Virtual Power Outage Detection Using Social Sensors

Konstantin Bauman
Stern School of Business
New York University
kbauman@stern.nyu.edu

Alexander Tuzhilin
Stern School of Business
New York University
atuzhili@stern.nyu.edu

Ryan Zaczynski
Stern School of Business
New York University
rjz226@stern.nyu.edu

***Abstract*—In this paper we describe a novel approach to detecting power outages that utilizes social media platform users as "social sensors" for virtual detection of power outages. We present the underlying methodology based on analyzing Twitter and other social media data that detects bursts in tweets related to the power outages. The proposed methodology was implemented and deployed by a major company in the area of enterprise solutions for social media aggregation for the electrical utility industry as a part of their comprehensive social engagement platform. It was also field tested on the Twitter users in an industrial setting and performed well during these tests.**

I. Introduction

Power outages constitute a serious problem around the world that disrupts our lives in the most unexpected ways. In the United States alone, there were $3,634$ power outages reported in 2014, affecting an estimated 14.2 million people [1], and over the period of 2008-2014 the US has averaged $2,987$ outages affecting 21.6 million people per year [1]. Power outage related losses for US businesses are estimated as being in excess of $\$150$ billion annually [1].

To address this problem, there have been extensive resources dedicated to detecting and reporting power outages with new technologies, such as smart sensors, meters and distribution devices [2]. Utilities' Outage Management Systems (OMS) vary in their composition of outage detection technologies. These systems typically include both traditional and, so-called, "smart" grid elements as means of power outage detection in the utilities' coverage regions. Unfortunately, the sensor and smart grid technologies, although useful in detecting power outages, are extremely costly when deployed at scale across the country. Full implementation of smart technologies is not expected until approximately 2030, with total costs estimated at a staggering $\$338$ to $\$476$ billion [1]. To address this important problem in the near term and with respect to budgetary constraints, it is necessary to develop alternative approaches to power outage detection, including usage of a different class of "sensors."

The ubiquity of smart phones and social networks has given rise to an entirely new class of sensor: the human "social sensor" [3]. Indeed, any individual with a networked device and a social media account has the potential to become a *"social sensor node"*. Social sensor nodes are capable of a wide range of functions, including producing unsolicited descriptive data about spontaneous events in real-time, acting as two-way channels for communication and information feedback, and functioning either independently or as a collective network. Social networks, such as Facebook and Twitter, consist of tens of millions of these would-be social sensor nodes. The Electrical Utility industry is uniquely positioned to immediately benefit from incorporating this new class of sensors into their existing Outage Management Systems. The data produced by social sensor nodes can be analyzed, modeled, and used to construct a virtual outage detection network for power outage events. A virtual outage detection network could function independently of, and in parallel with, utilities' existing Outage Management Systems.

The immense value of decentralized, crowd-sourced social data in a crisis event was underscored in the wake of 2012's Hurricane Sandy. The disaster, which crippled much of the Northeast's electrical power grid for several weeks in 2012, prompted the US Department of Homeland Security to declare social media as one of the "critical components of emergency preparedness, response, and recovery" and, furthermore, the agency noted that Twitter's use for reporting "issues, danger, and power outages" was regarded by many as a "lifeline" [4]. Other government agencies, such as the US Department of Energy (DoE) and Federal Emergency Management Agency (FEMA) have participated in the initiatives led by The White House to improve and standardize crisis management information [5]. These collaborations culminated in an event titled "The White House Innovation for Disaster Response and Recovery Initiative Demo Day" in July of 2014 which brought together a diverse group of "technologists, entrepreneurs, and members of the disaster response community to showcase tools that will make a tangible impact in the lives of survivors of large-scale emergencies" [6]. One of the authors of this paper was invited to attend this event alongside technologists from enterprises such as Google and Microsoft, emergency managers from states across the US, and representatives from eight of the US Agencies directly involved with crisis management. The co-author's experience served as the inspiration for this research. Enterprises, including General Electric, have also recognized the potential value of social media in the area of power outage detection [7]. Electrical utilities have taken notice as well, and they have dramatically increased their interest and investments in social media related technologies [2].

Twitter is one type of a social network that can particularly be useful in power outage detection. It has approximately 63 million active users in the United States alone [8], and recent empirical research has identified a strong correlation of geo-tagged Tweets across the world being in close proximity to the presence of electricity [9]. As such, the social sensor nodes needed to construct a virtual outage detection network are potentially already in place. Furthermore, our research has identified that a certain percentage of these Tweeting social sensor nodes are already actively reporting on outage events when they occur. Thus, having social sensor nodes already in place and actively reporting on electrical outage events, our research investigates how the information supplied by these Twitter users can serve as the foundation upon which this virtual outage detection network may be built.

The value of a virtual outage detection network to an electrical utility is twofold. First, it currently serves as a method of choice to improve Electrical Utilities' existing Outage Management Systems detection capabilities prior to full implementation of smart grid elements that are still years away from their full deployment that is projected to be done by utilities by 2030. Since the social sensor node information in the form of Tweets and other social media posts are free, the costs of implementing a virtual outage detection network constitutes a tiny fraction of the projected $338 to $476 billion needed for the full smart grid implementation, and it could be achieved much earlier than 2030. Furthermore, once smart grid elements are in place, a virtual outage detection network will continue to *augment* the effectiveness of the smart grid network elements by extending the reach of the Outage Management Systems' detection capabilities beyond the confines of the electrical grid to wherever social media users are located.

