Developing better Civic Services through Crowdsourcing: The Twitter Case Study

Srushti Wadekar
Dept. of Computer Science
Lakehead University
Thunder Bay,Ontario,Canada
swadekar@lakeheadu.ca

Komal Barge
Dept. of Computer Science
Lakehead University
Thunder Bay,Ontario,Canada
bargek@lakeheadu.ca

Kunal Thapar
Dept. of Computer Science
Lakehead University
Thunder Bay,Ontario,Canada
kthapar@lakeheadu.ca

Devanshu Mishra
Dept. of Computer Science
Lakehead University
Thunder Bay,Ontario,Canada
dmishra@lakeheadu.ca

Rahul Singh
Dept. of Computer Science
Lakehead University
Thunder Bay,Ontario,Canada
rsingh26@lakeheadu.ca

Sabah Mohammad
COMP5112WC Supervisor
Lakehead University
Thunder Bay,Ontario,Canada
mohammed@lakeheadu.ca

Abstract—**Civic technology is a fast-developing segment that holds huge potential for a new generation of startups. A recent survey report on civic technology noted that the sector saw $430 million in investment in just the last two years. It's not just a new market ripe with opportunity it's crucial to our democracy. Crowdsourcing has proven to be an effective supplementary mechanism for public engagement in city government in order to use mutual knowledge in online communities to address such issues as a means of engaging people in urban design. Government needs new alternatives -- alternatives of modern, superior tools and services that are offered at reasonable rates. An effective and easy-to-use civic technology platform enables wide participation. Response to, and a 'conversation' with, the users is very crucial for engagement, as is a feeling of being part of a society. These findings can contribute to the future design of civic technology platforms. In this research, we are trying to introduce a crowdsourcing platform, which will be helpful to people who are facing problems in their everyday practice because of the government services. This platform will gather the information from the trending twitter tweets for last month or so and try to identify which challenges public is confronting. Twitter for crowdsourcing as it is a simple social platform for questions and for the people who see the tweet to get an instant answer. These problems will be analyzed based on their significance which then will be made open to public for its solutions. The findings demonstrate how crowdsourcing tends to boost community engagement, enhances citizens ' views of their town and thus tends us find ways to enhance the city's competitiveness, which faces some serious problems. Using of topic modeling with Latent Dirichlet Allocation (LDA) algorithm helped get categorized civic technology topics which was then validated by simple classification algorithm. While working on this research, we encountered some issues regarding to the tools that were available which we have discussed in the 'Counter arguments' section.**

I. INTRODUCTION

Civic technology is technology that's spurring civic engagement, enhancing citizen communications, improving government infrastructure, or generally making government progressively powerful. "Civic technologies" can be considered as tools that people use to create, support, or serve public good. What is "civic"? Variously defined, "civic" portrays "citizenship"—or, better, the aspects of "living in society". Sure, governments are essential to making society function, but they're just one of several actors. Journalists and artists, government employees and entrepreneurs, parents and teachers, public health workers and neighborhood activists: the "civic sphere" exists where these and other roles meet. It's the public square, the variety of physical and digital spaces where communities connect with each other.

There is a lack of awareness of how 'civic tech' platforms are used and how they may be designed for maximum effectiveness. Numerous data collection methods are used to investigate a well-developed example of civic tech. efficient civic tech can enable extensive democratic participation to improve government services. [2] The overall aim of this research is to (1) Investigate the trending issues that people are posting on social media, (2) Understand the context in which the information has been posted, (3) Identify the nature of the data collected, (4) Get the public opinions or solutions on selected problems. As an outcome, the intention is to identify the trending issues on social media with the help of advanced algorithms. This is to provide guidance for the design of future civic tech platforms. Using twitter for gathering reference data will give us trending and relevant topics up to date. Gathering trending tweets from people can provide us real and impactful problems faced by citizens from the government. This will help government gather real time data from the citizens which will eventually help them.

II. PROBLEM DEFINITION

Civic technology is technology that promotes civic engagement or allows the government to provide citizens services and to establish partnerships with the public. 'Civic technology' describes all public sector and city life innovations, but government technology is a more fitting word for that broader group. Civic technology is where people generally voluntarily lend their talents to help governments do better work. In this paper, In order to bridge the gap between citizens and government we have used twitter as a platform for crowdsourcing. After analyzing the problem, citizens can voluntarily provide solutions for the problem which inclines to Civic tech purpose.

