion	6.9%
Clarify opinion	2.6%
Confirm fact	1.4%
Find info. source	0.7%
other	8.5%

As the table shows, around half of journalists' questions are labeled with the "Enrich Argument" type. Investigation of the raw labels shows that for all 758 instances where a tweet received two types by at least one annotator, "Enrich Argument" was used as one of the two types in 68% of the instances. Given such observation and the high prevalence of this type in the dataset, we think that this type was ill-defined in our taxonomy, or its definition was too general that annotators found it sometimes difficult to distinguish tweets that truly belong to it from those that belong to other types. One way to address this issue is to split this type into sub-types with narrower scopes; an experiment we leave for future investigation.

Interestingly, we found the type "Disseminate" to be the second most prevalent question type in our dataset. Investigating examples of questions under this category, we observed that journalists often use questions to publicize their own articles. Moreover, most of these tweets appeared in our questions dataset because journalists usually use titles of the articles they are sharing as the tweet content, and those titles are actually formulated as questions. Since such questions can appear to be real answer-seeking questions, they can add noise to a system that aims to detect and answer questions of journalists (since such questions were already answered in the articles shared through the tweets). Contrary to what we expected, Table 2 shows that questions aiming at verifying or finding sources of information are very rare in our dataset. This might indicate that journalists are not confident enough in performing such sensitive practices of their job through Twitter; yet more data must be collected to investigate this issue further.

Experimental Evaluation

To evaluate our methodology, we use LIBSVM (Chang and Lin 2011) to train a multi-class linear Support Vector Machine (SVM) for the eight question types using 10-fold cross validation. Normalizing the features did not improve the performance. Hence, we report only classification results without features normalization. This results in an average accuracy of 71.04%, which is a substantial improvement over the baseline of always choosing the majority class (781 / 1,858 = 42.03%). The left-most plot of Figure 1 shows a sharp increase in performance until the annotation size (which, in this case, is split between training and test, so those tweets not used for training are used for test) reaches about 40 tweets. The increase in performance then slows down.

Because the applications in which we are interested do not focus on the *Enrich Argument* and *Disseminate* types, we merge them with *Other*. We next train and evaluate the classifier on a total of six categories (i.e., five question types of interest, and the merged category of uninteresting types). We get an average accuracy of 81.65%. While this is an improvement (from the application perspective) over the previous accuracy of 71.04%, this score can be misleading. In fact, the new score is barely over the combined prevalence of 79.06% for the three uninteresting categories. The middle plot of Figure 1 shows that the average accuracy starts to plateau just after hitting the performance of the majority class classifier.

We now attempt to isolate any effect of the uninteresting categories by completely excluding their corresponding tweets from both training and test. The right-most plot of Figure 1 shows that we can achieve a modest average accuracy of 66.67% that is higher than that of the majority class classifier (44.76%).

However, we do not know from this result if the difference should be attributed to the characteristics of the ques-