Finally, virtual outage detection network can complement or maybe even replace the traditional phone-based reporting methods when customers call electrical utilities to inform about power outages. In particular, the younger populations, such as teenagers, extensively use Twitter, Instagram, Snapchat and other social-media platforms as their primary communication tools and, therefore, virtual outage detections are the natural method of choice for them. Also, trying to reach utilities via phone, especially in cases of extensive power outages, can be a daunting and a time-consuming task. In addition, people may not always have access to the phones or have willingness to call, and the phones may not be properly functioning in cases of serious emergencies.

This research investigates how this new data from social sensor nodes can be captured, analyzed, and delivered as validated, actionable information for use by the Electrical Utility industry's Outage Management Systems and used to build a virtual outage detection network. The proposed method presented in this paper uses key textual descriptions of power outages, filters the Tweets containing these concepts, builds a predictive model that identifies those Tweets referring to real power outages and detects bursts among these identified Tweets.

This method was implemented by a major company in the area of enterprise solutions for social media aggregation for the electrical utility industry. It is a part of that company's comprehensive social engagement platform under the leadership of one of the authors of this paper. It was also field tested on the Twitter users in real industrial settings. These test results show that, from all the power outages that our system detected, 93.7% and 97.6% of them referred to the real outages across the two validation mechanisms reported in the paper. Furthermore, our system was able to detect 74.1% and 69.8% of all the power outages mentioned in the tweets across these two validation mechanisms. Although this number is relatively low, it is actually a good detection result because many of the missed power outages were described by only one or two tweets, thus potentially producing inaccurate results.

The described system constitutes the very first power outage detection platform based on social media networks (such as Twitter) developed in the Electrical Utility industry. The use of social media information for power outage detection has created substantial excitement in the Electrical Utility industry. For example, this technology has been highlighted by a major electrical utility industry publication as playing an integral role in the "next generation of outage management" [2], and it has been regarded as a "new fascinating development" by a principle at one of the worldwide leading companies in this area[1].

Prior work in this area focused on using Twitter and other social media platforms for detecting earthquakes and other types of emergency events [3], [10], [11]. In particular, [3], developed an earthquake alert and report system that uses Twitter data for detecting and reporting earthquakes immediately after their occurrence. Similarly, [12], [13] presents the SMART-C framework for emergency detection and alert dissemination. Unfortunately, this framework focuses mainly on the architectural and privacy issues and does not cover implementation and deployment aspects in [12], [13].

In contrast to this prior work, the contributions of this project lie in focusing on the power outage application, in developing a novel power outage detection method, and in implementing it as a part of a comprehensive social engagement platform.

II. Overview of our Methodology

The proposed methodology of power outage detection is based on the idea of (1) identifying keywords for a particular application, (2) detecting tweets containing these keywords and (3) identifying bursts in the stream of these tweets. More specifically the overall approach is presented in Figure 1 and consists of the following steps:

1) (a) Specify the set of core key concepts K pertaining to power outages and (b) compute its closure C.

[1]Since we did not have a chance to obtain the explicit permission of that person to quote him in this paper, unfortunately, we cannot reveal his identity here.

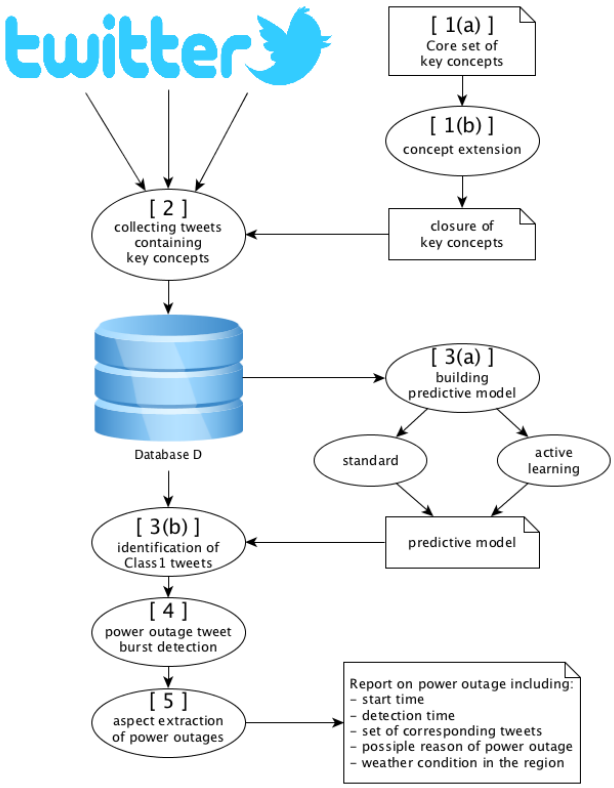

Fig. 1. Methodology of Power Outage Detection System.

2) Collect the tweets from the Twitter stream containing at least one key concept from set C and store them in database D.
3) (a) Build a predictive model that identifies if tweet x from D posted at time t refers to a power outage that occurred around time t in the region where tweet x originated (as determined by the tweet location);
(b) classify the tweets from database D into $Class\ 1$ specifying those tweets that include their GPS data or home location and that were posted by the individuals who witnessed real power outages and immediately tweeted about them, and all other $Class\ 0$ tweets.
4) Identify the bursts of $Class\ 1$ tweets in the stream generated in Step 3 (that refer to currently occurring power outages).
5) Extract the aspects of the power outages for each burst of $Class\ 1$ tweets, these aspects pertaining to the possible reasons and the weather conditions of the outage.