Topic modeling is a type of statistical modeling for discovering the abstract topics that occur in a collection of documents. Categorising topics which are vital to civic relations will help us find the problems in a systematic order. Using of Latent Dirichlet Allocation (LDA) for topic modeling to a particular topic from the documents. Arranging civic problems in groups or categorical format will help government officer select the problems in organised fashion which will help reduce the time for finding relevant and urgent topics from the vast variety of tweets. Latent Dirichlet Allocation (LDA) assign importance points to each words in a sentence or tweets in our case which then helps decide the topic of that sentence. This process of getting topic from a tweet will generate topic and put all the similar sentences in a similar topic forming groups based on topics.

III. RELATED RESEARCH WORK

Presto (2012) recognized five key components that support crowdsourcing activities (and that could in this manner be utilized as a reason for impact assessment): (1) Transparency refers to openness of information, yet in addition how commitments inside a citizen sourcing process are directed, gathered and then utilized. (2) Participation is portrayed as key, and specifically the contribution of diverse, delegate groups. (3) Collaboration is significant, either among residents and different residents, or on account of e-government, among residents and government. (4) Deliberation is especially significant for crowd sourcing including problem solving, and it is asserted that this constructs trust, limit and in the long run more prominent participation. (5) Responsiveness is significant in light of the fact that 'citizens will proceed to partake and participate in government activities just on the off chance that they believe they are being listened to'.

Investigations concerning civic tech have tended to focus on data quality, rather than the human part inside the framework, or a more extensive view on the effect and included value of the civic tech suggestion. The variety of activities that go under the wide flag of citizen sourcing or civic tech results about an absence of standard evaluation criteria and metrics, especially comparable to the effect of the activity. Corresponding to the general phenomena of humans going about as sensors or information providers, and assessing their value, Lathia (2013) depicts how techniques 'remain elusive'[8]. So also, even in a relatively established field, such as, volunteered geographic information, a key issue is the absence of standardized strategies and metrics for evaluation (Antoniou and Skopeliti 2015) [3]. In a specific study related to citizen science, Cox et al. (2015) built up a set of result-based metrics for the Zooniverse projects. The two key criteria used were (1) contribution to science, and (2) public engagement, and these were then separated into measures and proxy metrics [7].

Impact assessments specifically related to civic tech are simply rising. Uppström and Lönn (2013) explain an evaluation of a versatile platform for complaints and issue reporting focused at Swedish districts and the neighborhood occupants. They utilize the system of Nam (2012), however couldn't play out an outcome evaluation because of the designed platform 'not yet being in a productive state'. Lee, Almirall, and Wareham (2016) examined the expansive effect that civic applications had made, especially following open information activities and attempts to consolidate data into applications and services for public advantage. In spite of the fact that the system is vague, their decision was 'applications that had genuine impact for residents or government were few'. Misra et al. (2014) talk about crowd sourcing explicitly with regards to transport planning and conclude that despite the fact that there are issues with investment and data quality, well-planned activities and platforms have incredible potential for assisting with resolving transport issues. U.S Elections [8] research provides the U.S. participants with a geospatial overview of tweets and in 2012 Congress will help scientists perform comparable work in the future. Topic Maps (TM) technology is used for implementation to represent informal learning data in a formal curriculum structure. This led to a development of a framework adapted to formal and informal learning contexts [9]. Followers frequently take actions involved in tweets from leaders, and in many cases, tweets from leaders influence political opinions of followers rather than friends and family. To others, using Twitter leads to political discord, and there's always a discrepancy between what followers on Twitter demand from leaders and what those leaders give them [10]. Social media in government was "described as a technology community enabling public agencies to encourage interaction with people and other organisations". The aim of this study is to address the question: What role does government play on Twitter in public deliberations about civic tech?[11]. In election campaigns Twitter has become an omnipresent device. Candidates, groups, media, and an ever-increasing share of the public use Twitter to discuss, communicate and study public policy reactions. This article presents the findings of a systematic analysis of the literature of 127 research on the use of Twitter in election campaigns.[12]

