Harnessing Social Media for Environmental Sustainability: A Measurement Study on Harmful Algal Blooms

Vinay Boddula, Awani Joshi and Lakshmish Ramaswamy
Department of Computer Science
University of Georgia
Athens, Georgia 30602
Email: boddula,joshi,laks@cs.uga.edu

Deepak Mishra
Department of Geography
University of Georgia
Athens, Georgia 30602
Email: dmishra@uga.edu

***Abstract*—In recent years, social media has revolutionized citizen science activities. Given its popularity among people and communities, these social media services could be used effectively for environmental surveillance. However in social media, people use different terms to refer to same event for example, Blue Green Algae, Cyanobacteria, Algae Bloom and Red Tide refer to same event but one is very technical and other is more generic term. The technical terms are normally known to field experts or the domain scientists which inherently would mean more reliable information on social media but the more generic term is used by people of various backgrounds putting a question on the trustworthiness of the post. Moreover, the user base and the number of posts for more technical terms are relatively less compared to the generic terms. But the dichotomy is that the more common the term, the more noisy the data. One can say using generic terms to track the environmental events would be more effective. But the social media data has lot of flux thus using train once and classify ever model of machine learning will miss to classify many of the relevant events as shown in the paper. Our research seeks to explore the various opportunities, challenges and approaches in using social media for environmental monitoring.**

Keywords- (citizen science, environmental surveillance)

I. Introduction

In many scientific domain, involving lay citizens in scientific efforts also called as Citizen Science has been the key parameter to attain the larger goal of data collection. Until very recent, this was quite challenging as there was no scalable mechanism to involve people to collaborate from larger demographic area. However with the proliferation of Online Social Media services such as Twitter, Facebook etc. these services can be used as potential game changer as it provides very easy and cost effective platform for common people to participate in citizen science.

In the past, online social media has been used in multi facet social data analytics applications such as sentimental analysis, reporting health trends, reporting environmental disaster viz LITMUS [1] etc. Using the posts from the globally dispersed individuals, researchers have been able to monitor health and environmental trends. For example, Courtney, et al, Schmidt [2], Achrekar, et al [3] have successfully demonstrated the feasibility of tracking the trends of infectious diseases such as influenza, flu using Tweets from Twitter. Similarly Power, et al [4] and Sakaki, et al [5] have used Tweets to detect the environmental disaster viz. fire and earthquakes.

Another category of citizen science is using online social media for environmental sustainability which has not been explored widely in research community. When used successfully, it can fill the gap where the traditional methods of data collection for environmental sustainability via field trips, surveys, sensors or satellites face serious challenges. For example, field trips are not scalable to large areas, poor weather such as cloud cover can completely hinder the field and satellite data collection activity and the harsh environment can lead to sensor failures making the data collection infeasible. Online social media can fill the gap by expanding the observer base to include not only environmental scientists and restoration officials but also the lay community at large, including area residents and tourists to encourage more frequent and comprehensive environmental monitoring.

Events of interest could be tracked through keywords or hashtags. As observed, some of the more obvious hashtags such as #earthquake, #ebola and offhand hashtags such as #BlackLivesMatter, #occupycentral etc. becomes quickly popular and trending in the social media generating tens of thousands of related post which can be attributed due to the natural coverage by newspapers and televisions. Whereas for long term and slowly evolving disasters such as water quality degradation by harmful algal blooms or soil erosion, citizen science poses a different challenge because it doesnt generate media attention. Hence it is extremely hard to spread the hashtag created by environmental scientists among public for monitoring, leading one to depend on existing hashtags for citizen science.

However, there are some inherent issues in monitoring such hashtags as pointed out by our research. First, what term should one monitor for? Some phenomena are described by different terms. For example Cyanobacteria, Algae Bloom, Blue Green Algae and Red Tide are different keywords but they refer to similar organism that contaminate water. Second, lower reliability of popular but lesser technical term. The posts with a generic term are known widely among non-expert population hence the reliability cannot be guaranteed when a non-expert posts about it. Third, train once classify ever model doesnt work. The social media data has high flux leading to concept

drift. The train once and classify ever model would eventually fail to provide desired classification measure over time.

