Determining the Accuracy of Crowdsourced Tweet Verification for Auroral Research

Nathan A. Case[1,2,3], Elizabeth A. MacDonald[1,2], Sean McCloat[2,4], Nick Lalone[5], and Andrea H. Tapia[5]

[1]NASA Goddard Space Flight Center, Greenbelt, MD, USA

[2]New Mexico Consortium, Los Alamos, NM, USA

[3]Department of Physics, Lancaster University, Lancaster, UK

[4]University of North Dakota, Grand Forks, ND, USA

[5]College of Information Sciences and Technology, Pennsylvania State University, State College, PA, USA

n.case@lancaster.ac.uk

Abstract

The Aurorasaurus citizen science project harnesses volunteer crowdsourcing to identify sightings of an aurora (or the "northern/southern lights") posted by citizen scientists on Twitter. Previous studies have demonstrated that aurora sightings can be mined from Twitter but with the caveat that there is a high level of accompanying non-sighting tweets, especially during periods of low auroral activity. Aurorasaurus attempts to mitigate this, and thus increase the quality of its Twitter sighting data, by utilizing volunteers to sift through a pre-filtered list of geo-located tweets to verify real-time aurora sightings. In this study, the current implementation of this crowdsourced verification system, including the process of geo-locating tweets, is described and its accuracy (which, overall, is found to be 68.4%) is determined. The findings suggest that citizen science volunteers are able to accurately filter out unrelated, spam-like, Twitter data but struggle when filtering out somewhat related, yet undesired, data. The citizen scientists particularly struggle with determining the real-time nature of the sightings and care must therefore be taken when relying on crowdsourced identification.

Keywords: twitter, crowdsourcing, aurora, sightings, citizen science

Introduction

The citizen science project Aurorasaurus (MacDonald et al., 2015) has two main space weather related goals: improving the "nowcasting" of a visible aurora (commonly known as the "northern/southern lights") and the ability to accurately model both the size and strength of an aurora. To do this, it collects observations of the aurora made by the general public. These observations can be submitted directly to the project, via its website (http://aurorasaurus.org) and mobile apps, and are found by searching Twitter for possible sightings.

Twitter can be a useful resource of data for many citizen science projects, as information is freely shared by millions of users distributed around the globe. Indeed, previous studies have shown that Twitter users, who post short updates (of a maximum 140 characters in length) known as "tweets", will often share details about the conditions around them. This is especially true for large-scale events such as earthquakes (Earle et. al, 2010; Crooks et al., 2013), influenza outbreaks (Culotta, 2010; Lampos, De Bie & Cristianini, 2010), and service outages (Motoyama et al., 2010). Case et al. (2015a) showed that this was also true for the aurora by comparing the number of tweets relating to an aurora with auroral activity (or, rather, to common auroral activity indices). However, they also noted that Twitter data is particularly noisy and that many tweets which contain aurora-related keywords (e.g. "aurora" and "northern lights") are not actually sightings (instead they may be about a person or place, or about the desire to witness an aurora).

As such, the Aurorasaurus project enlists volunteers (registered and anonymous), who themselves can be thought of as citizen scientists, to sort through pre-filtered, aurora-related, tweets to identify and positively verify real-time aurora sightings. Whilst combining Twitter data with other citizen science data is quite rare, and this exact application of crowdsourcing may be new, many previous studies have demonstrated that crowdsourcing can be used for classification of data - often using Amazon's Mechanical Turk (Kittur, Chi & Suh, 2008; Ipeirotis, Provost & Wang, 2010). In fact, studies have shown that the crowd is sometimes even more accurate than the expert at identification tasks (Alonso & Mizzaro, 2009).

Once a tweet has been verified as a positive sighting by the Aurorasaurus volunteers, it is treated in the same way as a direct report via the project's website or apps. In this way, Aurorasaurus uses sightings on Twitter to supplement those citizen science reports submitted directly to it, thus maximizing the available data. The observations, both positively verified tweets and direct reports, are displayed on a real-time map, alongside a modeled auroral oval (i.e. the extent to which an aurora is visible directly overhead), which is shown on the project's homepage. These observations serve several different functions, including: demonstrating where the aurora is currently being observed (Priedhorsky, MacDonald & Cao, 2012), providing data points for scientific investigation (Case, MacDonald & Viereck, 2016), and providing the basis for a hybrid alert system (Lalone et al., 2015) that is analogous to disaster early warning systems (Tapia et al., 2014). The process of how the tweets are found, presented to the volunteers, and verified is discussed in the following section.

In this study, using real Twitter data, collected by an operational citizen science project, the accuracy of the volunteer crowd at filtering useful data from a stream of tweets is investigated for the first time. The results presented herein, whilst related to one natural phenomenon, can provide insights into the accuracy of the volunteer crowd in analysing Twitter data for other citizen science projects too.

Tweet Verification

Aurorasaurus exploits the Twitter Search API to identify publicly-accessible tweets that contain any one of several different aurora-related keywords (e.g. "aurora", "northern lights"). The returned tweets are then further filtered on the Aurorasaurus servers to exclude most retweets, tweets from Twitter users with "aurora" in their username (though a whitelist is maintained to allow tweets from some users through), and tweets containing profanity or other common "spam" terms.