According to representation and public participation, Escher (2011a, 2011b) considered two civic tech sites built by my Society and found that a significant extent of users were first-time users who may not in any case have occupied with civic action. 40% of those utilizing WriteToThem3 had never previously reached their Member of Parliament, and 60% of the users of TheyWorkforYou4 had never looked into data on their Member of Parliament. Thus, Lee, Almirall, and Wareham (2016) explored numerous of civic applications, and furthermore inferred that FixMyStreet5 (a comparative idea to FMT, empowering detailing or discussing about local issues) had the option to connect with 'another segment that would have been less inclined to report through traditional channels'. Interestingly, notwithstanding, Cantijoch, Galandini, and Gibson (2016) presumed that civic tech (and explicitly other mySociety sites) draws in people who are as of now occupied with civic engagement and utilize online platforms as a means of supplementing and extending their levels of civic or society engagement. Cantijoch also found that users were not representative of the general population, being older and predominantly male.[6]

IV. COMPUTATIONAL METHODS

Social media sites, and other networks offer streaming APIs to provide real-time data to web clients. The data generated at the time we run the collection script will be collected using the Streaming API. Tweets can be downloaded with unique keywords or hashtags. To use any of Twitter's APIs, first we need to collect a set of Twitter API keys that will be used to connect to these API. For Collecting tweets, we have used python Library called Tweepy to connect to the Twitter API and download the data. Tweets containing civic related information such as public issues, current affairs, transportation issues etc. can be obtained using tweepy filters. With the help of standard stream parameters, we can obtain more sophisticated filters such as tweets according to specific times, geographical locations etc. The downloaded data is stored into csv file with the help of python file script and punctuations are removed before

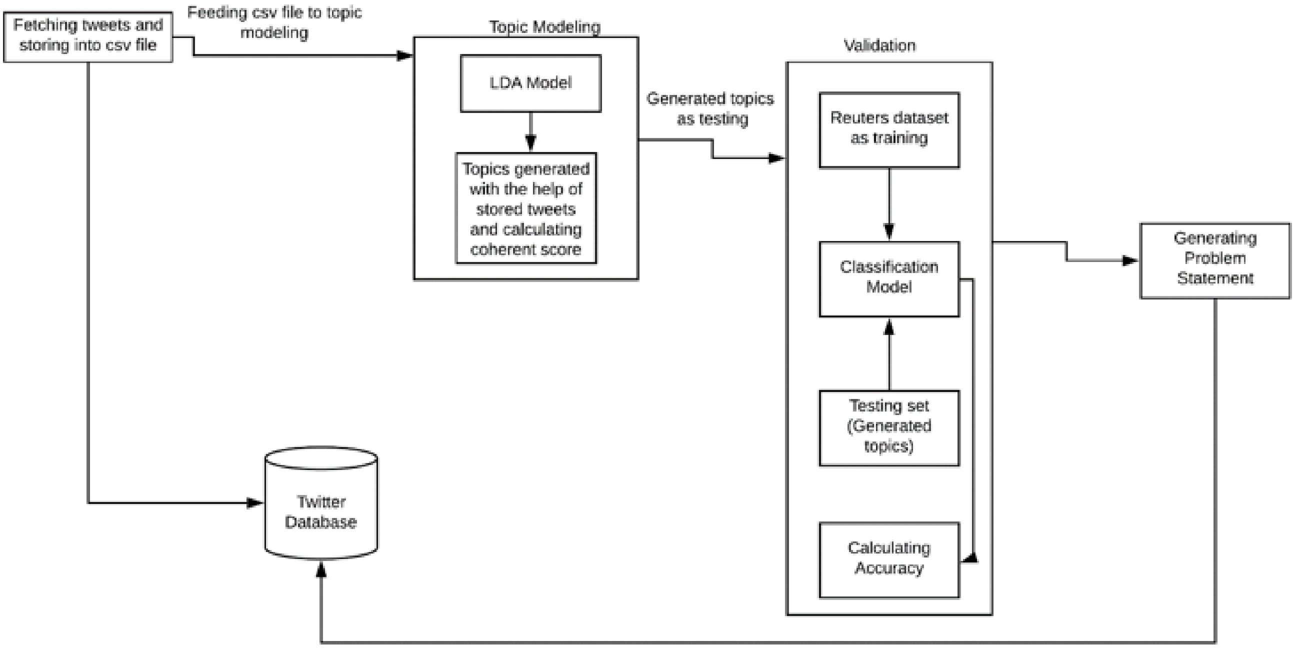
Fetching tweets and storing into csv file
Feeding csv file to topic modeling
Topic Modeling
LDA Model
Topics generated with the help of stored tweets and calculating coherent score
Generated topics as testing
Validation
Reuters dataset as training
Classification Model
Testing set (Generated topics)
Calculating Accuracy
Generating Problem Statement
Twitter Database

Fig 1: Overall Block Diagram

storing into file. The overall working flow of the project with the help of block diagram and use case diagram is shown in Fig 1 and Fig 2 respectively.