These five steps are further explained in detail in the rest of this section.

A. Building a Set of Key Concepts

We start the process of building the key concepts pertaining to power outages by first identifying the set of *core key concepts* K, such as "power outage" and "no power." Set K of these core concepts is specified "by hand" (i.e., manually by the expert on power outages). After that, we compute *closure* C of this set of core concepts K by finding other concepts that are "similar" to set K. More specifically, we proceed as follows. For each keyword in the concept, we find it synonyms using an online dictionary and a WordNet [14]. Then we construct the list of combinations of synonyms of individual words using all possible combinations of them. As a result, we obtain complete closure of all the possible synonyms of a key concept, although some of them may not make much sense. For example, if we use "force" as the synonym of "power" and "failure" as a synonym of "outage", then "force failure" is not really a synonym of "power outage." Nevertheless, this is not a problem in our case because the value of such irrelevant key concepts will be automatically reduced in the subsequent steps of our outage identification process (described in Section II-C). Therefore, we do not filter such meaningless concepts from the set C now because our main objective in this step is to achieve maximal coverage of all the possible concepts that may point in some way to the power outage.

In our study, we specified the following set of core key concepts $K = \{$"$power\ outage$," "$no\ power$," "$electric\ failure$"$\}$, and generated 110 key concepts as a result of computing its closure C. Some examples of the key concepts from this set C include "power failure," "electricity outages," "electrical blackout," "electrics out," "no energy."

B. Collecting Tweets

Given closure C of the set of power outage key concepts described to Section II-A, we use Twitter API to extract from the Twitter stream all the tweets containing at least one of the concepts from set C. We extract such Tweets in real time and store them in database D of all the tweets related to power outages. In our study, we have obtained the Twitter data from $08/18/2014$ to $1/25/2015$ using set C of the key concepts. As a result, we have collected $117,490$ tweets related to power outages over the period of 5 months.

Furthermore, we have also specified 281 regions of the United States corresponding to the regional power companies, each region being served by that power company and constituting the unit of analysis of power outages based on the tweets posted in that region. Therefore, this partitioning of the USA into regions allows us to identify power outages for each power utility.

C. Predictive Model

Clearly, not all the tweets in database D identify real power outages. For example, the following tweet "MDU Plans Power Outages details here–>http://t.co/abcd1234" is referring to a power outage that can possibly occur in the future and not to a currently occurring outage. Therefore, it is necessary to differentiate between the useful tweets and all other possible (and irrelevant) tweets. Furthermore, several types of tweets describe different aspects of a power outage but do not refer to the real power outages happening at that moment, such as tweets referring to past power outages, the tweets from the news agencies referring to power outages that happened in other regions, or the tweets that prepare population to

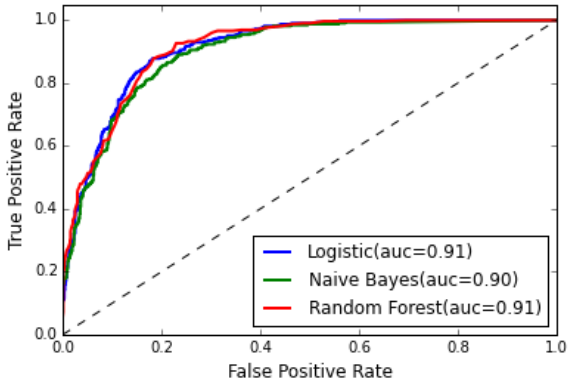

Fig. 2. Receiver Operating Characteristic curves

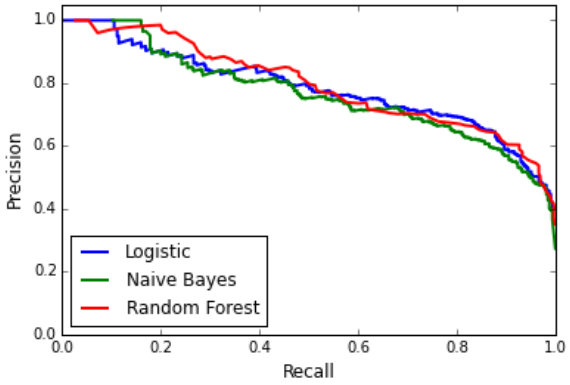

Fig. 3. Precision-Recall curves

possible power outages that can happen in the region due to the approaching storms or other similar reasons. In this project we focused on the identification of the tweets that have their GPS or home location and that were posted by the individuals who witnessed real power outages and immediately tweeted about them ($Class$ 1 tweets). We define all other types of tweets as $Class$ 0.

To identify such tweets, we construct a predictive model that determines if tweet x from database D posted at time t belongs to $Class$ 1. Such model helps us identify all the $Class$ 1 tweets in database D and classify the newly posted tweets as being of $Class$ 1 or $Class$ 0.

In this project we build a predictive model using the following two methods:

1) *The standard method,* in which we labeled a randomly selected subset of tweets from D by hand, and constructed a supervised predictive model identifying $Class$ 1 tweets using some of the standard machine learning classification algorithms.
2) *The active learning method,* in which we constructed a machine learning classifier having initially very few labeled examples. Then we iteratively identified new examples to be labeled in order to improve the prediction quality of the model generated during the previous iteration.