A location extraction process is then undertaken on these filtered tweets. The location is either determined using the embedded GPS meta-data, if the Twitter user has opted to share their location, or the tweet is run through the geo-parsing software CLAVIN (https://clavin.bericotechnologies.com), which attempts to extract a location for the tweet based upon its text (D'Ignazio et al., 2014). Using this process, approximately 15% of the tweets have a location associated with them (with extraction through CLAVIN accounting for approximately 80% of those). Further filtering then takes place to remove tweets whose location is determined as anywhere containing the term "Aurora" (e.g. Aurora, CO, USA).

These tweets, known as the "unverified tweets", are then presented to the Aurorasaurus community for verification. An example of such a tweet is given in Figure 1. The community is asked "Did they just see the aurora?" (where "they" refers to the tweet's author) and have only two choices for their vote ("yes" or "no"). This subjective task allows for automatic aggregation of the votes into a score and classification based upon that score (Iren & Bilgen, 2014).

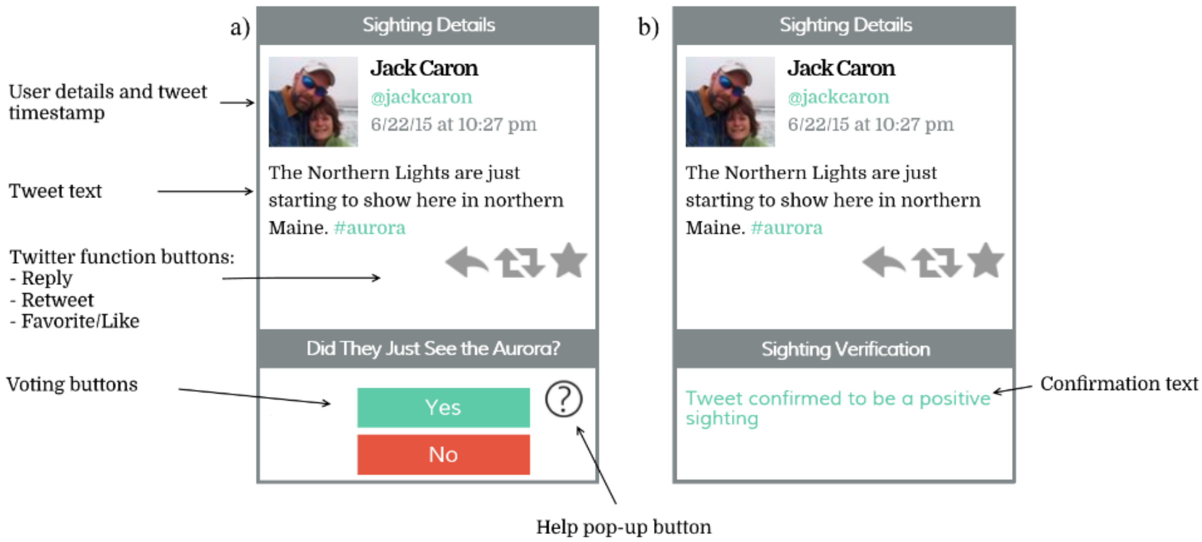

Figure 1. a) An example tweet, as presented to the Aurorasaurus community for verification. The volunteers are asked "Did they just see the aurora?" and are given the two simple options of "yes" (for a positive, real-time, aurora sighting) or "no". b) Once a threshold positive score is reached, the tweet is then confirmed as a "positive sighting" and becomes known as a "positively verified tweet". It is then no longer available for further voting.

For every "yes" vote a tweet receives, +1 is added to its score. Conversely, for every "no" vote a tweet receives, -1 is added to the score. Votes from registered users and anonymous users are treated equally (i.e. there is no weighting applied to the vote based upon the user or their credentials). Once the tweet's score reaches a certain positive threshold, it is categorized as a "positively verified tweet", its marker is updated on the map to show this new status, and votes are no longer taken on it. Similarly, once a tweet reaches a certain negative threshold, the vote is categorized as a "negatively verified tweet", the marker is removed from the map, and the tweet is no longer presented to the community for verification.

To reduce the barriers of entry for users to start verifying tweets, there is no compulsory training required. However, help in how to verify a tweet is provided by a pop-out help menu, which opens if the user clicks on the question mark in the tweet window (see Figure 1). Additionally, a blog post and a quiz are available which both guide the voter through examples of tweets and how they should be voted upon. Approximately half of the respondents to a recent Aurorasaurus survey indicated that they had read at least some of this guidance (Lalone, pers. comm., 2015).

In this study, the verified tweets posted during March and April 2015 are analyzed. This two month period represents a subset of the larger Aurorasaurus data set (which spans from November 2014 to present) and includes several large auroral events, including the largest auroral event this decade (Case, MacDonald & Patel, 2015). It is important to note that large auroral events, where an aurora can be seen from the mid-US and central Europe, are relatively infrequent and are dependent upon several factors - including solar activity, time of day/year, and local conditions (e.g. cloud cover). Additionally, an aurora can be a widespread phenomenon, with sightings of the same event spanning multiple continents (Case, MacDonald & Patel, 2015).