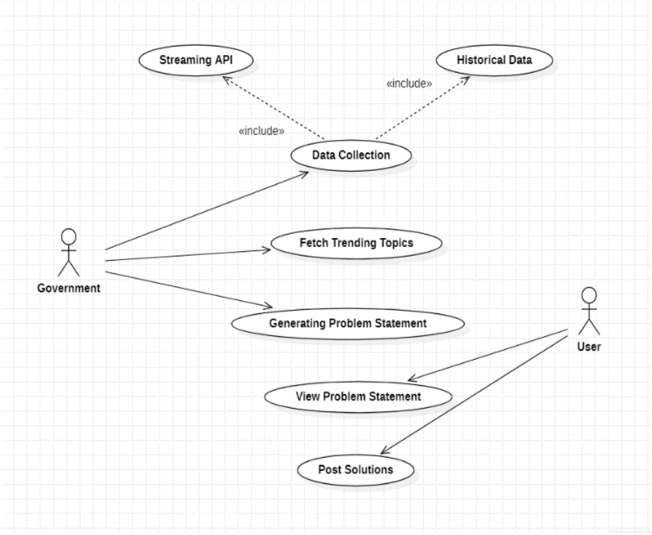

Fig 2: Use Case Diagram

A topic model is a type of statistical model for the discovery of the abstract "topics" present in a document set. Topic modelling is a widely used text-mining method for finding secret semantic structures in a body of text. Intuitively, given that a document is about a specific subject, one should expect similar terms to appear frequently in the document. The "topics" created by topic modelling techniques are clusters of similar words. A topic model expresses this concept in a statistical context that allows the analysis of a collection of documents and the exploration, based on the word statistics in each, of what the topics might be and what the balance of topics in each document is. Topic models are often referred to as probabilistic topic models, which refer to statistical algorithms for discovering an extensive text body's latent semantic structure. The latent Dirichlet allocation (LDA) is a generative statistical model in natural language processing that enables sets of observations to be interpreted by non-observed groups explaining why certain parts of the data are identical. For example, if observations are words collected in documents, it posits that each document is a mixture of a few topics, and that the existence of each word is due to one of the topics of the document. For classification of civic related topics, Reuters dataset have been used. Reuters is a benchmark repository for the classification of documents. The result shows the top trending topics generated with the help of topic modeling.

V. PROTOTOTYPING

For collecting reliable information, twitter streaming API is used in our system and for historical data a csv file containing all tweets related content. Consumer key and Token key are generated to establish API connection. While using twitter API, we must consider rate limit factor. Rate limiting of the standard API is primarily on a per-user basis or more accurately described, per user access token. If a method allows for 15 requests per rate limit window, then it allows 15 requests per window per access token. When using application-only authentication, rate limits are determined globally for the entire application. If a method allows for 15 requests per rate limit window, then it allows you to make 15 requests per window on behalf of your application. This limit is considered completely separately from per-user limits. Rate limits are divided into 15-minute intervals. All endpoints require authentication, so there is no concept of unauthenticated calls and rate limits. There are two initial buckets available for GET requests: 15 calls every 15 minutes, and 180 calls every 15 minutes. To overcome this rate limit problem, we have used twitter streaming API for fetching tweets. The Streaming API has rate limiting and access levels that are appropriate for long-lived connections. Leveraging the Streaming API is a great way to free-up your rate limits for more inventive uses of the Twitter API. Rate Limiting information for the Streaming API is detailed on Connecting to a streaming endpoint as shown in Fig 3. Different filters have been applied in order to fetch tweets and stored into CSV file for topic modeling.