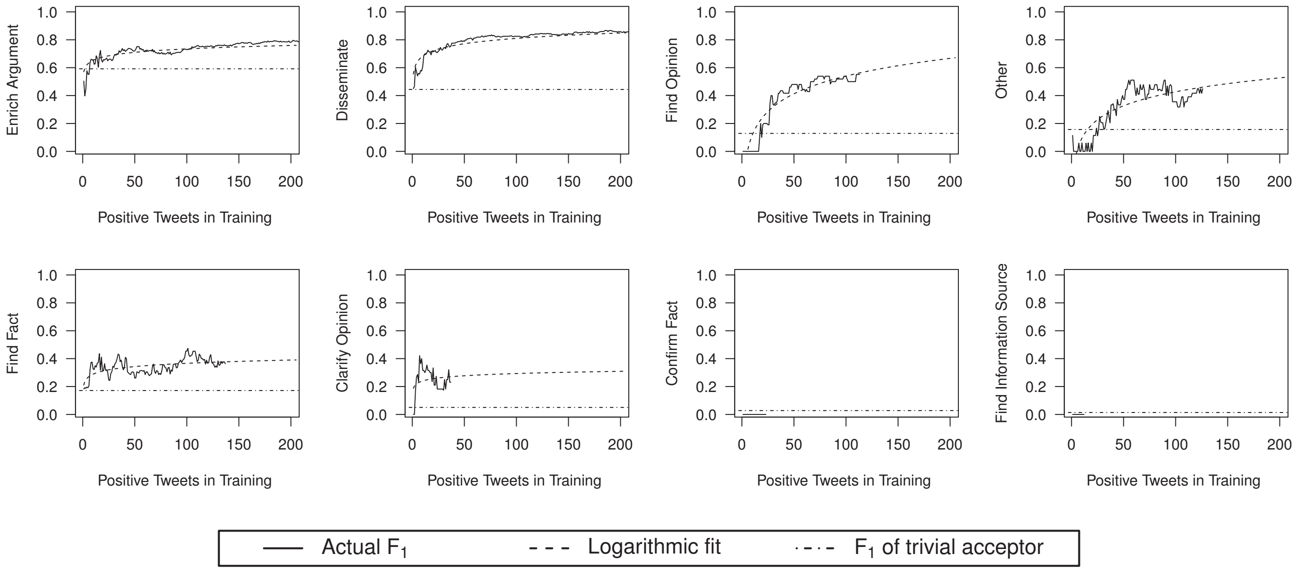

Figure 2: F_1 score as a function of the number of positive documents used to train an individual class.

tion types (e.g., they might be ill-defined), or to the number of tweets we have for each category. As the average accuracy over 10 folds and several classes hides some details, we next look at the usefulness of our features to detect each class individually. More importantly, we see whether adding more training tweets will help better identify the question types. To do so, we randomly split the tweets into a training set of 1,458 tweets, and a test set of 400 tweets. Then, independently for each type, we consider it to be a positive category, and the union of the other types to be a negative category. In this setup, the positive category will be a minority class in both training and test, and a measure other than accuracy could be more appropriate to evaluate the individual classifiers. For this, we choose F_1 as the evaluation measure, using the SVM^{perf} package (Joachims 2005), as it can optimize training directly for F_1 instead of accuracy. We fix the test set of 400 tweets, and gradually populate the training set through several stages, in a random order, such that at each stage we have one additional positive tweet (and eventually a few negative ones). We train, test and record F_1 before continuing to the following stage.

Figure 2 shows F_1 plots of the eight categories. We compare it against the naive classifier *trivial acceptor*, which always predicts tweets to be positive. In general, when sufficient positive tweets are present in the test set, performance appears to improve, at a logarithmic rate, as a function of the training-set size. For the *Enrich Argument and* Disseminate categories, the expected number of positive tweets (168 and 114, respectively) is large enough that increase in performance is fairly smooth. The next three categories, in terms of prevalence, are *Find Fact*, *Find Opinion* and *Other*. They also show increasing performance, but with relatively high variance, as their expected number of positive tweets in the test set falls between 27 and 38. We are practically hopeless as the prevalence drops below 3% for the categories *Clarify Opinion*, *Confirm Fact* and *Find Source*. It appears that both the classifier performance, and our ability to measure it are strongly impacted by the small number of positive tweets.

Conclusion

In this work, we have conducted the first study specifically focusing on analyzing the questions that journalists post in Twitter. We collected more than 49K journalists tweets that potentially contain questions. Working with a sample of these questions, we developed a 7-way taxonomy of journalists' question types. We used crowdsorucing to collect binary annotations for 10K of the potential question tweets based on whether they truly contain questions or not. Recruiting in-house annotators, we then collected question-type labels for 2.25K question tweets. Using 10-fold cross validation, an SVM classifier showed an average accuracy of 71% over the type-labeled question tweets. We also observed that classification performance is more effective with types of questions that are more prevalent in the labeled data. Thus, we hope that adding more labeled questions for the least common question types might further improve performance.

We plan to use active learning to gather more annotations for the questions types with low prevalence in our annotated data. Additionally, we hope to extend this work by conducting a comparative study with tweets of English speaking journalists.

Acknowledgments

This work was made possible by NPRP grant# NPRP 6-1377-1-257 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Bagdouri, M., and Oard, D. W. 2015. Profession-based person search in microblogs: Using seed sets to find journalists. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, 593–602.