Standard Method. To follow the standard approach, we labeled a random sample of $4,000$ tweets and divided this set in the $80/20$ proportion for the training and testing purpose.

We built our classification model using the following list of features:

1) Features based on text of the tweet
   - single words (True/False)
   - n-grams of words (True/False)
   - number of words in the tweet
   - number of symbols in the tweet
   - number of capital letters in the tweet
   - if the tweet contains a URL link (True/False)
   - if the tweet is a re-tweet (True/False)
   - sentiment of the tweet
   - if the tweet contains numbers (True/False)
   - number of verbs in the past tense
2) Features based on user name:
   - length of the user name
   - number of capital letters in the user name
   - if the user name contains certain special words, such as "news", "police", "power", etc.

Among all these features, "single word" is a Boolean feature specifying whether a particular word from the universe of all the words in database D is present in the tweet. Also, the n-gram feature specifies the same concept as the "single word" but for the n-grams. We used n-grams with $n = 2$ and $n = 3$ in our study. The sentiment feature of the tweet identifies its sentiment, as produced by the Python library TextBlob[2].

We compared several classical machine learning methods, including Naive Bayes (NB), Logistic Regression (LR) and Random Forests (RF). The results of these comparisons in terms of ROC and Precision-Recall curves are reported in Figures 2 and 3 respectively. We built these models and did the analysis using Python scikit-learn library [15].

As you can see in Figures 2 and 3, the performance results for all the three methods are comparable. We selected the LR method among the three alternatives since it is slightly better than the other approaches.

Furthermore, as our NB and RF models showed, the most powerful features were the features specifying the length of the tweet, such as the number of words and symbols in the tweet. This means that the true power outages happening in real time are usually specified by significantly shorter tweets (that are of $Class$ 1) than other types of tweets related to power outages (that are of $Class$ 0). Also, tweets of $Class$ 1 rarely constitute re-tweets and unlikely contain links.

Finally, the $F1 - measure$ performance of our Logistic Regression model on the test set is $F1_{Class1} = 0.74$ and $F1_{Class0} = 0.88$.

Active Learning Method. In the active learning method, we started with labeling a small set of n random tweets from our

[2]textblob.readthedocs.org

dataset D and built NB model using the same set of features as for the standard method but based on only these n tweets. In our study we started with $n = 10$. Further, at each iteration, we predicted labels for all the tweets in D and selected a new set of n tweets to be labeled next. We selected this new set according to the method described in [16], in particular we selected for labeling the set of unlabeled examples that generated the lowest expected error on all other examples. We did this process iteratively until we reached saturation. As we ran this iterative learning process, we checked the prediction quality of the resulting NB model on each iteration on the test set of 800 tweets. The authors in [16] maintain that their active learning procedure can reach sufficiently good results within the first 30 or 40 labeled items. Our study confirmed the same result since we reached convergence for our model using only 30 to 40 labels. In particular, the performance on the test set is $F1_{Class1} = 0.697$ for the NB model built on the first 30 labeled items and $F1_{Class1} = 0.712$ for the first 40 labeled items.

Although we could have used either type of classification model (the standard or the active one) since both of them perform well, we used the standard LR model in our study. This standard LR model classified all the tweets from database D and identified $34,436$ of them as $Class$ 1 tweets (we denote this set as W). Note that the tweets from database D containing the "irrelevant key concepts" described in Section II-A will be eliminated in this step, as being not relevant to real power outages according to our predictive model. Next, we use the tweets from set W for the identification of power outages as described in Section II-D.

D. Identification of Power Outages

Having a tweet of $Class$ 1 from set W described in Section II-C does not necessarily mean that there is a true power outage at the time of the tweet post since false positive rates of these tweets can be high. Therefore, a better method of identification of power outages is based on the analysis of the *time sequence* of tweets in set W and identification of *bursts* in that sequence.

Since the burst detection problem has been studied before [17], [18], [19], we decided to apply one of the prior burst detection methods to our problem. More specifically, we used the particular burst detection algorithm developed by Jon Kleinberg [18] where he considered a stream of emails and news articles that arrive continuously over time and detected bursts of discussion of certain topics among them using an infinite state automaton technique, in which bursts appear naturally as state transitions. This method is shown to be efficient and it deals with the underlying noises, thus not requiring usage of sliding windows or human interventions.

In our study we used Kleinberg's method for detecting bursts within the streams of tweets within each region by inspecting time periods between the consecutive tweets of $Class$ 1 and identifying those groups of subsequent tweets having abnormally short time periods between them. We have also extended the method from [18] by examining all the bursts detected by Kleinberg's algorithm and accepting only those of them as "true bursts" that have the number of tweets greater than some threshold value th. The threshold number th is selected by finding the right balance between the precision and recall measures of the model. In our case, we selected the threshold level of $th = 2$ by experimenting with different levels. Finally, we have applied this burst detection method to each of the 281 regions described in Section II-B in order to detect power outages in those particular regions.

The burst detection method described in this section is really an off-line detection method applicable to the historical tweeting data. In contrast, we needed to detect power outages in real time in our study. Therefore, we launch the Kleinberg's algorithm [18] each time we receive a $Class$ 1 tweet $t \in W$ and check if the algorithm has detected a burst at the time when tweet t was posted. Further, once we detect such real-time burst, we keep track of all the $Class$ 1 tweets from W corresponding to this burst until it subsides according to Kleinberg's algorithm. We denote this set of tweets corresponding to a particular i-th burst as S_i. In our study we identified $3,750$ such bursts across 135 regions.

