Tweet Semantic Classification in Civic Engagement Research

S. Compion, P. Croft, J. J. Li, K. Ngoy, and F. Qi

***Abstract*—This paper presents a proposal to apply Latent Semantics Indexing to automatically classify Twitter tweets into different categories, in order to create a location-based geographic map of students' civic engagement intensity and correlate it with social behavior. Since the work is at the proposal stage, the focus of this paper is on the proposed research methodology stemming from a pilot study we conducted with Facebook data. We implemented the methodology in a posting classification tool working with Facebook API. During our validation, the tool extracted 100 postings and classified them into five categories of politics, entertainment, science, technology, and daily life. Once adopted to analyzing tweets, we hope to contribute to the field by applying machine learning algorithms to the study of social behavior with focus on measuring youth civic engagement.**

***Index Terms*—Classifier, latent semantics indexing, machine learning, pattern recognition.**

I. Introduction

The advance of computer power enables the application of Artificial Intelligence (AI) and machine learning to various fields [1], [2], ranging from autonomous cars to the automation of scientific experiments. Deep learning uses gigantic neural networks to recognize and classify a large quantity of data such as text [3], sound [4] and images [5]. Besides application to scientific fields, more and more "deep learning" has been applied to social science for data modeling and prediction.

The widespread use of social media and explosive growth of publicly accessible content on the internet provides a rich source of data that is relevant to our study of social science, in particular, civic engagement. Real-world events, sentiments, and social landscapes leave trails or reflections online that can be collected and analyzed scientifically. The goal of our research is to "mine" such online data, and map out an approximation of the real world landscape of public opinions to find the most effective civic engagement. Among the popular social media platforms, Twitter is a microblogging service and one of today's largest social networks. It has more than 336 million monthly active users [6] and generates over 500 million "tweets" per day [7].

We propose to use Twitter data to model and virtually map the college educated Millennials' tweets. Their expressions of civic concerns about campus gun-violence on a democratic social media platform are collected and analyzed to show their correlations to social effects. We will compare New Jersey college students to peers across the country to determine differences in their viewpoints and styles of expressing civic engagement.

U.S. Millennials include those born after 1990, who have come of age in an era marked by rising social fragmentation, heightened surveillance, public safety "lock-downs," and active shooter drills. Often called Digital Natives [8], Millennials consume and produce information in vast quantities and, while accused of being self-absorbed and selfie-snapping, they also engage as citizens in creative digital ways. They use social media and click-activism to crowdsource knowledge and to mobilize resources for civic purposes. This was exemplified by the recent youth-driven response to a school shooting at Marjory Stoneman Douglas High School in Parkland, Florida, after which students used social media to galvanize public interest that helped them lead a delegation to lobby legislators at Capitol Hill resulting in political support for gun policy reform and improvement.

We plan to mine and utilize Twitter tweets to map out the virtual landscape of U.S. Millennials' views about gun violence and school shootings, and to explore, by state, the impacts of these tweets in mobilizing civic action (such as sending a student delegation to lobby legislators). Such research has not been done before and the results of the research will help us better understand the Millennials' civic engagement, as well as identifying effective technologies and methods for promoting political agendas through correlating opinion tweets with social effects.

Our research focuses on the role of Twitter as a democratic social media platform where Millennials find space for civic expression and mobilize political reform. Particular attention will be given to issues of widespread public interest such as that related to school shootings. The social effects such as student delegation and news annotation are also announced through Twitter. Thus our first research task is to classify gun policy related tweets into categories of opinions and effects.

In the future, our research can be further extended into views about environmental justice, climate change, and social inequality. By analyzing and mapping the social media data, we hope to answer three questions: 1) How do the virtual patterns of expressed opinions about gun-violence overlap with the U.S. geophysical and socio-political landscape? 2) Are there differences in the opinions and

Manuscript received Augugst 24, 2018; revised October 15, 2018. This work was supported in part by the Kean StemPact research grant.

S. Compion is with Interdisciplinary and Global Studies, Kean University, Union, NJ 07083 USA (e-mail: scompion@kean.edu).