Results

During March and April 2015, 227,280 aurora-related tweets were collected with 39,636 (17.4%) having a location associated with them and thus were available for the Aurorasaurus community to vote on. Of these, the Aurorasaurus community verified 4,547 (11.5%) tweets: 475 positively (10.4%) and 4,072 negatively (89.6%). There were 70,331 votes cast: 49,495 by logged-in users (70.4%) and 20,836 by anonymous users (29.6%).

The distribution of the tweets and their verified status is shown in Figure 2. The number of each type of tweet ("total", "with location", "positively verified", "negatively verified" and “unverified”) is shown by the filled bars. Note the logarithmic scale on the y-axis.

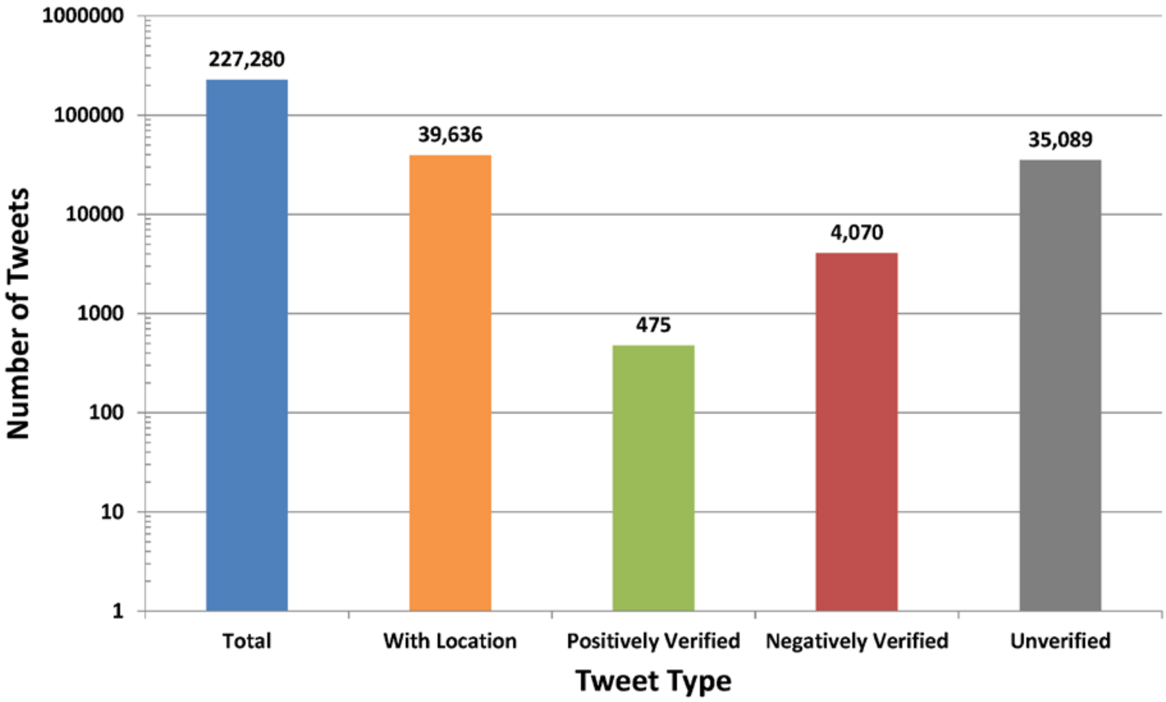

Figure 2. The distribution of the tweets collected during March and April 2015. The first (blue) bar indicates the total number of tweets collected. The second (orange) shows the number of tweets with a location associated with them, and thus available for the Aurorasaurus community to vote on. The third (green) bar shows the number of positively verified tweets whilst the fourth (red) shows the number of negatively verified tweets. The final (gray) column is the number of tweets that were not verified (i.e. “unverified”).

Each of the positively verified tweets was then independently manually inspected by two members of the Aurorasaurus team. This inspection involved analyzing the text of the tweets in detail to identify any signs of non-originality and comparing the location and time of the supposed sighting with auroral models and other citizen science observations.

The verified tweets were categorized primarily into "valid" (where the tweet was indeed a real-time aurora sighting made by the tweet's author) or "invalid" (where the tweet was incorrectly positively verified by the users). Using an open-coding method, the following categories for the invalid positively verified tweets were created:

- "Not real-time": a sighting of an aurora by the tweet's author, however, the tweet was posted at least several hours after the sighting took place (often the next morning).

- "Not original": the sighting was not made by the tweet's author (usually "retweets" or "mentions" of someone else's tweet).
- "Overlap": the sighting was both not real-time nor was it made by the tweet's author. This would often be the retweeting of someone else's aurora photograph.
- "Wrong location": the location extraction algorithm (CLAVIN) failed to determine the location correctly. These failures are particularly difficult for the voters to spot since the location of the tweet is not shown on the tweet (see Figure 1).
- "Not positive sighting": the tweet did not contain a sighting of an aurora but may have been related to one (e.g. "Seeing an aurora is on my bucket list").
- "Junk": these tweets had nothing to do with an aurora (e.g. "Went to Aurora last night").

The distribution of these categories is shown in Figure 3.