In this paper, we do an empirical study on an important issue of environmental sustainability viz. Cyanobacterial Harmful Algal Blooms (CyanoHABs). CyanoHABs are a class of bacteria found particularly in lakes, ponds and ocean water which produces cyanotoxins. These toxins can cause shellfish poisoning, fish kills and can also be fatal to animals and humans. We collected the data from Twitter for four keywords representing CyanoHABs - Red Tide, Cyanobacteria, Algae Bloom and Blue Green Algae. Our empirical study finds that keywords have impact on effectiveness of environmental monitoring through citizen science as certain keywords are not useful for environmental sustainability. We also characterize the users and show the usage of keywords as per the degree of technicality of the terms. We also studied effectiveness of machine learning algorithms in classification of relevant and irrelevant tweets on these keywords and show that there is a significant concept drift over a period of time which can be contoured with small amount of retraining to significantly increase the performance of the algorithms.

II. BACKGROUND

Citizen science involves mechanisms through which the non-scientist citizen can meaningfully contribute to gather the data for scientific research. Due to the limitations of traditional data collection infrastructures, citizen science has been quite helpful in projects related to species exploration, water/air quality monitoring, weather forecasting through social sensing where sensors havent been placed yet etc.But most of these research studies need people to be already aware of such projects and report the collected data at a particular destination thus insufficiently taking advantage of citizen science.

Fig. 1: Cyanobacteria Bloom in Georgia Pond.

Cyanobacterial Harmful Algal Blooms and Meghdoot Project:-

CyanoHABs are a major water quality and public health issue in the inland waters and estuarine environments where they can hamper recreational activities, degrade aquatic habitats through fish kills, and potentially affect human and animal health via their toxic impact. Of particular concern are a diverse range of toxins produced by cyanobacteria, cyanotoxins, which are hazardous to human, animal and aquatic ecosystem health. Thus, CyanoHABs have significant economic and sociocultural impacts worldwide. The current method of monitoring cyanobacteria involves expensive and unscalable methods such as in-situ water sample lab analysis or deployment of spectroradiometer sensors or monitoring via satellite images which would not provide data for smaller lakes due to poor spatial resolution. Despite the risks posed to environmental, human and animal health, there is no established rapid monitoring program to periodically evaluate the spatial distribution of cyanobacterial blooms.

In the meghdoot/ cyanotracker [6] project at the University of Georgia, we are integrating social media, sensors and satellite imagery data towards an early warning system for monitoring harmful algal bloom (also termed as blue green algae, cyanobacteria, red ride) in lakes and ponds across Georgia, USA. Since these blooms can be observed by naked eye (Figure 1), it is easy for common people to post messages on such blooms on social media. We not only plan to monitor the tag created for the project i.e. #cyanotracker, but also track other existing tags and keywords which refer to the same phenomena viz Algae Bloom, Blue Green Algae, Red Tide, #CyanoHABs, #cyanobacteria, #microsystin thus taking advantage of social sensing.

Challenges:-

The following are a few of the challenges posed while using citizen science for monitoring harmful algal blooms –

First, some of these keywords and hashtags are known and used by experts only (cyanobacteria, microcystin) whereas some are more generic terms (algae blooms, red tide) and widely known thus preferred by majority of people according to their familiarity of the word. The dichotomy of amount of data that can be collected and the number of users tweeting with a generic term vs the noise associated with it is studied in this research.

Second, the posts generated by a non-expert would not be as reliable as he/she may not be sure enough to correctly identify the cyanobacterial bloom event because of lack of expertise but still tweet a positive incident by reporting it. Thus the collected data would be noisy and less trustworthy. Inherently, an experts post can be considered more trustworthy and relevant although we cannot quantify it through experiments. One of the whitepapers published by San Antonio based market research firm Peer Analytics (Kelly [7]) which studied Twitter posts show that in 40% of the tweets are pointless babbles and just 4% of the tweets are actually related to some kind of news. This makes the citizen science through non-project-aware citizen further more challenging.

Third, machine learning algorithms are popularly used for classification. Once a model is built using the training data, it could be used to classify the tweet into different labels. But our research highlights that the social media data has high flux leading to concept drift. We show that model once and classify ever would eventually fail to provide desired classification measure and hence would not work as social media data evolves with time needing one to continuously monitor the variations in these evolving data. Any incremental or periodic retraining machine learning algorithm cannot completely take out the human involvement. With our experimental results, we could say that rather than manually labelling 100% of the new tweet every time for retraining, one effective approach would be to label just a small percentage of tweets from each month and retrain still achieving the desired classification measures.