E. Aspect Extraction of Power Outages

In addition to detecting power outages, we also try to identify the following two aspects of these outages: (a) its reason, such as equipment failure or public accident, and (b) the weather condition at the time of the outage. We accomplish this task as follows.

Each type of reason of a power outage is defined by a set of its characteristic keywords. In our study, we have used the standard reasons for power outages, as adopted by a major electrical utility company:

- *Vegetation*, such as a tree falling on a power line
- *Equipment Failure*, such as a problem with a power substation, or a power line pole falling down
- *Public Accident*, such as a car accident involving power equipment
- *Wildlife*, such as a squirrel cutting a power line.

Then we define a set of keywords for each of these four categories of power outages. For example, we defined the following set of keywords for the Vegetation category: "tree," "limb," "branch," "vines," and "trunk." Then our system scans the set of tweets S_i corresponding to a particular burst of $Class$ 1 tweets and identifies for each keyword in each of the four categories the set of tweets in S_i containing that keyword. For example, the keyword "tree" in the Vegetation category may have appeared in tweets t_1, t_5 and t_9, and the keyword "branch" may have appeared in tweets t_2, t_5 and t_8. After that, for each category we combine all the tweets containing at least one of its keywords. For example, category Vegetation will have the union of the tweets corresponding to the keywords "branch" and "tree," i.e., t_1, t_2, t_5, t_8 and t_9.

We also use a very similar method for the weather identification process. In particular, we used the standard weather classification codes adopted by a major electrical utility company, such as: *Rain, Wind, Calm, Snow,* etc. For each of

these weather conditions, we have a set of corresponding keywords, as for the case of the power outage reasons. For example, for the weather condition *Snow* we have identified the following keywords: "snow," "snowfall," "ice," "sleet," "drifts," "melting," etc. Then we proceed in the same way as for the power outage reasons and identify the set of tweets in S_i where these keywords occur.

In conclusion, our system reports not only the occurrence of a power outage in a certain region at a certain time, but it also outputs all of its defining tweets S_i together with the list of power outage reasons and the weather conditions, and the set of tweets corresponding to these reasons and weather conditions. In Section III we evaluate the performance of our power outage detection method described in this section.

III. Results

We evaluate performance of our system based on two types of data. First of all, we compare the tweet bursts detected by our system with the information about real power outages, as observed and recorded by a power company. Secondly, we compare the power outages detected by our method with the outages that were identified *and* tweeted by some of the reliable sources, such as police departments, news agencies, power companies, etc. We describe these two evaluations in Sections III-A and III-B respectively.

A. Utility Power Outage Data

We have obtained information about the observed power outages for one of the major utility companies operating in a large municipal region in the US for the time period from 10/25/2014 to 1/25/2015. For confidentiality reasons, unfortunately, we cannot reveal neither the name of the company nor the geographical region where it operates. Therefore, we refer to this Utility and to the region where it operates as *XYZ* in this paper.

The unit of power outage for the XYZ company is the event defined by the loss of power on a single unit of power supply, called *"feeder"*. A complete power outage is a complex event that is defined by a combination of individual power outages on one or several feeders that occur, roughly, at the same time. Therefore, a complete power outage consists of one or several outage "fragments," each fragment corresponding to a single feeder. Each "feeder outage" event is defined by its starting time, ending time, feederID, outage type, location, cause, category, equipment type, the number of affected customers and some other characteristics.

Figure 4 presents an example of the histogram of tweets posted in the XYZ region on an hourly basis on 1/18/2015. The blue histogram in Figure 4 shows the number of tweets per hour containing at least one of the keywords from set C described in Section II-A. The green line corresponds to the histogram recording the quantities of $Class$ 1 tweets over time. Finally, the red line defines the number of tweets corresponding to the detected bursts of power outage tweets. As you can see from Figure 4, the red histogram shows that

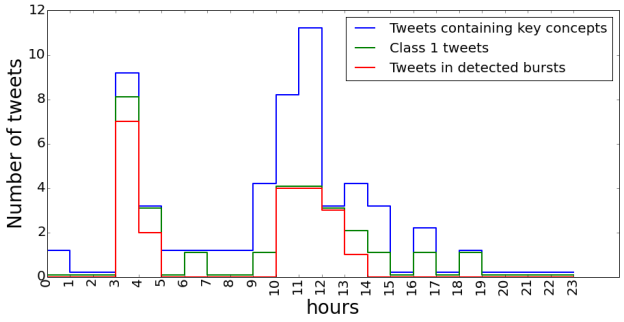

Fig. 4. Numbers of tweets by hours in Utility XYZ region on 1/18/15

our system has detected two power outages in the XYZ region on January 18, 2015.

We measure performance of the power outage detection method presented in Section II using the precision and recall measures as follows.

Precision Calculations. For each observed burst in tweets, we identify if it has a registered power outages on *at least one* feeder that match this burst by time and by location (as defined by the same region). The precision measure is defined as follows in this case:

$$Precision = \frac{\#True\ Positive}{\#True\ Positive + \#False\ Positive} \quad (1)$$

where $\#True\ Positive$ is the number of bursts that matched at least one power outage event, and $\#False\ Positive$ is the number of bursts that did not match any power outage event.