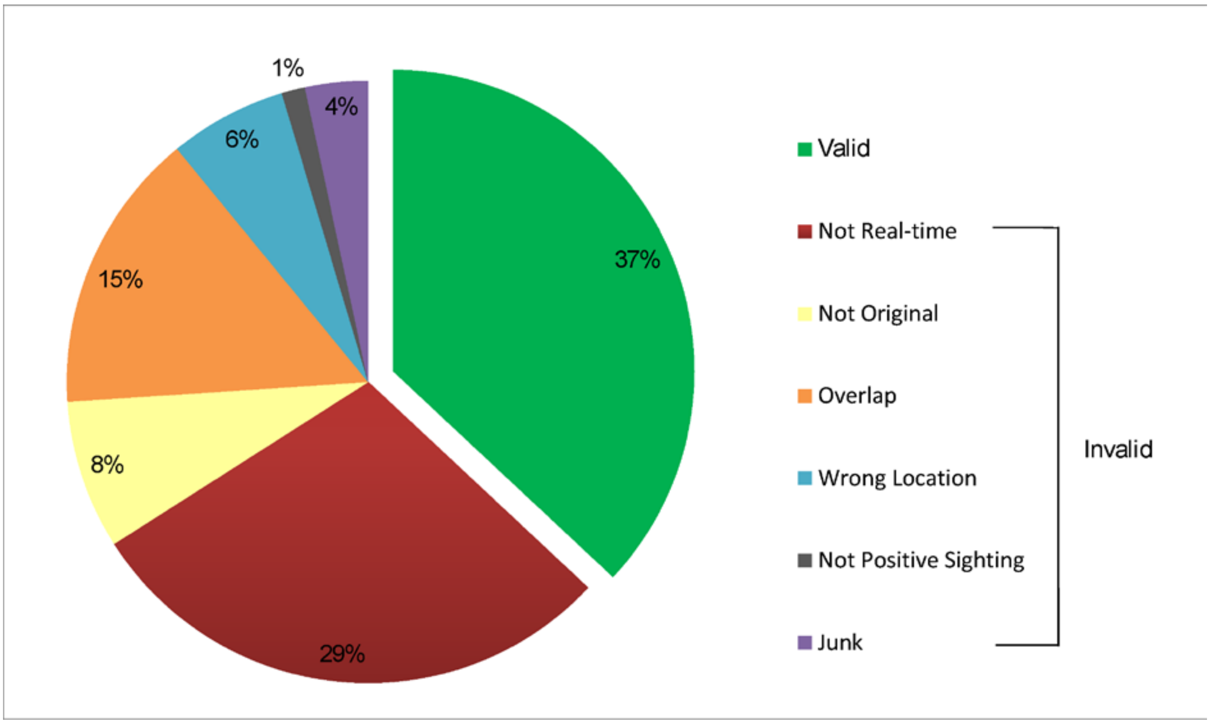

Figure 3. The distribution of the positively verified tweets collected during March and April 2015. The tweets are grouped by the previous categories: valid (green), not real-time (red), not original (yellow), overlap (orange), wrong location (blue), not a positive sighting (black) and junk (purple).

Of the 475 positively verified tweets, 176 (37%) are valid. The precision, or positive predictive value (PPV), as calculated using Equation 1, of the positively verified tweets is therefore 37.1%.

$$PPV = \frac{\Sigma TP}{\Sigma TP + \Sigma FP} \quad (1)$$

where ΣTP is the number of true positives (i.e. positively verified tweets that are valid) and ΣFP is the number of false positives (i.e. positively verified tweets that are invalid).

The process was then repeated for a sample of the negatively verified tweets. This randomly selected sample included 475 negatively verified tweets (chosen to match the number of positively verified tweets). All but two of the tweets in the sample were correctly identified as negatively verified tweets. Thus, the "negative precision", or negative predictive value (NPV), as calculated using Equation 2, was 99.6%.

$$NPV = \frac{\Sigma TN}{\Sigma TN + \Sigma FN} \quad (2)$$

where ΣTN is the number of true negatives (i.e. negatively verified tweets that are not valid sightings) and ΣFN is the number of false negatives (i.e. negatively verified tweets that are actually valid sightings).

The overall accuracy of the verified tweets, in which all of the positively verified tweets and a same-sized sample of negatively verified tweets are included, can now be determined. Using Equation 3, the overall accuracy is found to be 68.4%.

$$ACC = \frac{\Sigma TP + \Sigma TN}{N} \quad (3)$$

where N is the total number of verified tweets in this sample (i.e. N=950).

Furthermore, these results can be decomposed based upon periods of when auroral activity was particularly elevated (which is when most sightings would be expected to occur). There were three such events during this time period: March 01 - 03, March 17 - 19 and April 10 - 12. The distributions of the previous categories are shown, for each of these periods, along with the distribution of "non-elevated" periods, in Figure 4.