III. EMPIRICAL STUDY

Our empirical study has three main goals. A. Performing a keywords characterization to show popularity and quality of the various keywords. B. Performing user characterization to study the contribution of various category of people and C. Studying the effectiveness of machine learning algorithms on classification of these keywords and contouring the concept drift phenomena.

Data Collection

We used twitter posts (Tweets) as our social media platform to study the challenges as described in the paper. Using the Python library Tweepy [8], we extracted the tweets for four hashtags: Red Tide, Cyanobacteria, Algae bloom and Blue Green Algae. For experiments we collected the data for a period of seven months starting from September 2014 to March 2015 for the keywords Red Tide, Cyanobacteria and Algae Bloom.

Hashtags				
Month	Red Tide	Cyanobacteria	Algae Bloom	Blue Green Algae
September	3891	210	319	—
October	2274	330	148	—
November	3060	311	262	—
December	1894	91	164	—
January	1676	401	255	—
February	1027	501	263	487
March	2463	326	216	439
April	—	—	—	297
Total	16285	2170	1627	1223

TABLE I: Total Number of Tweets obtained for each hastag every month.

However, for the keyword Blue Green Algae, the data extraction was started late from February 2015. Table I shows the number of tweets obtained for each hashtag. The tweets that were extracted had a lot of noise, particularly for keywords that are used for multiple events. For instance, the keyword Red Tide is also associated with a musical band, US political elections etc. Such tweets are being referred to as noise as they do not refer to the algal bloom phenomenon growing on water bodies. So in order to classify the tweets as relevant or irrelevant to algal bloom, we asked a group of human evaluators to label the tweets as relevant or irrelevant. Figure 2 shows the number of relevant and irrelevant tweets for each hashtag.

A. Keyword Characterization:

Investigation: Our study demonstrates the impact of technicality of a term and its usage in Social Media - Twitter. The analysis includes the statistical study of tweets, users from both the relevant and irrelevant category and also studies the profile description of the relevant users provided on their twitter profile.

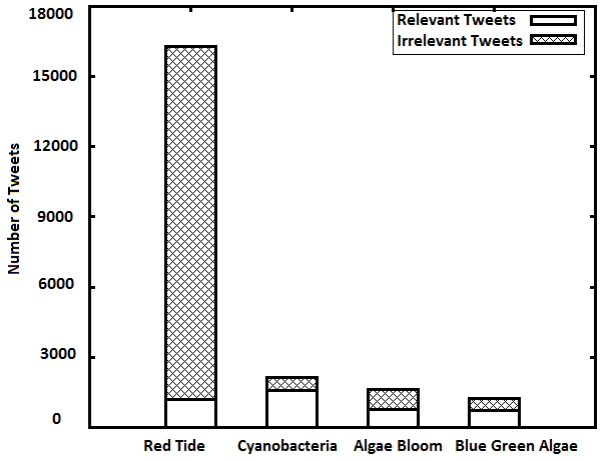

Fig. 2: Total Number of Relevant and Irrelevant Tweets

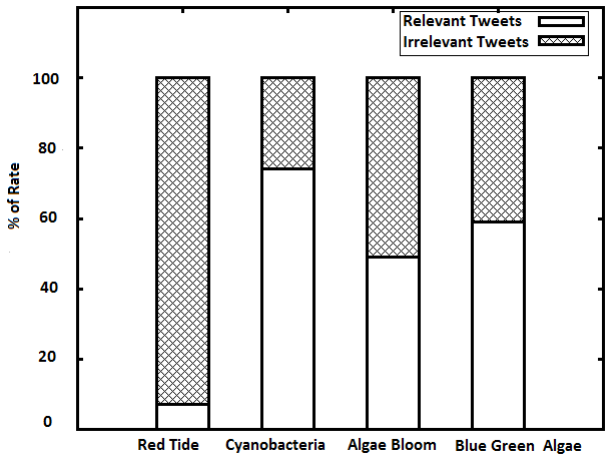

Fig. 3: Total % of Relevant and Irrelevant Tweets

Tweet analysis: As stated previously, tweets for Blue Green Algae were collected for a duration of 3 months and the other hashtags were collected for a duration of 7 months. Hence for the analysis, we have calculated the rate of relevant tweets by dividing the total relevant tweet count by total tweet count and similarly we calculated the rate of irrelevant tweet by dividing the total irrelevant tweet count by total tweet count. Figure 3 shows the percentage divide of relevant and irrelevant tweets rates for these hashtags. From Figure 2 and Figure 3, we can see that though Red Tide being generic term and known to larger crowd produces maximum number of tweets but it has a lot of irrelevant tweets corresponding the noise. The hashtag Cyanobacteria and Blue Green Algae which correspond to more technical term has lesser number of tweets when compared to Red Tide but the relevant rate percentage is comparatively higher which means that the noise decreases as we shift from monitoring the generic term to more technical term.