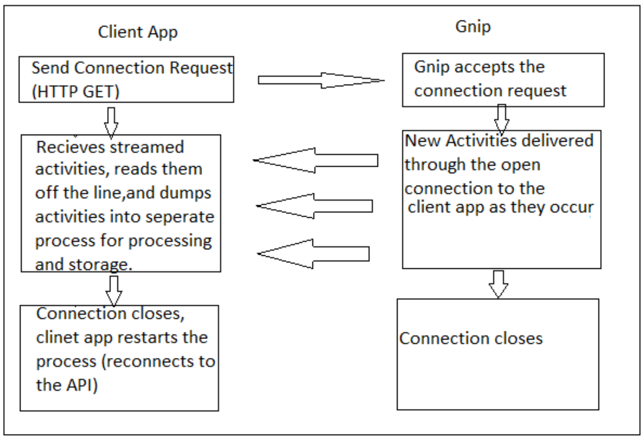
Client App
Gnip
Send Connection Request (HTTP GET)
Gnip accepts the connection request
Recieves streamed activities, reads them off the line,and dumps activities into seperate process for processing and storage.
New Activities delivered through the open connection to the client app as they occur
Connection closes, clinet app restarts the process (reconnects to the API)
Connection closes

Fig 3: Streaming API Flow

Topic models offer an easy way to examine large quantities of unlabeled data. A topic is a collection of terms mostly grouped together. A method for the topic modelling takes one text (or body of texts) and searches for trends in the use of the words. The topic models are computer algorithms which identify latent word patterns using word distribution in a document collection. The consequence is a selection of subjects composed of word-clusters co-existing according to certain trends in these texts. Hence, in our research, topic modelling will be used to fetch the trending issues based on Twitter data. Collection of current tweets with historical tweets will become the dataset for topic modeling considering each tweet as one sentence. Topic modelling using Mallet will be applied on the twitter data corpus to find the trends in the data. If, without necessarily reading any document, you have hundreds of documents from an archive and would like to understand something of the archive, then modelling topics can be a good approach. Topic models represent a computer family that extracts topics from texts. A list of words in statistically significant way is a subject for the machine. An email, a blog post, a chapter of a novel, an essay in the newspaper, and a journal report may be a file that includes some unstructured text. We say that no machine understandable annotations inform the device the semantic importance of the terms in the document. Topic modeling programs do not know anything about the meaning of the words in a text. However, they believe that every text (by an author) is constructed by choosing terms from potential word baskets, where each basket is equal to a topic. If that is true, then it is possible to decompose a text mathematically into likely baskets from which the words came first. This method is pursued through the device, until it concludes that the word is most definitely represented in containers, which we consider themes.

In this paper, Topic modeling starts by removal of punctuations and symbols and tweets are converted to lowercase with the help of regular expression library. Topic map is implemented with the help of tokenization and lemmatization preprocessing. Tokenization is assigning importance value to each word in the tweets. Importance value can be assigned based on number of occurrences, relevance and use of word in the tweet. Normalized words are then mapped to its integer id's. These collections of words are then added to bag of words. Bag of Words is a way of extracting features from text for use in models. Any information about the order or structure of words in the document is discarded. The model is only concerned with weather known words occur in the document, not where in the document. LDA is a type of unsupervised learning that considers documents as bags of words (i.e., no matter how order is). LDA works by first making a key assumption: picking a set of topics and then picking a set of terms for each topic was the way a document was produced. To do this it does the following for each document m:

1. Assume there are k topics across all of the documents
2. Distribute these k topics across document m (this distribution is known as α and can be symmetric or asymmetric, more on this later) by assigning each word a topic.
3. For each word w in document m, assume its topic is wrong but every other word is assigned the correct topic.
4. Probabilistically assign word w a topic based on:
   a. what topics are in document m
   b. how many times word w has been assigned a topic across all the documents (this distribution is called β, more on this later)
5. Repeat this process several times for each document.

The LDA model is then used to determine the topic along with the word count. The model complexity and coherence calculation are done to determine the performance of the model. To customise the model as per the requirements we can tweak value of α (a matrix where each row is a document and each column represents a topic) and β (a matrix where each row represents a topic and each column represents a word). The Reuters dataset is used for the classification of the document and is used to train the model. Reuters dataset is usually used for document classification. It has 90 classes, 7769 training documents and 3019 testing documents. The average number of words per document, grouped by gender, is from 93 to 1263. The training model of Reuters has vocabulary size of 35247.This model is then used to test the topic modelling document generated to test the accuracy of the generated model. The twitter data is fed as a testing sample on the model trained with Reuters dataset. The topic coherence is used for the evaluation of topic models: methods for generating topics automatically by latent variable models from a document set. -- of the topics produced consists of words, and the topic is consistent with the topic N words This model is then used to test the topic modelling document generated to test the accuracy of the generated model. Now, Importation of word cloud library is performed which helps in joining the different processed titles together and visualization of the word cloud. Word Cloud is a technology for data processing used to display text data that displays its frequency or value in the size of each word. A word cloud can be used to illustrate important textual data points. Finally, the top 10 most common words related to civic technology are shown graphically to help in analysis or study purpose.

