Exploring the Role of Context to Distinguish Rhetorical and Information-Seeking Questions

Yuan Zhuang and Ellen Riloff
School of Computing
University of Utah
{yyzhuang, riloff}@cs.utah.edu

Abstract

Social media posts often contain questions, but many of the questions are rhetorical and do not seek information. Our work studies the problem of distinguishing rhetorical and information-seeking questions on Twitter. Most work has focused on features of the question itself, but we hypothesize that the prior context plays a role too. This paper introduces a new dataset containing questions in tweets paired with their prior tweets to provide context. We create classification models to assess the difficulty of distinguishing rhetorical and information-seeking questions, and experiment with different properties of the prior context. Our results show that the prior tweet and topic features can improve performance on this task.

1 Introduction

Questions are common in social media forums, but they can serve many pragmatic functions. Questions are often information-seeking, but social media posts also frequently contain questions that do not expect any information for what the question literally asks about. We will use the term *rhetorical question (RQ)* broadly to refer to all questions that do not seek any information. For example, rhetorical questions can express criticism (e.g., *"Can't you do anything right?"*), sentiment (e.g., *"How fun is that?"*), sarcasm (e.g., *"Who knew?"*), and agreement/disagreement (e.g., *"Is the pope catholic?"*). Distinguishing rhetorical and information-seeking questions is important for dialogue processing and conversational analysis, but only recently has begun to receive attention in the NLP community.

Our research has two main contributions. First, we created a new resource for this understudied problem. We have compiled a collection of nearly 5,000 tweets that contain a question that is responding to an initial tweet, and labeled the questions as *rhetorical* or *information-seeking* with crowdsourcing. We found that 53% of the questions are information-seeking (IQ) and 47% are rhetorical (RQ), confirming that both types of questions are prevalent in social media.

Second, our research examines whether the initial tweet prior to a question can help to predict whether a question is information-seeking or rhetorical. Most prior work has focused only on the question itself, but we investigate whether the topic of the discussion may be a valuable indicator too. Our intuition was that rhetorical questions are common in contexts associated with argumentation and debate, such as politics. Conversely, we expect information-seeking questions to be prevalent in contexts about products and services, where people are actively seeking information.

In this paper, we first describe our Twitter dataset and human annotations. Next, we present classification models that exploit both the question and the initial tweet prior to the question. We explore several ways of extracting topic information from tweets to capture the prior context. Our results show that the prior context does improve performance for this task.

2 Related Work

Rhetorical questions have been studied in linguistics, primarily focused on linguistic properties and pragmatic functions (Sadock, 1971; Schmidt-Radefeldt, 1977; Frank, 1990; Gutierrez-Rexach, 1998; Han, 2002; Schaffer, 2005). However there has been relatively little work on rhetorical questions in the NLP community until recently. Work by Zhao and Mei (2013) identified the information need of questions in Twitter by extracting features from the question tweets. However, their work did not explore the usefulness of prior context in dis-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 306–312
July 5 - July 10, 2020.

tinguishing rhetorical questions from information-seeking questions. Ranganath et al. (2016) modeled the contextual overlap between a question and the most recent status message (MRSM) of the same user in Twitter, with the hypothesis that a rhetorical question shares context with its MRSM more than a random question with its MRSM. Bhattasali et al. (2015) found that n-gram features from utterances immediately preceding and following a question could help identify rhetorical questions. Our work differs from both works in several aspects. First our evaluation dataset contains human-assigned gold labels and a rich mix of both RQ and IQ. In contrast, Ranganath et al. (2016) automatically assigned their dataset with labels according to some heuristic rules, which may be noisy, and Bhattasali et al. (2015) used the Switchboard Dialog Act Corpus (Godfrey et al., 1992), where only 5% of questions are rhetorical. Second, neither of these works took preceding context and topic information into account.

Oraby et al. (2016) studied rhetorical questions in the context of sarcasm in debate forums, but they did not study the problem of distinguishing rhetorical questions from information-seeking questions. In contrast, we focused on distinguishing the information need of general questions in Twitter. Oraby et al. (2017) further explored distinguishing rhetorical questions from information-seeking questions. But their gold standard data consists of rhetorical questions automatically extracted from debate forums using heuristic rules. In contrast, our gold standard data consists of questions that have been manually labeled as rhetorical or information-seeking. Another difference is that Oraby et al. (2017) did not consider the prior context for questions, which we focus on in this work.