	Unique Users	Unique Users Rate	Reliable Users	Reliable Users %	Reliable Users Rate	Active Users	Active Users Rate	Patron Users %
Red Tide	10556	1508	958	9.05	137	126	18	13
Cyanobacteria	1295	185	914	70.5	131	160	23	17
Algae Bloom	1084	155	626	57.7	89	96	14	15
Blue Green Algae	934	311	532	56.91	177	94	31	18

TABLE II: Unique, Reliable, Active and Patron Users for each hastag.

B. User Characterization:

In Table II, we present the total number of Unique, Reliable and Active Users for each of hashtag. Unique Users is the combination of users posting relevant and irrelevant tweet, Reliable Users is defined as the users who posted only relevant tweet and Active Users is defined as users who have posted a minimum of two tweets for that hashtag. We show Rates column to tackle the problem with uneven duration of data collection.

Though Red Tide has more number of Unique Users, the Reliable Users % is very less. This shows that in spite of the term being more popular in the social media, the relevant contributors of information are very less. Whereas for technical terms such as Cyanobacteria and Blue Green Algae, although the total number of users is less as compared to Red Tide, they have a decent percentage of Reliable Users. Also the Reliable Users % is maximum for Cyanobacteria from which it can be inferred that, the more technical term, the more relevant or trustworthy information.

From this we could conclude that for technical terms though the total number of contributors are less, majority of them are relevant information providers who could be experts in this area of study. The information coming from these users will have high level of trustworthiness because of their expertise. Whereas for generic terms, though the total number of contributors are more, common people the level of the number of relevant information providers are less. And since most of the contributors are trustworthiness would be comparatively less since they may or may not be the expert in this field.

Active Users: Table II we show the Active Users % for each hashtag. The users who have posted tweet atleast twice for the same hashtag were considered to be active users. Again, for the uneven duration of data collection, we determined the Active Users Rate. We divide Reliable Users Rate by Active Users Rate to determine which hashtag could yield maximum information about algal bloom which we call as Patron User %. We see that the Blue Green Algae has the maximum number of Patron Users % whereas Red Tide has the least number of Patron Users % among the four hashtags. From this we can infer that for mildly technical terms, the Patron Users are more as compared to generic terms. Since a very technical term like Cyanobacteria is known to lesser crowd, the contributors for these terms are less and the same set of users often post tweets using these highly technical terms.

Profession of Users: In order to determine the level of expertise of the users posting the tweets for each hashtag, we parsed the user description for the Reliable Users to find about their profession. Due to privacy concerns and misleading or hypothetical descriptions provided by majority of users, it was difficult to find the profession of these twitter users. So we did a word count of the profile descriptions for each of the hashtags Reliable Users. Table III shows the top 10 frequent words appearing in the description of Reliable Users for each hashtag. We can see that for most technical term in the group i.e. Cyanobacteria, the top words are PhD, scientist, professor, university, student, research etc. indicating that such term are used in social media by researchers, scientists or students dealing with the environment. The top words for Blue Green Algae and Algae Bloom indicates that it might be more commonly used by media reporters, blog writers or people concerned about environment or community. Whereas for a very generic term Red Tide we see that the users are common people or users of twitter and have words like love, music, happy, people etc. as the frequent words in their description.

HASTAG	TOP-10 POPULAR WORDS
Red Tide	News, love, life, follow, people, just, music, one, like, world
Cyanobacteria	News, water, science, university, resesarch, phd, life, scientist, sudent, twitter
Algae Bloom	News, science, water, community, media, life, latest, tweets, follow, environmental
Blue Green Algae	News, water, local, media, science, life, things, health, social, #4h2o

TABLE III: Top-10 Popular words.

C. Effectiveness of Machine Learning:

Social media data streams have a lot of flux. For example, the term landslide which is a geological phenomenon can also be used for landslide victory in politics or landslide board game. Another aspect of Twitter being the noisy data varies as per the popularity of the hashtag.