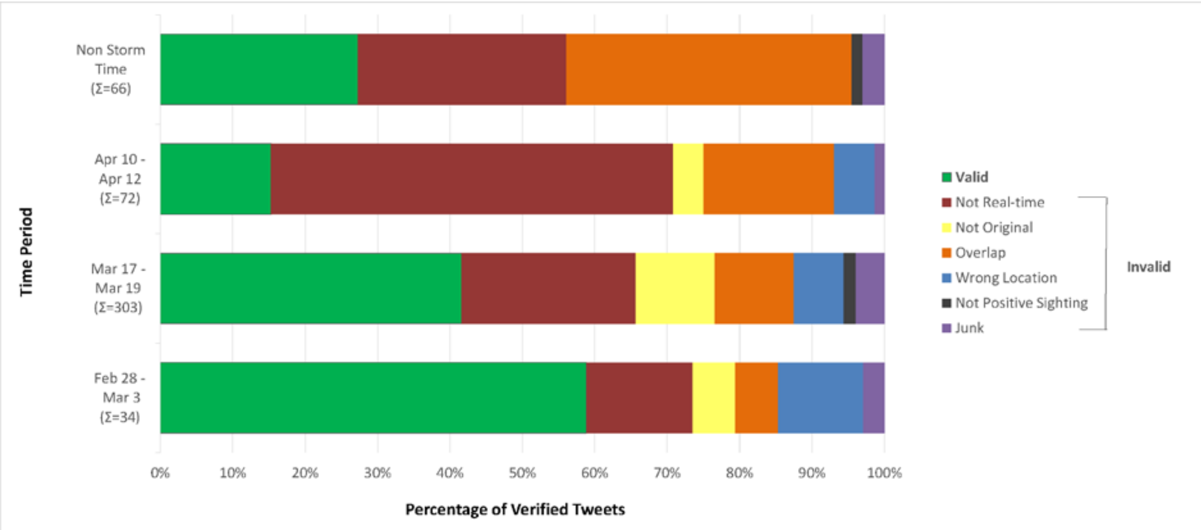

Figure 4. The positively verified tweets have been split into three active auroral time and one non-storm time periods. For each period, the percentage share of each of the categories listed earlier is shown.

The negatively verified tweets were also split by storm period. Both of the invalid negatively verified tweets occurred during the March 17-19 storm (which is not particularly surprising due to the majority of tweets occurring during this time). The PPV, NPV and ACC are calculated for each of these storm periods and are presented in Table 1.

Period	N	N_{pos}	N_{neg}	PPV (%)	NPV (%)	ACC (%)
March 01 - 03	44	34	10	58.8	100.0	79.4
March 17 - 19	461	303	158	41.6	98.7	70.2
April 10 - 12	117	72	45	16.7	100.0	58.4
Non-Storm Time	328	66	262	27.3	100.0	63.7
Overall	950	475	475	37.1	99.6	68.4

Table 1. Tweet numbers and verification accuracy, split by periods of auroral activity.

Discussion

Approximately 17.4% of the 227,280 tweets collected during this case study had a location associated with them, which is consistent with other studies (e.g. Vieweg et al., 2010). Thus, nearly 40,000 tweets were available for the Aurorasaurus community to vote on. Approximately 75% of the locations obtained were determined using the CLAVIN geo-location extraction algorithm and thus only a small percentage of the total tweets contained an embedded GPS location. Again, this result is consistent with other studies (e.g. Cheng, Caverlee & Lee, 2010; Lee, Srivatsa & Mohapatra, 2013).

The community cast over 70,000 votes and verified over 4,500 tweets. The vast majority, around 80%, of those verified tweets were negatively verified, i.e. the Aurorasaurus community voted that the tweet was not a real-time sighting of an aurora made by the tweet's author. This result is perhaps unsurprising, since it is generally only when auroral activity is high (which occurred three times during this case study) that more people tweet sightings of an aurora (Case et al., 2015a). Indeed, as can be determined from Table 1, the percentage of positively verified tweets (i.e. N_{pos}/N) rises from around 20% during non-storm times to around 70% during active times.

Notably, nearly 90% of the tweets with locations went unverified (i.e. they were not positively or negatively verified). These tweets are most likely not aurora sightings, rather they are tweets that contain aurora-related keywords, however, we cannot be certain that this set does not contain sightings that have simply been overlooked. Whilst this does not affect the accuracy of the verification system, it does mean that some scientifically useful observations, such as rare sightings during low auroral activity, might be being missed. Further investigation into the exact nature of the unverified tweets, and what effect this may have on citizen science data collection on Twitter, should therefore be undertaken.

Verification Accuracy

The Aurorasaurus community was able to negatively verify tweets with extremely high accuracy. In fact, of the 475 negatively verified tweets analyzed, only two were incorrectly classified - resulting in an overall NPV of nearly 100%. The community was, however, much less accurate when positively verifying tweets. The overall PPV (or precision) was 37%, though there was significant variance in the PPVs when splitting by event (with the highest PPV of 59% occurring during the March 01-03 storm and the lowest PPV of 27% occurring during the April 10-12 storm). At this time, there is no known reason for this variance except, perhaps, for differences in the sample sizes.

The overall accuracy of the verification system, in this case study, was 68%. Though had all of the negatively verified tweets been analysed, and subsequently used in the accuracy calculation, the overall accuracy would probably be much higher. However, since the number of negatively verified tweets was so much greater than the number of positively verified tweets, a representative sample was instead chosen. It is also important to note that the positively verified tweets (i.e. actual sightings) hold the most scientific value and so the PPV is more important, perhaps, than the NPV or overall accuracy.