Recall Calculations. For each power outage event observed on a feeder, we identify if it has the corresponding burst of tweets reported at approximately the same time and the same location (region) as the power outage. The recall measure is defined as follows in this case:

$$Recall = \frac{\#True\ Positive}{\#True\ Positive + \#False\ Negatives}. \quad (2)$$

where $\#True\ \ Positive$ is the number of power outage events that matched at least one detected burst, and $\#False\ Negative$ is the number of outage events that did not match any detected power outage discussion burst on Twitter.

Although our system has identified 325 tweets posted in the XYZ region between 10/25/2014 and 1/25/2015 that contain the key concepts, only 163 of them were classified as Class 1 tweets. From these 163 $Class$ 1 tweets, our system has identified 16 bursts of outage discussions using the methods reported in Section II-D.

The performance results are reported in Table I. As you can see, 15 out of the total of 16 power outages identified by our method in the Utility XYZ region between 10/25/2014 to 1/25/2015, have been confirmed by the XYZ company data, thus producing the precision performance of 93.7%. Furthermore, the 16th power outage not confirmed by the XYZ data corresponded to some lights going off on certain floors of a City hotel, as recorded by a few tweets posted by the hotel

Identified PO bursts	16
Detected PO confirmed by utilities	15
Precision	93.7%
PO in validation data	298
PO identified by system	109
PO discussed on Twitter	147
Total Recall	36.5%
Twitter based Recall	74.1%

TABLE I
PERFORMANCE RESULTS BASED ON THE UTILITY POWER OUTAGE (PO) DATA

customers. Since this power outage was so small, it was not detected by the XYZ company.

The recall performance measure is recorded in Table I. Out of 298 of the total "feeder outage" events reported by XYZ between 10/25/2014 to 1/25/2015, 109 of these events corresponded to some bursts in the tweet discussions detected by our system. This result corresponds to the *total recall* performance measures of 36.5%.

Although such levels of recall are usually considered to be "low" in many applications, it is, nevertheless, are "reasonable" performance measure in our case because not all the power outages are usually discussed on Twitter. For example, some power outages occurring in the rural and less "social-media friendly" regions, as well as in the more industrial (vis-a-vis more residential) areas may not be recorded on Twitter. To test this hypothesis, we examined how many of the 298 total power outage events have really been actually recorded in out Twitter data. And it turned out that only 147 were recorded. This means that our system has captured 109 out of the total 147 *recorded* outages, which constitutes 74.1% for the *Twitter-based recall* measure. The rest 25.9% of the power outage events mentioned on Twitter were represented by only 1-2 tweets and, thus, were missed by our system.

One nice property of this data and the performance measures is that they correspond to the *actual* power outages. However, this data corresponds to only one geographical region of Utility XYZ (out of 281 regions in the US) and only to the three months of collected data. Therefore, it is also important to test the performance of our method on all the regions and over a longer time period. In the next section, we present such a study where we evaluate the performance of our system against the power outages, as recorded by various "authoritative sources."

B. Outage Data Based on Reliable Twitter Accounts

In this section we evaluate performance of our system vis-a-vis power outages detected by tweets coming from various *reliable sources,* such as news organizations, police departments and other "official" accounts whose job is to report the news and different emergency events through various venues, including Twitter. Unlike the previous case, such reports are filed across most of the US regions, albeit covering not *all* the power outages, but focusing only on the major and most "newsworthy" ones. We construct this set of "reliable tweets" as follows:

1) We, first, identify the set of reliable Twitter user names, such as "NBCNews," "NantucketPolice," "Edison_Electric," etc. We constructed this set based on the list of "reliable" keywords, such as "news", "police", "power", "electricity", "weather", "alert," etc., and collected all the user names from the database D containing these keywords as their substrings.
2) For the set of reliable user names identified in Step 1, we collect all the tweets from database D that these users have generated. We assume that the tweets in the resulting set R are all "true," i.e., are reliable in the sense that the power outages mentioned in them indeed happened in real life. For example, if ABCNews posted a tweet about a power outage in a region, we trust that power outage indeed happened in accordance with the ABCNews tweet.

We next use this generated set R to calculate the precision and recall measures as follows.

Precision Calculations. For each identified burst of tweets (corresponding to a power outage in a certain region), we collect a set of tweets from set R posted around the time of this burst. In our study, we assumed "around" to be one day before and one days after the burst. Further, we manually examine the identified bursts and label them as: (a) *"news"* if the burst is confirmed as an outage by one of our reliable users, such as a news agency or a police department; (b) *"manual"* if the burst is not confirmed by any of our reliable users, but a sufficient number of tweets corresponding to this burst describes a power outage, and they were posted within a small geographic area and a short time period; (c) *"false"* if the burst doesn't relate to a real power outage, i.e., the set of tweets corresponding to the burst are not really about the currently occurring power outage.

For the proposed method, we compute two precision measures, $Precision_{news}$ and $Precision_{manual}$ based on the previously generated labels using formula (1). When calculating $Precision_{news}$, by $True\ Positive$ we assume the set of bursts detected by our system and confirmed by the reliable sources, such as news agencies, and, therefore, they are labeled as *"news"*. For the $Precision_{manual}$ case, by $True\ Positive$ we assume all the detected bursts that are confirmed by either the reliable sources or have been confirmed by manual inspection and, therefore, labeled as *"news"* or *"manual"*. The number of $(\#True\ Positive + \#False\ Positive)$ in the denominator is the same in both cases and is equal to the total number of all the bursts detected by our system.