Therefore while using social media for environmental monitoring, it becomes necessary to filter out the irrelevant tweets from the relevant ones for better results. For small amount of data we could make use of human evaluators to classify the tweets as relevant or irrelevant. But in real time streaming environment where huge amount of data comes at high volume, it is not feasible to have human evaluators due to time and cost ineffectiveness. In such a situation, supervised machine learning classification algorithms can be used to achieve a desired F-Measure for unseen data.

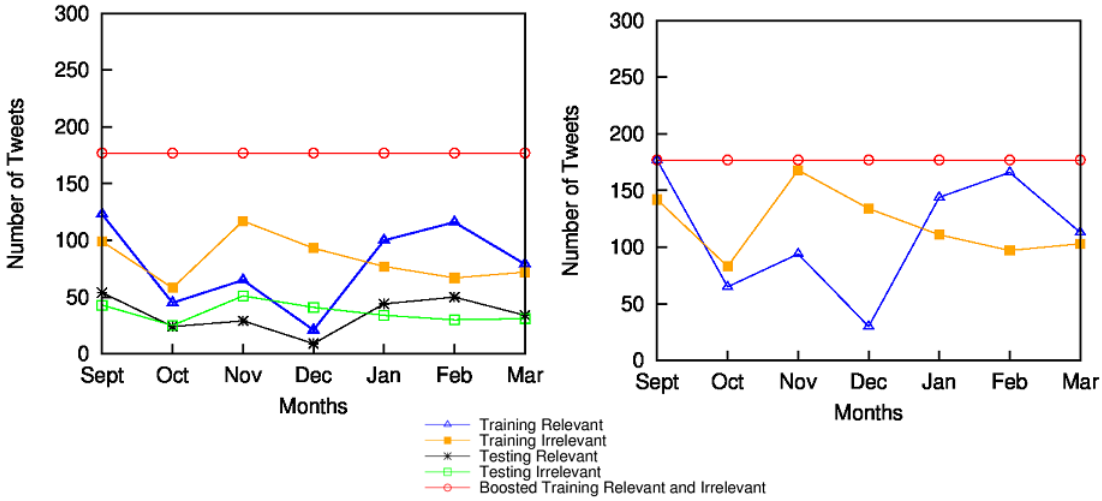

Fig. 4: Testing and Training Dataset split for Algae Bloom for same and different months

Machine Learning Algorithms such as Nave Bayes, Decision Tree, K Nearest Neighbors, Neural networks, Random Forest etc. are some of the supervised classification algorithms that generate a model on the labelled data and using this generated model it classifies unlabeled data. Such a model can also be referred to as train once classify ever model where the one time generated training model is used to assign labels to all future unseen dataset. Since we deal with highly evolving data streams in Twitter, train once and classify ever model would see a decreasing trend as the time gap between the training and testing dataset increases.

In the following, we have described the experimental setup required to analyze the performance of machine learning algorithms on evolving Twitter data streams for three hashtags namely Red Tide, Cyanobacteria and Algae Bloom.

C.1 Experimental Setup

For performing the machine learning classifications, we divided the tweets collected for the three hashtags Red Tide, Cyanobacteria and Algae Bloom according to the months in which they were posted. All the retweet were removed from the dataset. Figure 4 shows the number of tweets for both same month and different month experiments for the hashtag Algae Bloom. The labels were assigned by a group of human evaluators who were instructed to distinguish between the relevant and irrelevant tweets. We have analyzed the performance of machine learning algorithms by increasing the monthly gap between the training and testing datasets.

We have used the traditional train once and classify ever way of classification and studied its performance on twitter data. For instance, when September data is used as the training dataset, the training model generated for September is used for testing the dataset from October then November and so on till March. The training months were also gradually increased from September to February. In this way, the monthly gap between the training and testing dataset is gradually increased and the performance of the Machine Learning algorithms was analyzed. For the monthly analysis, the experiments consist of two sub types namely, same month experiments and different month experiments. Same month experiment had training and testing datasets from the same month. In this case, the 70-30 split mechanism for training and testing dataset was used, where 70% of the relevant monthly tweets go to training and 30% go to testing dataset. Similarly for irrelevant tweets, 70-30 split was carried out.

For different month experiments, tweets from the entire month were used as training and testing dataset, for example for the Sept-Mar experiments, all the training dataset were taken from September and the testing dataset were taken from March. As the number of tweets posted every month varies, we boosted the training and testing dataset to make the number of relevant and irrelevant tweets equal. For this, we determined the maximum number of relevant and irrelevant tweets for each hashtag and then boosted all the relevant and irrelevant training data to match this maximum number. For example, the maximum number of irrelevant tweets is 142 that occurs in the month of September and maximum number of relevant tweet is 177 which also occurs in the same month of September. So we boosted all of the relevant and irrelevant data in training dataset to 177.