J. J. Li is with School of Computer Science, Kean University, Union, NJ 07083 USA (e-mail: juli@kean.edu).

P. Croft, K. Ngoy and F. Qi are with the School of Sustainable and Environmental Sciences, Kean University, Union, NJ 07083 USA (e-mail: {pcroft|kngoy|fqi}@kean.edu).

doi: 10.18178/ijmlc.2018.8.6.751

sentiments toward gun-violence between Millennials with or without a college campus experience? 3) Does the number of tweets and their reached audience have a positive impact on the effect of the civic engagement?

The project is envisioned as an interdisciplinary initiative to the study of youth citizenship practices and civic engagement in the USA, and to examine the role of the campus-experience in the process of civic education. The goal of this broader endeavor is to give voice to cultural experiences, forms of civic expression, and styles of collective engagement as shared by Millennials. The long term emphasis on studying civic engagement necessitates an interdisciplinary approach to research, creative, and scholarly works.

The rest of the paper describes how to automatically classify tweets into different categories through our previous work with Facebook postings. We will also use the data to correlate opinion postings with effect postings and identify the most effective civic engagement methods.

II. Description of the Underlying Technology

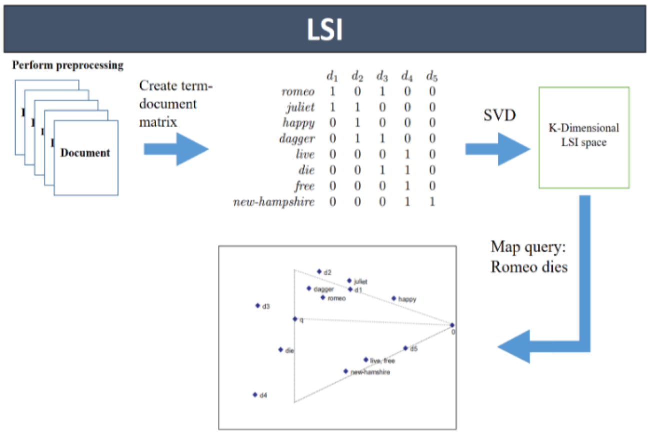
Research has been conducted to use twitter data to model electoral behavior [9], forecast unemployment rates [10] and even flu outbreaks [11], and map hate speech across the US [12]. In our work, we will collect, mine, and map twitter data that is relevant to civic engagement. We will use Latent Semantics Indexing (LSI) to identify relevant tweets and classify them into different categories. The advantage of LSI method is that it understands the semantics of posting and would be able to identify postings as gun control related even if the posting does not include the key word “gun”.

When searching a tweet for a category, lexical methods may fail to provide accurate results. Some words in the tweets under search may be related to a category through synonymy or context. A tweet may not match a category word for word, but its content does match semantically. As humans we can easily distinguish that there is a match between the category: “fishing” and a tweet containing the words: “rod”, “river”, “bait”, and “lure”. Another example is that words such as “lockdown”, “alert” and “suspicious” could be related to school shooting without words such as “gun” or “shooting”. A computer searching goes strictly for a match between the category word and the posting will fail to return the tweet as a relevant one for further study. This situation is where LSI comes in as the best machine leaning algorithm for semantic classification of documents ort postings or tweets.

LSI tries to overcome the problems of lexical matching by using statistically derived conceptual indices instead of individual words for retrieval and classification. LSI assumes that there is some underlying or latent structure in word usage that is partially obscured by variability in word choice. A truncated singular value decomposition (SVD, a well-known matrix manipulation in linear algebra) is used to estimate the structure in word usage across a tweet posting. Classification is then performed using the database of singular values and vectors obtained from the truncated SVD. Performance data shows that these statistically derived vectors are more robust indicators of meaning than individual terms [13].

A. LSI workflow

Fig. 1 shows a workflow of LSI method [14] for text classification. The LSI method begins by performing some preprocessing on a text corpus, as implemented in our LSI tool for Facebook postings and yet to be extended for tweets. Our LSI implementation starts by stripping the postings under study off the “stop words”, which include articles (e.g. a, an, the), prepositions (e.g. at, by, in, to, from, with), and conjunctions (e.g. and, but, because). In many situations, transitional words, which don’t add meaning to sentences, can also be removed, such as “therefore”, “but”, “so”, and “subsequently”. So other words such as those representing sounds can also be excluded. We also remove all ASCII accents from each posting under study so that our software will work for any text in any format.



	d_1	d_2	d_3	d_4	d_5
romeo	1	0	1	0	0
juliet	1	1	0	0	0
happy	0	1	0	0	0
dagger	0	1	1	0	0
live	0	0	0	1	0
die	0	0	1	1	0
free	0	0	0	1	0
new-hampshire	0	0	0	1	1

Fig. 1. LSI workflow illustration.