Recall Calculations. The recall measure in our case is the ratio of all the power outages identified by our system and the total number of power outages reported by the reliable sources. In order to compute the recall measure, we aggregate all the tweets from set R posted in the same region per day. For each (region, day) pair containing at least one tweet from set R we collect: (a) all the tweets posted in this region two days prior and two days after the specified day; (b) all the bursts identified in this region within the same time interval as in (a). Further, we manually examine all the (region, day)

Identified PO bursts	3750
Detected PO confirmed by utilities	152 (from 300)
Detected PO confirmed by inspection	296 (from 300)
Precision	54.7%
Precision Manual	97.6%
PO in validation data	4205
PO identified by system	169 (from 300)
PO discussed on Twitter	242 (from 300)
Total Recall	56.3%
Twitter based Recall	69.8%

TABLE II
PERFORMANCE RESULTS BASED ON RELIABLE TWITTER ACCOUNTS

pairs and label them as: (a) *"identified"* if the power outage discussed by one of the reliable users was also identified by our system; (b) *"missing"* if our system did not identify the power outage while some tweets did; further, we collect the number of tweets referring to this power outage; (c) *"no data"* if our system missed the power outage and there were no tweets posted by the individual users about this power outage.

We also compute the recall measure in the standard way using (2) but in two different "flavors," $Recall_{complete}$ and $Recall_{data}$. For the $Recall_{complete}$ measure, we assume that $False\ Negative$ is the total number of power outages mentioned on Twitter by reliable sources that were missed by our system. These power outages were labeled as either *"missing"* or *"no data"*. For the $Recall_{data}$ measure, we assume that $\#False\ Negative$ is the number of power outages that were mentioned on Twitter not only by the reliable sources but also by individual users, and that were missed by our system. These power outages were labeled as *"missing"*. In both cases, by $True\ Positive$ we assume the set of real power outages that were mentioned in tweets from the reliable sources and were identified by our system. These power outages were labeled as *"identified"*. Note that the size of this set is equal to the size of the set used in calculating the precision measure above (i.e., the number of bursts detected by our system and confirmed by the reliable sources) since both of these two sets refer to the same thing.

The performance results are reported in Table II. In our study, we have identified $3,750$ tweeting bursts over the time period from $08/18/2014$ to $1/25/2015$ (5 months in total). When we labeled 300 of these bursts using the techniques described above, we produced 152 "news," 144 "manual" and 4 "false" labels. Based on these numbers, the precision performance measures are: $Precision_{news} = 0.506$ and $Precision_{manual} = 0.976$. When computing the recall measures over the time period of five months, we, first, obtained $4,205$ (region, day) pairs containing news about power outages. Then we determined the labels for 300 of these pairs and obtained 169 "identified", 73 "missed" and 58 "no data" labels. Based on these numbers, the recall performance measures are: $Recall_{complete} = 0.563$ and $Recall_{data} = 0.698$. Further, although, the $Recall_{data}$ measure is only 0.698 in our case, the average number of tweets posted by the individual users about power outages that our system missed is equal to 1.82. This means that our system missed mostly those power outages that are mentioned in only less than two tweets on average and, therefore, are really hard to detect.

Finally, our system has detected possible reason of power outage in 201 cases of twitter bursts (out of the total of $3,750$ bursts). Also, it has detected the discussed weather conditions in 387 twitter bursts (out of the total of $3,750$ bursts). These numbers mean that in certain cases, people not only report about actual power outages but also provide the reasons for and the weather conditions during these outages.

Note that we haven't compared our system with any baselines. This is the case because it is the very first system for detecting power outages using Twitter data and, therefore, there is no other similar system with which it can be compared.

IV. CONCLUSIONS

In this paper we presented a novel power outage detection method that we have developed as a part of a comprehensive social engagement platform deployed by a major company in the area of enterprise solutions for social media aggregation for the electrical utility industry. The proposed method presented in this paper uses certain predefined key concepts as textual descriptions of power outages, filters the tweets containing these concepts, builds a predictive model that identifies those tweets referring to real power outages, and detects bursts among these identified tweets. These bursts are subsequently tested to see if they really correspond to actual power outages. The detected power outages are reported to the users together with the possible reasons of the outage and the weather conditions in the region at that time.

The proposed method was implemented in our system, tested on the Twitter users, and validated on the power outage data provided by the Utility XYZ from Large Municipality and on the power outage data reported by the news media, police departments and other similar types of outage sources (that we call collectively as "reliable sources"). The validation results show that the precision of our method constitute 93.7% for the XYZ case. For the "reliable sources" case, 54.7% of the detected power outages were confirmed by at least one of these reliable sources. Furthermore, 97.6% of the detected power outages were identified manually as real power outages.

Out of all the power outages identified by the XYZ company, we managed to discover 36.5% of them using our system, giving us the recall value of 36.5%. Furthermore, we managed to identify 74.1% of all the power outages discussed on Twitter in the XYZ region during the 3-month time period. Finally, our system has identified 56.3% of all the power outages reported by the "reliable sources" between 8/18/2014 and 1/25/2015, and has also identified 69.8% of the power outages discussed on Twitter by individual users.