C.2 Results and Graphs

We employed Nave Bayes and Ensemble learning algorithm to classify the evolving data. In order to plot the graphs, we determined the average of the F-Measure, Correctly Classified % and Recall obtained from the above two machine learning algorithms. Figure 5 show the F-Measure, Correctly Classified % and Recall for the three hashtags Vs the difference in months. We see that for Red Tide and Algae Bloom, the average measures exhibit almost decreasing or inconsistent trend as the time gap increases between the trained dataset model and the testing dataset. However, for Cyanobacteria the average measure has a mild increase even though the time gap increases between the trained dataset model and the tested dataset. From this we can infer that, when we use dataset from the past to classify the present data, the accuracy or performance of the machine learning is hampered for generic terms which are associated with multiple entities and are

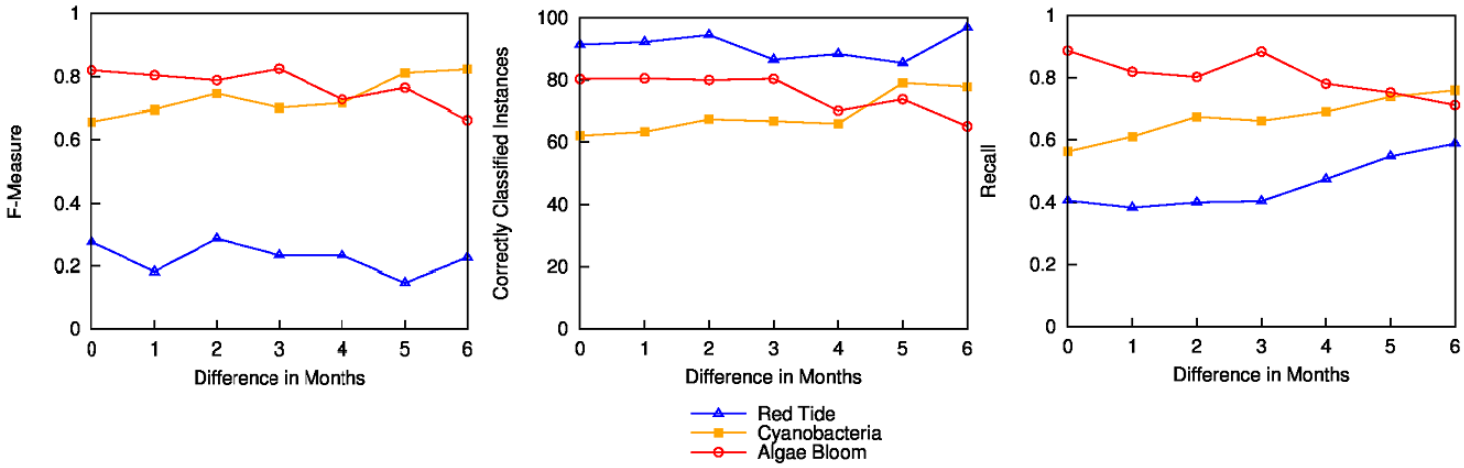

Fig. 5: Average F-Measure, Correctly Classified and Recall vs Difference in months for Red Tide Cyanobacteria and Algae Bloom

known to majority of the population.

One other reason for this is that an event that has occurred in a particular month might not have occurred after few months and there could be totally new discussion topic few months later. So there might be unseen data in the testing dataset which were never part of the training dataset. With lesser monthly gap between the training and testing dataset, more similar events coexist due to which machine learning algorithms gave good classification results. But for the hashtag cyanobacteria which is more technical term and contained lesser noise, the average performance of machine learning algorithms have an increasing trend even though the monthly gap between the training and testing dataset increased.

The trend for these hashtags can be justified by seeing the Jaccards coefficient between different months which would basically give the overlapping words between the dataset of these individual monthly gaps. From the entire dataset, we first removed the stop words and determined unique words. We used the same combination of training and testing dataset months as used in the machine learning experiments and calculated the Jaccards coefficient as follows:

$$J(M1, M2) = \frac{J(M1 \cap M2)}{J(M1 \cup M2)}$$

,where M1 and M2 are months

$M1 \cap M2$ is the overlap of words between the two months M1 and M2, and $M1 \cup M2$ is the sum of unique words in M1 and M2. From Figure 6, we can see that for Red Tide, there is a steep drop in the Jaccard coefficient as the monthly gap increases. This shows that there is a decrease in similarity of events or tweets for a generic term is more as the monthly gap increases which might be one of the reason behind the decrease in performance of the machine learning algorithms.