What affected the community's precision?

It is relatively easy to spot spam-like tweets that have nothing to do with a sighting of an aurora. It is much harder, however, to differentiate between tweets that are real-time sightings of an aurora from those that are just related to the aurora or are true sightings that occurred several hours ago. Indeed, our analysis showed that the primary reason the community wrongly positively verified a tweet was that community incorrectly identified the tweet as being real-time.

Identifying whether or not a sighting posted in a tweet is real-time can be a complex task - even for the Aurorasaurus team members. The tweet, of course, has a timestamp associated with it but the tweet's author may be posting about a sighting that occurred several hours ago or perhaps even the day before. Unless the author explicitly uses words or phrases that chronological identify when the sighting occurred, e.g. "just seen" or "spotted 10 mins ago", it is difficult to know when exactly the sighting occurred. In fact, even if the author includes a time, e.g. "aurora seen at 21:30", the verifier would need to know the offset between their current time zone and the time zone of the tweet's author to determine how long ago that time was. Such detailed investigation is probably too much for most of the community to engage in, especially when they are voting on many tweets in one go.

The second most common reason for incorrectly positively verifying a tweet was that the sighting was "not original". From this category we identified two themes: the tweet was of someone else's aurora photograph (85%) or the tweet was a retweet of somebody else's sighting (15%). Both of these errors would seem to stem from some members of the community being unfamiliar with the particular nomenclature of Twitter. For example, most of the "not original" tweets contained signs of the non-originality, i.e. the text "RT" (an acronym for retweet) or tagging of other users (which will always start with the @ symbol). We note, however, that many original real-time sightings may also tag other users, often as a way of alerting those other users, so this method to determine originality cannot be used on its own.

Improving the voting system

It is assumed that when the community has incorrectly positively verified a tweet it is the result of an "honest mistake" rather than "cheaters" (i.e. those with malicious intent) since there is no gain to poor verification (Hirth, Hoßfeld & Tran-Gia, 2013; Iren & Bilgen, 2014). Therefore, one of the primary ways to improve the accuracy of the crowd is to improve the information provided about the task and the desired outcome (Iren & Bilgen, 2014). Aurorasaurus currently provides its community with instructions/guidance via a help page, blog post, and a quiz (where members of the community can test their voting skill and receive feedback on their choices). However, these are all "hidden elements" in that the user may not have even seen them before they start voting. Indeed, a recent survey of Aurorasaurus users showed that 40% of users did not know that instructions on how to verify tweets were available (Lalone, pers. comms., 2015).

Additionally, enforcing training upon the community before they are able to vote has shown to improve the quality of their voting (e.g. Le et al., 2010). In some implementations, such training results in a pass/fail that screens out untrustworthy or inaccurate users (Downs et al., 2010, Le et al., 2010). In others, the scoring mechanism of each voter is weighted based upon how well they performed during the training (Sheng et al., 2014). We note, however, that these studies often employ contributors through Amazon's Mechanical Turk rather than volunteers through citizen science projects.

Since the Aurorasaurus project, like all citizen science projects, is reliant on volunteers, adding such compulsory activities might reduce the number of people who are willing to participate. Therefore, training that is not compulsory but that could be used to better inform the voting system on a user's trustworthiness might be desirable. For example, votes from anonymous users might be weighted to score 1, votes from registered users who have not taken the training might be weighted to score 2, votes cast by those who have taken the quiz but did not score highly might be weighted to 3 and votes from users who scored highly in the quiz might be weighted to 5. Project staff, or trusted super-users, might then have an even higher voting weight. This approach has the benefit of determining a pseudo confidence level for each vote whilst also not erecting barriers to participation.

Vuurens, de Vries & Eickhoff (2011) demonstrated that a "combined consensus algorithm", which generally used a majority vote but then took into account the voters' trustworthiness in a tie situation, consistently gave the most accurate results. A tied result, with respect to the Aurorasaurus crowdsourcing system, would be where the number of votes is over the verification threshold, however, the score has not exceeded said threshold (i.e. 10 users vote: five yes and five no, thus resulting in a score of 0).

The training, and perhaps subsequent vote weighting, is likely to be a one-time effort (though, in practice, users could be allowed to complete it more than once). This may lead to situations where the user forgets what they have been taught or their voting is affected by other factors (e.g. fatigue or lack of concentration). In this case, an adaption of the "majority decision" cheat detection method (Hirth, Hoßfeld & Tran-Gia, 2013) could be employed. If a member of the community votes against the current majority decision, or perhaps the decision of a trusted voter (e.g. staff or super-user), they are advised of this in real-time and are offered training/guidance on how they should vote. The frequency to which a user matches, or rather does not match, the majority can be stored, allowing for a hybrid voting reputation to be built (Voyer et al., 2010). Based on this reputation voting weights could again be applied.

In addition to improving the voting mechanism itself, another way to perhaps increase the quality of the verification process is to improve the chance of a tweet being a valid sighting before presenting it to the community for validation. The current system is somewhat basic in that it simply uses a set of keywords for searching and another set for filtering. Machine learning, based on either a gold standard set or the community's voting, might improve the quality of the tweets being served to the community (Wang, 2010; Becker, Naaman & Gravano, 2011; Truong et al., 2014). This approach was tested early on in the Aurorasaurus project, however, it failed to yield any noticeable improvements (MacDonald, pers. comms., 2015), indicating that further refinement may be needed on such an approach before it could be successfully applied to this task.