3 Data

We began by collecting tweets that contain question marks from January to December 2014.[1] We then applied a few filters to remove tweets that (1) are not in English (based on Twitter's language code), (2) contain < 5 words, (3) are retweets or have quotation marks around the question, because these questions did not originate with the tweeter, (4) contain URLs or media (e.g., photos), because the question may refer to the linked content, (5) contain multiple questions, which could be difficult to tease apart, or (6) were posted by a VIP ("verified") account. Questions posted by VIP accounts (entities in the public interest) were predominately rhetorical questions in advertisements, and we did not want these to skew our data. We will refer to the resulting tweets as **Question Tweets (QTweets)**.

We also collected the preceding tweets, which we will refer to as **Prior Tweet (PTweets)**. Our hypothesis is that the preceding context can be important because (a) the question alone can be ambiguous, and (b) knowledge about the topic of discussion can affect the likelihood that a question will be rhetorical or information-seeking. Consequently, we only kept question tweets that responded to a prior tweet. We also required that the prior tweet was the initial tweet in the conversational thread because conversational threads often have topic shifts and questions may refer back to earlier comments. Detangling discourse threads is a challenging problem in its own right.

This process produced 5,064 question tweets[2], each paired with its prior tweet as context.

3.1 Manual Annotation

We hired three annotators from Amazon's Mechanical Turk[3] to label each of the question tweets (coupled with its prior tweet) with one of the following three labels:

Information-seeking Question (IQ): The main purpose of the question is most likely to seek some information about what it literally asks.

Rhetorical Question (RQ): The main purpose of the question is most likely *not* to seek any information about what it literally asks. Instead, the speaker uses the question mainly for some other purpose, such as suggestion or criticism.

Incomprehensible (I): The annotation sample is not in English or it is hard to understand.

We emphasized in the annotation guidelines that some questions are ambiguous and could indeed have multiple purposes at the same time. One example is the question *"The sunset is great, isn't it?"*, which may convey the speaker's admiration of the sunset and also seek the hearers' agreement at the same time. We advised the annotators to choose the most likely primary purpose of a question, according to their instincts. To further assist

[1] We intentionally collected tweets from several years ago because their continued presence on Twitter suggests that they are likely to remain available, so other researchers can easily reacquire our data.

[2] We originally collected 5,200 tweets, but a pre-processing error allowed 136 tweets with < 5 words to slip through so they were later discarded.

[3] https://www.mturk.com/

the annotators, we provided several examples of both rhetorical and information-seeking questions in the annotation guidelines, along with explanations for why each question belongs to its corresponding category.

The pairwise inter-annotator agreement scores using Cohen's kappa were: κ = .67, .67, .68. Of the 5,064 questions, 67 (1.3%) were annotated as Incomprehensible by at least 1 annotator and discarded. The rest were assigned a gold standard label using majority vote. The final annotated dataset contains 4,997 question tweets, with 2,332 (47%) labeled as rhetorical and 2,665 (53%) labeled as information-seeking. The final gold standard dataset is available for download at the authors' website.

4 Classifying Questions as Rhetorical or Information-seeking

We designed a variety of classification models to assess the difficulty of distinguishing rhetorical and information-seeking questions, and to examine the role of prior context for this task.

First, we applied the CMU Twokenizer (Gimpel et al., 2011), removed URLs and hashtags, and replaced acronyms with their corresponding full words or phrases using a Twitter acronym list[4]. Next, we applied the Stanford CoreNLP parser (Manning et al., 2014) to obtain lemmas and part-of-speech tags. For the embedding features, we used GloVe vectors (Pennington et al., 2014) pre-trained on 2B tweets. We experimented with both 25 and 100 dimensional vectors, and show the best results in Section 5. We then extracted three sets of features: word features, question features, and topic features.

4.1 Word Features

We explored both unigrams and embedding vectors to capture the meaning of the words in a tweet.

Unigrams: Each word is a feature with a TF-IDF value. We only include unigrams that occur $\geq$ 3 times in the training set.

Embedding (Embed): We create an embedding vector for a tweet by averaging the embedding vectors for all words in the tweet.

[4]https://sproutsocial.com/insights/social-media-acronyms/

4.2 Question Features

We suspected that rhetorical questions and information-seeking questions may be phrased differently. Hence we developed 3 features to capture the question form.

Question Attributes: (1) One feature represents the leading bigram of the question (e.g., a leading "How to" may be more likely to seek a solution), (2) one feature indicates the WH-category of the leftmost question word: {*who*, *when*, *what*, *where*, *which*, *why*, *how*}. (3) one feature counts the number of negations in the question, as rhetorical questions may have more negations (e.g., *"why don't you try this ?"*).