D. Retraining:

As seen in section C.2, the performance of the machine learning algorithms deteriorates as the monthly gap between

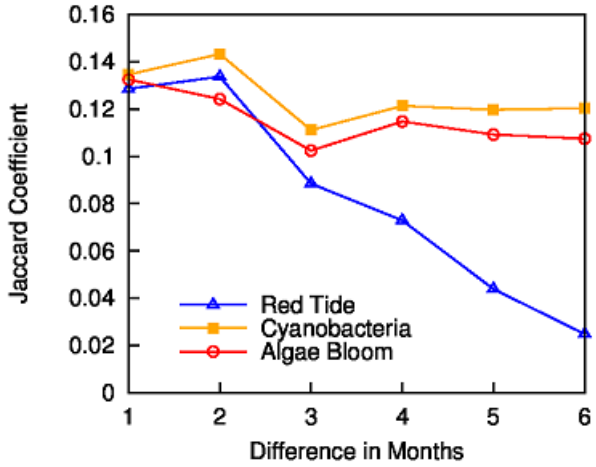

Fig. 6: Overlapping of words as time evolves.

the training and testing dataset increases. This shows that the train once and classify ever model is not suitable for the data from online social media sites. One of the possible solutions to this problem would be to use either incremental or batched retraining methods wherein a newer model is built as soon as unseen data is collected. In this way, the training dataset will have latest tweets along with the tweets from past. Such retrained training helps in improving the performance of the machine learning algorithms. But, not only such methods are time and cost inefficient but they also need high level of intuition and domain knowledge as its unclear whether a new data point should become part of model, if yes then should it replace any other old data that was part of the model.

One of the possible ways would be to use only a percentage of new data from each of the months for retraining. Figure 7 shows the increase in performance of machine learning algorithms when only part of each month data was used to build the retraining model. We have plotted graphs to compare

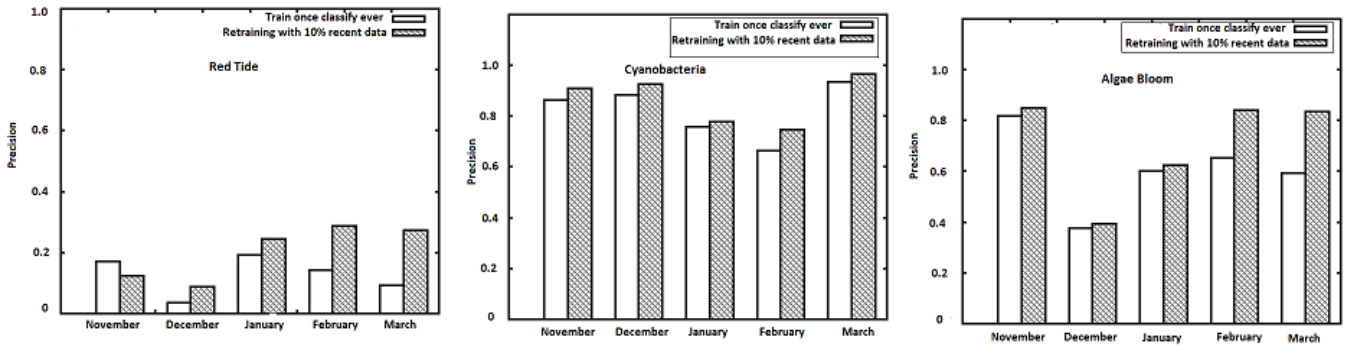

Fig. 7: Increase in precision after using Retraining Model

the performance of retraining model and train once classify ever model. The horizontal axis gives the months that were used for testing. For each testing month in the graph, the left bar denotes precision when the month of only September was used for training and the right bar denotes retraining model where the training dataset was made from portion of data for all months from September to a month prior to the testing month.