Although some of these performance measure results can be viewed as "low" in other data mining applications that enjoy higher levels of precision and recall, we maintain that these results are very good in our power outage virtual detection application for the following reasons. First, news media, police departments and other "reliable sources" discuss *only major*

power outages in the social media, whereas many $Class\ 1$ tweets refer to smaller types of outages, such as lights going off on certain floors of a hotel. This necessarily brings the precision levels down (such as the 54.7% number) for the "reliable sources" case in our study. Similarly, not all the power outages are being discussed on Twitter. For example, those in the rural and in the less "social-media-friendly" regions may not be captured by tweets. This means that the recall levels are expected to be low in our application, as the numbers of 36.5% and 56.3% demonstrate this. Furthermore, among the power outages that were mentioned by individual users but missed by our system (i.e., 74.1% and 69.8% of them across the two validation mechanisms), an average size of the tweet reference corresponding to those outages was only 1.82 tweets. This means that those outages missed by our system were not significant in terms of the tweeting activities and therefore hardly detectable.

Finally, our system identified a possible reason of power outage in 5.36% of the detected cases, and identified the weather condition in the region at that time in 10.32% of the detected cases. These low numbers are primarily due to the fact that most of the tweets simply state the fact of a power outage and do not report any reasons of why it happened.

As explained in Section I, our system was favorably received by some of the key Electric Utility industry players, who have recognized its role in the future of Outage Management Systems [2] and regarded it as a "new fascinating development" in the industry (see footnote 1). Furthermore, the continued collaboration by one of this paper's co-authors with the US Department of Energy as part of The White House Innovation for Disaster Response and Recovery Initiative Demo Day initiative should ensure this type of technology is afforded the opportunity for significant exposure to numerous US Government Agencies, emergency mangers, and leading technology firms. The prospect of constructing a validated, effective virtual outage detection network at a tiny fraction of the cost of full implementation of the smart grid (with high-end projected costs approaching half a trillion dollars over the next 15 years) solidifies the business case for this technology.

As a part of future work, we plan to adopt our system to other types of power outages and beyond, such as computational advertising, mass transit applications and political campaigns. In addition, although the focus of this work was on the detection of power outages, the proposed method can be extended to other types of outages, such as water, gas, cable and Internet outages. To apply our method to these other types of outages, one should specify a new list of key concepts corresponding to these outages (in Step 1 of Figure 1), as opposed to power outages. Once it is done, all other steps in Figure 1 can be implemented in a somewhat similar fashion. Moreover, the predictive modeling Step 3(a) can and should be done using the active learning approach since, as our study showed, we can build a good predictive model with only 40 learning examples (as opposed to $4,000$ examples, as was reported in Section II-C). Finally, the aspects of these new types of outages will be different from our case, and therefore Step 5 in Figure 1 needs to be adjusted to the new application.

REFERENCES

[1] "Power outage annual report," in *Blackout Tracker United States Annual Report 2014*. Eaton Corporation PLC, 2014.

[2] Malcolm, P. Wade, and B. Lyke, "Collaboration strengthens the customer connection," in *Transmission and Distribution World*, 2015, pp. 24–27.

[3] M. Avvenuti, S. Cresci, A. Marchetti, C. Meletti, and M. Tesconi, "Ears (earthquake alert and report system): A real time decision support system for earthquake crisis management," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014, pp. 1749–1758.

[4] "Lessons learned: Social media and hurricane sandy," in *Virtual Social Media Working Group and DHS First Responders Group*. US Department of Homeland Security, 2013.

[5] "Fact sheet: White house innovation for disaster response and recovery demo day." THE WHITE HOUSE Office of Science and Technology Policy, 2014.

[6] "Announcing the white house innovation for disaster response and recovery initiative demo day." THE WHITE HOUSE Office of Science and Technology Policy, 2014.

[7] D. Keseris, "Social media could provide early warning of power outages." The Telegraph (UK), 2014.

[8] "Number of monthly active twitter users in the united states from 1st quarter 2010 to 4th quarter," 2014.

[9] P. Meier, "Digital humanitarians: How big data is changing the face of humanitarian response." CRC Press (Taylor and Francis Group), 2015.

[10] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 851–860.

[11] P. Earle, D. Bowden, and M. Guy, "Twitter earthquake detection: earthquake monitoring in a social world," *Annals of Geophysics*, vol. 54, no. 6, 2012.

[12] N. Adam, J. Eledath, S. Mehrotra, and N. Venkatasubramanian, "Social media alert and response to threats to citizens (smart-c)," in *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on*, Oct 2012, pp. 181–189.

[13] N. Adam, B. Shafiq, and R. Staffin, "Spatial computing and social media in the context of disaster management," *Intelligent Systems, IEEE*, vol. 27, no. 6, pp. 90–96, Nov 2012.

[14] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] N. Roy and A. Mccallum, "Toward optimal active learning through monte carlo estimation of error reduction," in *in Proceedings of the International Conference on Machine Learning*, 2001.

[17] L. Araujo, J. Cuesta, and J. Merelo, "Genetic algorithm for burst detection and activity tracking in event streams," in *Parallel Problem Solving from Nature - PPSN IX*, ser. Lecture Notes in Computer Science, T. Runarsson, H.-G. Beyer, E. Burke, J. Merelo-Guervs, L. Whitley, and X. Yao, Eds. Springer Berlin Heidelberg, 2006, vol. 4193, pp. 302–311.

[18] J. Kleinberg, "Bursty and hierarchical structure in streams," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. New York, NY, USA: ACM, 2002, pp. 91–101.

[19] R. Ebina, K. Nakamura, and S. Oyanagi, "A real-time burst detection method," in *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*, Nov 2011, pp. 1040–1046.