Fig. 1 shows that our LSI tool then creates a term-document matrix where rows correspond to unique terms in the text corpus, and columns correspond to postings in the text corpus. The rows are related to a list of keywords and the columns are the documents denoted as d1, d2, d3, d4, d5, which indicates 5 postings under study. This allows every term and posting to be expressed as a vector; where every element in a vector represents the degree of participation of the posting or term in the corresponding concept [15] of a category. This technique allows us to expose document-document, document-term, and term-term similarities to be used for categorization. The term-document matrix represents a dimensional space of posting context.

The word list will expand as more documents or postings are added because all new key words will be included in the word list. Long documents might result in a long list of keywords. The limit on the size of tweets is 280 characters which makes LSI extremely suitable for analyzing tweets without the concern of the word list becomes too large.

After obtaining the term-document matrices, our LSI tool then performs SVD, a software included in most of Mathematics packages, on the matrix. LSI method is actually an application of performing SVD on a term-document matrix to represent the matrix in a lower dimensional space for classification. SVD is a mathematical algorithm that decomposes the factorization of a matrix into three other matrices, one of which is a diagonal matrix that can be collapsed into a vector representing the identity of the matrices.

Without performing SVD, the dimensional space of the original term-document space is too large to accurately return similarities of text belonging to the same category. SVD creates a low-rank approximation for a value k that is much smaller than the original rank of dimensions. This is similar to human viewpoints of documents with abstraction.

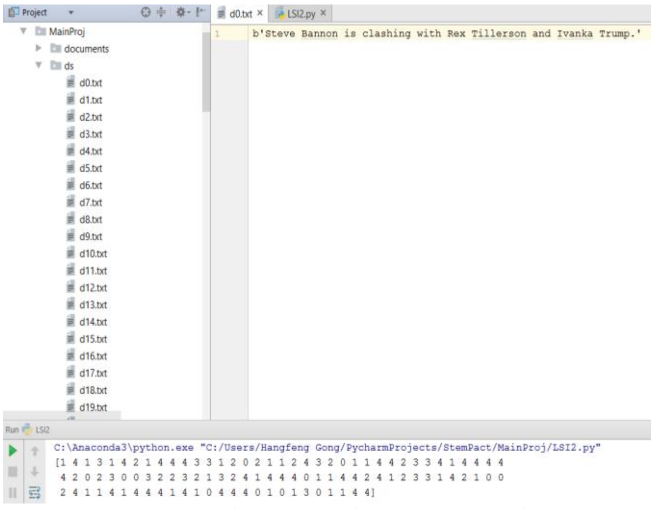
As shown in the matrices of Fig. 1, when a human reads the posting one which has both words of “Romeo” and “Juliet”, (s)he will most likely think of the category of “William Shakespeare”. SVD factorization has the similar effect of obtaining the language structure of the posting by abstracting the identity matrix.

After SVD, we now have a k-dimensional LSI space onto which we can map a classifier’s vector representation onto. The classifier’s vector may now be compared to all other document vectors by performing a cosine similarity analysis. The closer the distance between a category and a text document, the more they are related. At this point, each category of posting is represented in a vector. All new postings that need to be classified will first be represented in a vector and then to be compared with the vector of each category. The closest one is the category of the posting.

B. LSI Tool

Our contribution to LSI is our extension of LSI as a machine learning technology for text classification of Facebook postings or Tweets, rather than its original intended usage for simple queries. With our LSI classification and its corresponding tool implementation, we can now classify in real-time documents, postings, or even tweets in the future. Here are some of our initial results applying our LSI algorithm and tool to Facebook posting classification.