For instance, for the testing month January there are two models. One generated by just using September dataset and another generated by using equal parts of the data from following months September, October, November and December. For uniform comparison between the two training models, the number of tweets in the training dataset was kept the same in both the models. The experiments results are plotted in Figures 7 from which we can clearly see an increase in the performance of the machine learning algorithms when the retrained model is used in the training dataset. This shows that for evolving data source like Twitter where the data contents keep varying, machine learning performance can be improved by using such a retraining model. In this way, we will not only get latest tweets in the retraining model but will also reduce human effort by having to label only a fraction of the new incoming tweets rather than labelling all newly extracted tweets.

IV. CONCLUSION

In this paper we did an empirical study on effectiveness of using online social media for slowly evolving environmental phenomenon. We show that phenomenon can be referred with several hashtags in online social media by taking the example of harmful algal blooms. Some of these hashtags are highly technical and known to scientists and domain experts whereas some hashtags are quite generic in nature and known to majority of people thus they bring in high noise in the social posts. We have studied the hashtag characteristic and characteristic of the users of these hashtags. Later in the study, we show the concept drift in the social media posts and the degradation in performance of these machine learning algorithms with time and illustrate that the performance can be increased by retraining with just a small percentage of data from each month.

We say that the posts using highly technical hashtags are more trustworthy as the study showed they were preferred by scholars and the commonly known thought is that trustworthiness and expertise go hand in hand. But this may or may not be true as the scientist may or may not be a domain expert on cyanobacteria. So his/her expertise in this field is questionable. Also, we believe that if the tweets were extracted for a longer duration, the decreasing trend in performance of the machine learning algorithms would have been clearly visible. Though these findings are related to cyanobacteria, we believe that these challenges are not specific but wider in scope which is applicable to many scientific domains.

V. RELATED WORK

Many researchers ([1] [2] [3] [4] have focused on social media to detect events such as earthquake, swine flu, Ebola, landslide etc. which receive considerable media attention. The social media data for such events comes in burst and then eventually fades away and majority of these kind of events can be tracked using a very standard hashtag which are known to lay citizens. However, certain type of events are associated with multiple hashtags and are slowly evolving which doesnt receive much of social media attention. Our work tries to study the feasibility of using social media for detecting such slowly evolving events which doesnt receive media and large public attention. Our work also compares the concept drift of these hashtags over a period of time and the degradation in quality of relevant post for these different hashtags which we believe hasnt been studied by any of the work on event detection in social media.

ACKNOWLEDGMENT

This research has been partially funded by the National Science Foundation under Grant Number CCF-1442672. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the NSF.

REFERENCES

[1] A. Musaev, . D. Wang, S. Shridhar and C. Pu , "Toward a Real-time Service for Landslide Detection: Augmented Explicit Semantic," in IEEE International Conference on Web Services, 2015.

[2] M. Guy, . P. Earle, . C. Ostrum, K. Gruchalla and S. Horvath, "Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies," Advances in intelligent data analysis, vol. IX, pp. 42-53, 2010.

[3] S. A. P. P. Signorini A, ”The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic.,” PLoS One, 2011.

[4] C. W. Schmidt, ”Trending now: using social media to predict and track disease outbreaks,” Environmental Health Perspectives, vol. 120, no. 1, pp. 30-33, 2012.

[5] T. Sakaki, M. Okazaki and Y. Matsuo, ”Earthquake Shakes Twitter Users:Real-time Event Detection by Social Sensors,” in WWW, 2010.

[6] J. Ritterman, M. Osborne and E. Klein, ”Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic,” in 1st international workshop on mining social media, 2009.

[7] R. Power, B. Robinson, J. Colton and M. Cameron, ”A Case Study for Monitoring Fires with Twitter,” in Proceedings of the ISCRAM , 2015.

[8] R. Kelly, 12 August 2009. [Online]. Available: http://web.archive.org/web/20110715062407/www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf.

[9] C. D. Corley, . A. R. Mikler, K. P. Singh and D. J. Cook, ”Monitoring Influenza Trends through Mining Social,” in BIOCOMP, 2009.

[10] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu and B. Liu, ”Predicting flu trends using twitter data.,” in IEEE INFOCOM WKSHPS, 2011.

[11] ”Weka 3 - Data Mining with Open Source Machine Learning Software in Java,” [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/.

[12] ”Tweepy,” [Online]. Available: http://www.tweepy.org/.

[13] ”CyanoTRACKER - The University Georgia,” The University of Georgia, [Online]. Available: http://www.cyanotracker.uga.edu/.

[14] M. Guy, . P. Earle, C. Ostrum, . K. Gruchalla and S. Horvath, ”Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies,” dvances in intelligent data analysis, vol. IX, 2010.