Conclusion

Like many citizen science projects, Aurorasaurus is heavily reliant upon a community of volunteers for both providing data and for validating/classifying data. To compliment the aurora sightings reported directly to the project, Aurorasaurus also systematically searches for observations of an aurora posted on Twitter, using the Twitter Search API and several rudimentary filters. A location is required for all sightings and so those tweets that do not contain an embedded location are passed through a location extraction algorithm which attempts to resolve a location for the tweet based upon its text. This process, whilst not always accurate, increases the number of usable tweets four-fold. Using a similar location extraction process is therefore recommended for other citizen science projects needing location data from tweets. Including Twitter as a data source has increased the number of observations for the Aurorasaurus project by nearly 100%. Exploiting Twitter as an available data source is therefore recommended for other citizen science projects that collect observational data.

Twitter observations are noisier than traditional citizen science reports, however, and so need more curation by both the volunteers and project staff alike. The Aurorasaurus community is therefore encouraged to verify these potential sightings using a simple crowdsourcing scoring system. The community is rewarded for their participation by a leader board, where each votes earns the volunteer 5 points, and by increased accuracy in localized auroral visibility alerts.

This Aurorasaurus case study has shown that volunteer citizen scientists are extremely adept at filtering out spam-like tweets and other non-aurora sightings. These tweets tend to form the majority of tweets presented to the Aurorasaurus community - especially during time where there is little auroral activity. For the random sample studied, the NPV of the "negatively verified" tweets was almost 100%. A good NPV is perhaps unsurprising, as it is a relatively easy task, though such a high score was somewhat unexpected. The volunteer community proved to be less accurate when identifying the true aurora sightings, however. The PPV, or precision, of the positively verified sightings was somewhat poor at 37%. The most common reason for the community incorrectly positively verifying a tweet was that the tweet was not real-time, followed by the tweet not being an original sighting.

Whilst positively verifying tweets requires more detailed investigation than filtering out spam-like tweets, it would seem that the PPV achieved could certainly be improved. As discussed, it is likely that any incorrect identifications were the result of honest mistakes and so the primary way to reduce these is to provide training for the community. Aurorasaurus does provide some level of training, though it is not compulsory. The "verifying tweets quiz", which is the only interactive training offered, is detached from the verification process itself in that it is a completely separate entity and is not linked to when verifying tweets. Making any training compulsory will likely have a detrimental effect on the number of users who then participate in the verification process (Lintott, pers. comms., 2015). This is a quality control cost that many projects have to deal with (Iren & Bilgen, 2014). However, small improvements, such as providing a link to the quiz during the verification process, are likely to increase the community's accuracy, even if just a little, without affecting the number who are willing to participate.

Larger, systematic improvements, such as implementing vote weighting algorithms or the adaption of a real-time majority decision cheat detection system, are likely to significantly improve the quality (particularly the PPV) of the community's verification efforts. Such improvements will obviously take time and resources to implement but should be on the future road map for the project.

The results of this case study suggest that other citizen science projects which plan on using volunteer crowdsourcing for data validation, especially when the data are particularly noisy (e.g. tweets), also consider using some of the aforementioned training or quality control methods to increase the validation accuracy of the crowd. The information provided on Twitter by citizen scientists, and then verified by other volunteers, can be extremely useful. However, as this study shows, consideration must be given to the training of those volunteers who validate the data else the accuracy of the crowd can be poor.

Acknowledgments

This material is based upon work supported, in part, by the National Science Foundation (NSF) under Grant #1344296. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

Funding for SM was kindly provided by the North Dakota Space Grant Consortium.

References

Alonso, O and Mizzaro, S. 2009 Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation, Vol. 15. 16.

Becker, H, Naaman, M, and Gravano, L. 2011 Beyond Trending Topics: Real-World Event Identification on Twitter, *ICWSM 11*, 438–441.

Case, N. A, MacDonald, E. A, Heavner, M, Tapia, A. H, and Lalone, N. 2015 Mapping auroral activity with Twitter. *Geophys. Res. Lett.,* 4, 3668–3676. DOI: http://dx.doi.org/10.1002/2015GL063709

Case, N. A, MacDonald, E. A, and Patel, K. G. 2015 Aurorasaurus and the St Patrick’s Day storm. *Astronomy & Geophysics,* 56, 3, 3.13–3.14. DOI: http://dx.doi.org/10.1093/astrogeo/atv089

Case, N. A, MacDonald, E. A, and Viereck, R. 2016 Using citizen science reports to define the equatorial extent of auroral visibility, *Space Weather*, 14, DOI: http://dx.doi.org/10.1002/2015SW001320.

Cheng, Z, Caverlee, J, and Lee, K. 2010 You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 759–768.

Crooks, A, Croitoru, A, Stefanidis, A, and Radzikowski, J. 2013 # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS,* 17, 1, 124–147.