Post-Question Attributes: We observed that rhetorical questions in Twitter are often followed by another sentence (suggesting a self-answer) or emoji. So we created three post-question features: (1) a feature indicating whether the question is followed by additional words, (2) a feature indicating whether the question is followed by emoji and (3) a feature that counts the number of emoji after the question.

Subjectivity Features: Rhetorical questions often express an opinion (e.g., criticize), agreement/disagreement, etc. So we hypothesized that recognizing subjective language may be a helpful clue for identifying rhetorical questions. We extracted 5 features associated with subjectivity: (1) the number of elongated words (e.g., *"looooove"*), (2) the number of entirely upper case words (e.g., *"YAY"*), (3) the number of exclamation marks, (4) the number of strongly subjective words found in the MPQA lexicon (Wilson et al., 2005), and (5) the number of weakly subjective words in the MPQA lexicon.

4.3 Topic Features

Our research explores whether the topic of discussion can help distinguish rhetorical and information-seeking questions, so we created four types of features to capture topic information.

Nouns Embedding (NounEmbed): The set of nouns in a tweet, in aggregate, might sufficiently reflect the topic of a tweet. So we created a composite nouns embedding vector by averaging embeddings of all the nouns.

Specificity (Specific): Information-seeking questions often focus on a specific entity or object, so we created features to capture specificity using the MRC resource (Brysbaert et al., 2013),

which assigns words with familarity and concreteness scores from 100 to 700. One feature counts the number of words with familiarity score ≥ 400, and the other feature counts the number of words with concreteness score ≥ 400.

Latent Dirichlet Allocation (LDA): We created an LDA model (Blei et al., 2003) from our training data, after removing stopwords, with $k = 25$ as the number of topics. Given input text, we extracted the latent topic distribution as k features.

Google Topic Categories (GTopic): Google's Content Classifier[5] labels text with respect to **700+** topic categories in its content hierarchy. Our dataset is small, so we only used the 27 general categories in the top level of its hierarchy. Given an input text, we used 27 features to capture the topics assigned by Google's Content Classifier.

Initially we extracted topic features directly from a tweet. But topic models and classifiers perform better on longer texts, so we also tried giving a tweet to Google as a query, and extracting the summary snippet for the top-ranked web page.[6] The resulting snippet is usually longer but similar in topic. We will call the snippet retrieved by a QTweet its **QSnippet**, and the snippet retrieved by a PTweet its **PSnippet**. In our experiments, we tried extracting the topic features from the tweet alone, its snippet, and from the tweet combined with its snippet. For the sake of brevity, we only show the best-performing results.

4.4 Learning Models

We created two types of classifiers: 1) a linear SVM (Chang and Lin, 2011) with $C = 0.1$, and 2) a 4-layer BiLSTM, implemented using PyTorch[7], with a hidden size of 100 and ReLU. We set the learning rate of the BiLSTM to 0.0001 with a dropout rate of 0.1 (Srivastava et al., 2014). For both models, we use GloVe embeddings pre-trained on 2B tweets of size 25 or 100 dimensions (Pennington et al., 2014).

5 Experimental Results

We randomly split our data into 3 partitions: training (3,200), development (797), and test (1,000). All classifiers were trained on the training set and tuned with the development set. We report results on the test set as Precision, Recall, & F1 scores, all macro-averaged over the RQ and IQ classes.

First, we evaluated models that used features derived only from the QTweet (QT). The first two rows of Table 1 show the performance of SVMs trained with word embedding vectors and unigrams, respectively. The third row shows that the BiLSTM outperforms the SVMs, achieving an F1 score of 70.8. However, the fourth row shows that adding the Question Features to the SVM performs better than the BiLSTM, yielding an F1 score of 73.5.