The overall pseudocode for the case study is as follows:

1. Fetching tweets from twitter database and storing into csv file.
2. The file is then fed for topic modelling.
   a) LDA model is used to for the generation of topics.
   b) The generated topics are then used for calculating coherent score.
3. Further, Generated topics are fed as testing for validation.
   a) Reuters dataset is used for the training purpose.
   b) The testing set is then used in the classification model.
   c) Accuracy is calculated for validation.
4. Based on the topics, problem statement is generated.

VI. EXPERIMENTAL RESULTS

Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. Coherence measure defines if the topics generated support each other. It has various measures like C_V, C_p, C_uci, etc. Our model has been evaluated by using the 'c_v' coherence measure. Coherence measure is used to validate the results which are the topic generated from LDA model. For our streaming tweets, coherence score of 88% was observed for the topics generated through LDA model. To validate further the results of our model, a classifier was used for validation of topics generated. Classification was done by considering reuters dataset for training process. The classifier was seen to have an accuracy of 81% while validating the topics received from the LDA model.

VII. CONCLUSION

The case study of gathering current and historic tweets and passing the tweets in topic modelling to get the relevant topics from the tweets. Topics gathered from topic modeling with LDA method resulted with informative civic tech topics which can be converted into necessary problem statements. These problem statements can be then published at any popular public sites. Then solutions can be obtaining from public about the relevant problems. This cycle of gathering data from crowd to solve problems faced by crowd is the basic idea behind this paper. We have also validated the results from the topic model with the help of classifier. We have trained the classifier with many news related dataset. In future we hope to improve the dataset and can also improve the model to get trending topic which will help in solving civic problems.

VIII. REFERENCES

[1] Andrew May, Tracy Ross, "The design of civic technology: factors that influence public participation and impact", Journal of Ergonomics, Volume 61, 2018.

[2] Antoniou, V., and A. Skopeliti, "Measures and Indicators of VGI Quality: An Overview", ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume 3, October 2015.

[3] Avinoam Baruch, Andrew May, and Dapeng Yu, "Computers in Human Behavior The Motivations, Enablers and Barriers for Voluntary Participation in an Online Crowdsourcing Platform", Journal Computer in Human Behaviour, Volume 64, November 2016.

[4] Naill Bolger,Angelina Davis, and Eshkol Rafaeli, "Diary Methods: Capturing Life as lived", Annual Reviews of Philosophy, Volume 54, November 2018.

[5] Joe Cox, Eun Young Oh, Brooke Simmons, Chris Lintott, Karen Masters, Gary Greenhill, Kate Holmes, "Defining and Measuring Success in Online Citizen Science", Computer Science & Engineering, Volume 17, August 2015.

[6] Lathia N, "The Human Sensor: Bridging between Human Data and Services", In: Michelucci P. (eds) Handbook of Human Computation, Springer, November 2013.

[7] Marta Cantijoch, Silvia Galandini, Rachel Gibson, "A mixed-method study of civic websites and community efficacy", Journal of New Media & Society, October 2016.

[8] A.J. Milion, Bradley Wade Bishop, Sean P. Goggins, "An Exploration of "Localness" on Twitter during the 2012 U.S. Elections", 2012

[9] Tien-Chi Huang, Chia-Chen Chen, "Animating civic education: developing a knowledge navigation system using blogging and topic map technology", Journal of Education Technology & Society 16 (1), 79-92, 2013

[10] Parmelee, John H., and Shannon L. Bichard, "Politics and the twitter revolution: How tweets influence the relationship between political leaders and the public.", Lexington Books, 2011

[11] Charles, Crystal R., and J. Ramon Gil-Garcia., "Government Engagement with the Civic Tech Community on Twitter: The case of the New York City School of Data.", EGOV-CeDEM-ePart 2018 (2018): 61

[12] Jugherr, Andreas., "Twitter use in election campanings: A systematic literature review.", Journal of information technology & politics 13, no. 1 (2016): 72-91.