Culotta, A. 2010 Detecting influenza outbreaks by analyzing Twitter messages. *CoRR*. arXiv:1007.4748.

Dawid, A. P and Skene, A. M. 1979 Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, 20–28.

D’Ignazio, C, Bhargava, R, Zuckerman, E, and Beck, L. 2014 Cliff-clavin: Determining geographic focus for news. NewsKDD: Data Science for News Publishing, at KDD 2014.

Downs, J. S, Holbrook, M. B, Sheng, S, and Cranor, L. F. 2010 Are your participants gaming the system?: screening mechanical turk workers. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2399–2402.

Earle, P, Guy, M, Buckmaster, R, Ostrum, C, Horvath, S, and Vaughan, A. 2010 OMG earthquake! Can Twitter improve earthquake response? *Seismological Res. Lett*. 81, 2, 246–251. DOI: http://dx.doi.org/10.1002/2015GL063709

Hirth, M, Hoßfeld, T, and Tran-Gia, P. 2013 Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling.* 57, 11–12, 2918–2932. DOI: http://dx.doi.org/10.1016/j.mcm.2012.01.006

Ipeirotis, P. G, Provost, F, and Wang, J. 2010 Quality management on amazon mechanical turk. In Proceedings of the ACM SIGKDD workshop on human computation. ACM, 64–67.

Iren, D and Bilgen, S. 2014 Cost of Quality in Crowdsourcing. *Human Computation* 1, 2, 283–314. DOI: http://dx.doi.org/10.15346/hc.v1i2.14

Kittur, A, Chi, E. H, and Suh, B. 2008 Crowdsourcing User Studies with Mechanical Turk. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’08). ACM, New York, NY, USA, 453–456. DOI:http://dx.doi.org/10.1145/1357054.1357127

LaLone, N, Tapia, A. H, Case, N. A, MacDonald, E. A, M., H, and Heavner, M. 2015 Hybrid Community Participation in Crowdsourced Early Warning Systems. In Proceedings of the ISCRAM 2015 Conference.

Lampos, V, De Bie, T, and Cristianini, N. 2010 Flu Detector - Tracking Epidemics on Twitter. In Machine Learning and Knowledge Discovery in Databases, JosÃl’Luis BalcÃa¸zar, Francesco Bonchi, Aristides Gionis, and MichÃ´lle Sebag (Eds.). Lecture Notes in Computer Science, Vol. 6323. Springer Berlin Heidelberg, 599–602. DOI: http://dx.doi.org/10.1007/978-3-642-15939-8_42

Le, J, Edmonds, A, Hester, V, and Biewald, L. 2010 Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In SIGIR 2010 workshop on crowdsourcing for search evaluation. 21–26.

Lee, K, Ganti, R, Srivatsa, M, and Mohapatra, P. 2013 Spatio-temporal provenance: Identifying location information from unstructured text. In Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on. IEEE, 499–504.

MacDonald, E. A, Case, N. A, Clayton, J. H, Hall, M. K, Heavner, M, Lalone, N, Patel, K. G, and Tapia, A. H. 2015 Aurorasaurus: a Citizen Science Platform for Viewing and Reporting the Aurora. *Space Weather*. DOI: http://dx.doi.org/10.1002/2015SW001214

Motoyama, M, Meeder, B, Levchenko, K, Voelker, G. M, and Savage, S. 2010 Measuring online service availability using twitter. *WOSN’10*, 13–13.

Priedhorsky, R, MacDonald, E, and Cao, Y. 2012 First Solar Maximum with Social Media: Can Space Weather Forecasting be Improved?. In AGU Fall Meeting Abstracts, Vol. 1. 2324.

Sheng, K, Gu, Z, Mao, X, Tian, X, Gan, X, and Wang, X. 2014 Answer inference for crowdsourcing based scoring. In Global Communications Conference (GLOBECOM), 2014 IEEE. IEEE, 2733–2738.

Tapia A. H, Lalone, N, MacDonald, E. A, Hall, M, Case, N. A, Heavner, M. 2014 AURORASAURUS: Citizen Science, Early Warning Systems and Space Weather, In Second AAAI Conference on Human Computation and Crowdsourcing.

Truong, B, Caragea, C, Squicciarini, A, and Tapia, A. H. 2014 Identifying valuable information from twitter during natural disasters. *Proceedings of the American Society for Information Science and Technology* 51, 1 (2014), 1–4. DOI: http://dx.doi.org/10.1002/meet.2014.14505101162

Vieweg, S, Hughes, A. L, Starbird, K, and Palen, L. 2010 Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 1079–1088.

Voyer, R, Nygaard, V, Fitzgerald, W, and Copperman, H. 2010 A hybrid model for annotating named entity training corpora. In Proceedings of the fourth linguistic annotation workshop. Association for Computational Linguistics, 243–246.

Vuurens, J, de Vries, A. P, and Eickhoff, C. 2011 How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR’11). 21–26.

Wang, A. 2010 Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach. In Data and Applications Security and Privacy XXIV, Sara Foresti and Sushil Jajodia (Eds.). Lecture Notes in Computer

Science, Vol. 6166. Springer Berlin, Heidelberg, 335–342. DOI: http://dx.doi.org/10.1007/978-3-642-13739-6_25