Classifiers for QTweet	**Prec**	**Rec**	**F1**
SVM Embed(QT)^{100D}	67.6	67.8	67.6
SVM Unigrams(QT)	68.8	68.9	68.9
BiLSTM(QT)^{100D}	71.0	70.9	70.8
SVM Unigrams(QT) + QFeatures(QT)	**73.5**	**73.5**	**73.5**
Adding Topic Features			
+ NounEmbed(QT)^{25D}	72.9	72.9	72.9
+ Specific(QT)	73.1	73.1	73.1
+ LDA(QT)	73.4	73.4	73.4
+ GTopic(QT)	73.5	73.6	73.5
+ ALL topic features (**QT-SVM**)	**73.9**	**74.0**	**73.9**

Table 1: Results using only QTweet (QT)

The lower portion of Table 1 shows results when adding each type of topic feature (not cumulatively) to the best SVM model. None of them improved performance on their own, but adding them all together (shown in the last row) increased the F1 score to 73.9. We observed that the topic is often unclear from the question itself, which may explain the minimal gains. We will refer to the best model in Table 1 as QT-SVM.

Classifiers for QTweet+PTweet	**Prec**	**Rec**	**F1**
BiLSTM(PT + QT)^{100D}	70.3	70.1	70.2
QT-SVM+Embed(PT)^{25D}+Sbj(PT)	74.4	74.5	74.5
+ Specific(PT)	74.7	74.8	74.8
+ LDA(Psnippet)	74.7	74.8	74.8
+ GTopic(Psnippet)	74.8	74.9	74.8
+ NounEmbed(PT + Psnippet)^{25D}	75.1	75.2	75.2
Best Combination: + Specific + LDA + GTopic	**75.4**	**75.5**	**75.5**

Table 2: Results using QTweet (QT) & PTweet (PT)

Table 2 shows results for classifiers that used features derived from both the QTweet and PTweet. The first row shows the BiLSTM model trained with the PTweet words followed by the QTweet words. This model performs slightly worse than the BiLSTM trained on QTweets alone. The next row shows results for QT-SVM with added features representing the PTweet as a 25D embedding vec-

[5] https://cloud.google.com/natural-language/

[6] We filtered snippets from Twitter.com or any website with 'dictionary' in its url or title, because the snippet from Twitter.com is usually the tweet itself, and online dictionaries just provide definitions of the words.

[7] https://pytorch.org/

	RQ		IQ	
	Prec	**Rec**	**Prec**	**Rec**
QT-SVM	71.3	72.4	76.5	75.5
QT-SVM+Embed(PT)^{25D}+Sbj(PT)	71.7	73.3	77.1	75.7
+Topic Features	72.9	74.2	77.9	76.8

Table 3: Breakdown of Precision and Recall Scores of Different Models for Each Question Class

tor[8] and Subjectivity (Sbj) Features[9] extracted from the PTweet. The additional PTweet information improved the SVM performance from 73.9 to 74.5. The following rows show results when adding each type of topic feature extracted from the PTweet (not cumulatively). Each of them slightly improved performance. We also experimented with combining them and the last row shows the best-performing combination, which achieved an F1 score of 75.5. We conjecture that each topic feature itself does not necessarily capture useful topic information across all questions, but combined they become complementary to each other and are more useful for the classifier.

Table 3 shows the breakdown of precision and recall scores for rhetorical questions and information-seeking questions separately. Overall the scores for rhetorical questions are consistently lower than for information-seeking questions, which means that it is harder to identify rhetorical questions. This is not surprising as it often requires complex commonsense knowledge to understand that a question is not seeking information, and we will show some examples in Section 6.

Between the first row and the second row, the recall for rhetorical questions increases by about 1%, while the precision for information-seeking questions goes up. This shows that the embedding and subjectivity features from the PTweet help discover rhetorical questions that were previously mislabeled as information-seeking. In the third row, recall and precision improves for both categories as the topic features are added. This implies that the topic features from the PTweet help to identify both rhetorical and information-seeking questions that were previously mislabeled.

6 Analysis

To better understand how topics interact with rhetorical and information-seeking questions, we analyzed the distribution of RQ and IQ over topics, based on the topic labels produced by the Google Content Classifier applied to the PSnippets from our training and development sets. Table 4 shows the four topics most highly correlated with each question category. The second column shows the total number of questions identified for each topic, and the third and fourth columns show the percentages of rhetorical and information-seeking questions in each topic.

Topic	Total	**RQ%**	**IQ%**
Computers & Electronics	85	20	80
Internet & Telecom	70	23	77
Games	94	24	76
Autos & Vehicles	34	29	71
Home & Garden	41	56	44
Law & Government	74	58	42
Books & Literature	42	60	40
Sensitive Subjects	41	71	29

Table 4: Topic Associations for RQs and IQs

The four topics most highly correlated with information-seeking questions are *Computers & Electronics*, *Internet & Telecom*, *Games*, and *Autos & Vehicles*. Our analysis found that this is because people tend to ask about the details of products and services in Twitter (e.g., sale price, features of computers, and release dates of games). On the other hand, the four topics most highly correlated with rhetorical questions are *Sensitive Subjects*, *Books & Literature*, *Law & Government*, and *Home & Garden*. We inspected examples from these topics and found that they are usually related to opinion expressions and debates (e.g., debates about race and politics, and assessment of books' quality), which lead to more rhetorical questions.