Bagdouri, M. 2016. Journalists and twitter: A multidimensional quantitative description of usage patterns. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*, ICWSM '16. To appear.

Brautovi, M.; Milanovi-Litre, I.; and John, R. 2013. Journalism and Twitter: Between journalistic norms and new routines. *Medianali* 7(13):19–36.

Bruns, A.; Highfield, T.; and Burgess, J. 2013. The Arab spring and social media audiences English and Arabic Twitter users and their networks. *American Behavioral Scientist* 57(7):871–898.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Efron, M., and Winget, M. 2010. Questions are content: A taxonomy of questions in a microblogging environment. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, 27:1–27:10.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5):378.

Forte, A.; Dickard, M.; Magee, R.; and Agosto, D. E. 2014. What do teens ask their online social networks?: Social search practices among high school students. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, 28–37.

Hasanain, M.; Elsayed, T.; and Magdy, W. 2014. Identification of answer-seeking questions in Arabic microblogs. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, 1839–1842.

Hermida, A. 2013. #journalism: Reconfiguring journalism research about Twitter, one tweet at a time. *Digital Journalism* 1(3):295–313.

Joachims, T. 2005. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, 377–384.

Larkey, L.; Ballesteros, L.; and Connell, M. 2007. Light stemming for Arabic information retrieval. In *Arabic Computational Morphology*, volume 38 of *Text, Speech and Language Technology*. 221–243.

Lasorsa, D. L.; Lewis, S. C.; and Holton, A. E. 2012. Normalizing Twitter: Journalism practice in an emerging communication space. *Journalism Studies* 13(1):19–36.

Li, B.; Si, X.; Lyu, M. R.; King, I.; and Chang, E. Y. 2011. Question identification on Twitter. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, 2477–2480.

Liu, Z., and Jansen, B. J. 2012. Almighty twitter, what are people asking for? *Proceedings of the American Society for Information Science and Technology* 49(1):1–10.

Liu, Z., and Jansen, B. J. 2015. Subjective versus objective questions: Perception of question subjectivity in social Q&A. In *Social Computing, Behavioral-Cultural Modeling, and Prediction*, volume 9021. 131–140.

Lotan, G.; Graeff, E.; Ananny, M.; Gaffney, D.; Pearce, I.; and Boyd, D. 2011. The Arab spring| The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication* 5:31.

Morris, M. R.; Teevan, J.; and Panovich, K. 2010. What do people ask their social networks, and why?: A survey study of status message Q & A behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, 1739–1748.

Mourtada, R.; Salem, F.; Al-Dabbagh, M.; and Gargani, G. 2011. The role of social media in Arab womens empowerment. *Arab Social Media Report series.*

Mourtada, R.; Salem, F.; and Al-Shaer, S. 2014. Citizen engagement and public services in the Arab world: The potential of social media. *Arab Social Media Report series.*

Noguera-Vivo, J. M. 2013. How open are journalists on Twitter? trends towards the end-user journalism. *Communication&Society* 26(1).

Parmelee, J. H. 2013. Political journalists and Twitter: Influences on norms and practices. *Journal of Media Practice* 14(4):291–305.

Paul, S. A.; Hong, L.; and Chi, H. 2011. What is a question ? crowdsourcing tweet categorization. In *Workshop on Crowdsourcing and Human Computation at the Conference on Human Factors in Computing Systems (CHI).*

Revers, M. 2014. The twitterization of news making: Transparency and journalistic professionalism. *Journal of Communication* 64(5):806–826.

Schifferes, S.; Newman, N.; Thurman, N.; Corney, D.; Göker, A.; and Martin, C. 2014. Identifying and verifying news through social media: Developing a user-centred tool for professional journalists. *Digital Journalism* 2(3):406–418.

Sim, J., and Wright, C. C. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy* 85(3):257–268.

Vis, F. 2013. Twitter as a reporting tool for breaking news. *Digital Journalism* 1(1):27–47.

Zhao, Z., and Mei, Q. 2013. Questions about questions: An empirical analysis of information needs on Twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, 1545–1556.