We also manually inspected some questions that seemed to be difficult for our system to label correctly. Table 5 shows some examples from our development dataset that were mislabeled by our best model. Example 1 requires the model to know that the question serves as a joke. In Example 2, the question, despite its simple question structure and lack of explicitly negative words, expresses a negative emotion. But without recognizing the implicit sentiment, it is hard to determine that the question is rhetorical. In Example 3, the model needs to understand the interaction between the prior context and the question to know that the question serves as a sarcastic response. Example 4 was classified as RQ probably because it is syntactically not in a complete question form (e.g., *"Are you going to*

[8] We also tried adding unigrams but the embedding worked better.

[9] None of the Question Features other than the Subjectivity Features are applicable to PTweets

prom or nah?"). The model mislabeled Example 5 probably because the question contains only a (complex) noun phrase and thus looks like a suggestion, which is a more common phenomenon in rhetorical questions.

Rhetorical Questions	
1	**PTweet**: Welcome anytime. You know where I live. **Tweet**: At the bottom of a sinkhole?
2	**PTweet**: Son of a .. How many blocked FG's do we have to endure. Going out of my mind. #smh **Tweet**: How does this keep happening?!
3	**PTweet**: Hans is a piece of crap. **Tweet**: Where were you like 4 months ago with that?
Information-seeking Questions	
4	**PTweet**: Have faith just have faith. **Tweet**: you going to prom or nah?
5	**PTweet**: Definitely going to send me that picture are you? Haha! **Tweet**: The one of your cheese on toast?

Table 5: Examples of RQ and IQ Mislabeled by the Best Model

7 Conclusions

A contribution of this work is that we have created a new dataset containing nearly 5,000 question tweets labeled as rhetorical or information-seeking coupled with their prior tweets. To our knowledge, this is the first Twitter-based dataset for studying rhetorical questions that has both human-generated gold labels and includes prior context for each question. We also presented classification models to benchmark performance on this task, and showed that including the tweet prior to a question improves performance. We also showed several ways to capture topic information, and that topic information represented in the preceding context seems to be useful for this task. Our hope is that this work will lead to further research on the role of context for recognizing rhetorical and information-seeking questions in social media.

Acknowledgement

We thank Tao Li (University of Utah) who provided expertise that assisted this work. We are also very grateful to Harald Illig, Changchen Chen, Ruijia Zhu and Fan Wu for their help in the preliminary data analysis.

References

Shohini Bhattasali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic identification of rhetorical questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 743–749. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Jane Frank. 1990. You call that a rhetorical question?: Forms and functions of rhetorical questions in conversation. *Journal of Pragmatics*, 14(5):723 – 738.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, HLT '11*, volume 2, pages 42–47.

John Godfrey, E.C. Holliman, and J McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 517–520.

Javier Gutierrez-Rexach. 1998. Rhetorical questions, relevance and scales. *Revista Alicantina de Estudios Ingleses*, 11:139–156.

Chung-hye Han. 2002. Interpreting interrogatives as rhetorical questions. *Lingua*, 11:201–229.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical questions and sarcasm in social media dialog. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*,

pages 310–319, Saarbrücken, Germany. Association for Computational Linguistics.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Suhas Ranganath, Xia Hu, Jiliang Tang, Suhang Wang, and Huan Liu. 2016. Identifying rhetorical questions in social media. In *Proceedings of the 10th International AAAI Conference on Web and Social Meidia*, pages 667–670.

Jerrold Sadock. 1971. Queclaratives. *Papers from the 7th Regional Meeting of the Chicago Linguistic Society*, pages 223–232.

Deborah Schaffer. 2005. Can rhetorical questions function as retorts?: Is the pope catholic? *Journal of Pragmatics*, 37(4):433 – 460.

Jürgen Schmidt-Radefeldt. 1977. On so-called 'rhetorical' questions. *Journal of Pragmatics*, 1:375–392.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Zhe Zhao and Qiaozhu Mei. 2013. Questions about questions: An empirical analysis of information needs on twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1545–1556, New York, NY, USA. ACM.