We tried our LSI tool on Facebook postings fetched through APIs provided by Facebook. Fig. 2 shows the running of the tool implemented in Python and the graphical user interface of the tool.



```
b'Steve Bannon is clashing with Rex Tillerson and Ivanka Trump.'
```

```
C:\Anaconda3\python.exe "C:/Users/Hangfeng Gong/PycharmProjects/StemPact/MainProj/LSI2.py"
[1 4 1 3 1 4 2 1 4 4 4 3 3 1 2 0 2 1 1 2 4 3 2 0 1 1 4 4 2 3 3 4 1 4 4 4 4
 4 2 0 2 3 0 0 3 2 2 3 2 1 3 2 4 1 4 4 4 0 1 1 4 4 2 4 1 2 3 3 1 4 2 1 0 0
 2 4 1 1 4 1 4 4 4 1 4 1 0 4 4 4 0 1 0 1 3 0 1 1 4 4]
```

Fig. 2. Screenshot of our LSI tool for document classification.

C. LSI Initial Results

Fig. 2 shows that the interface of the tool includes three sections. The left is the list of postings under classification. The right is an example of a very simple posting and the bottom is the running and the output of the tool with vectors being displayed for debugging and validation purpose.

Our experimental implementation and results indicate that our LSI tool was able to classify Facebook postings into 5 categories (politics, entertainment, science, technology, and daily life) even if the posting doesn't have the keyword but is only related semantically. Our validation is based on human understanding of the concepts of those 5 categories. We propose to further validate the method and the tool through the study of tweets.

The window at the bottom of Fig. 2 is a vector representing a new document to be classified based on the previously trained documents. The display of the vector allows us to keep track of the number and appearance of words.

Fig. 3 also shows the results of running LSI on 100 Facebook posts from the New York Times’ Facebook page. After performing LSI, each posting was clustered into 5 categories then projected onto a graph for visual representation.

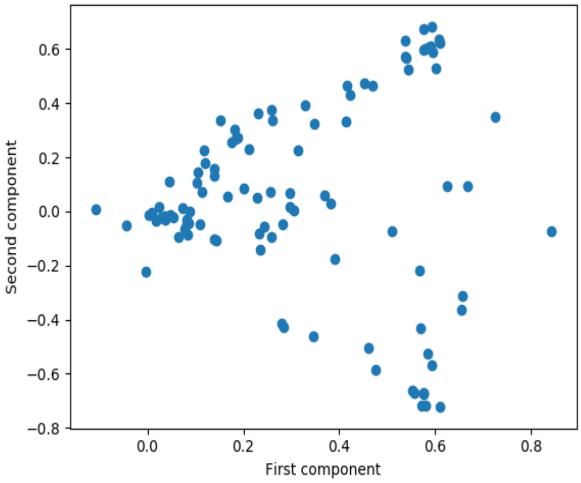


Fig. 3. 100 Postings projected onto a 2-dimensional space.

Fig. 3 shows there are some clusters in the postings and also some blurry boundaries between postings. Besides the vector representing characteristics of each posting, our LSI tool also calculates the cosine distance between the vectors so that they can be used in the second layer of clustering to group postings into various categories.

The second layer of conventional machine learning on numerical values can be performed to identify postings in divided up categories. For example, support vector machine can be used to put the dots in Fig. 3 into categories. K Nearest Neighbor (KNN) may also be used if there is no mathematical model and when the classification needs to be carried out in real-time for instant response to users. We can also use decision tree with words as conditions to divide up categories. Fuzzy logic can also be applied to decision trees to allow for each posting to be belong to multiple categories.

III. Future Work

The LSI algorithm described in the previous section allowed us to build an automatic tool to classify Facebook postings into categories. We will extend it for putting tweets into various categories, and then perform the following research to obtain meaningful information. For example, we will map the tweet categories into physical locations using geotagged tweets, and virtual communities using tweet-retweet-mention relationships. We will count the number of tweets in each category for each geo-location and virtual community.

So far, our LSI has been applied to Facebook postings. In order to expand it for tweets, we need to conduct the following steps of research.

Literature review and identification of keyword search terms for the categories of related or not-related to gun policies.

It is necessary to conduct background literature/desktop research to identify a few focal keywords that can be used as category terms for data mining. For example, for school shooting, keywords may include “school shooting”, “gun control”, “shooting”, “AR-15”, and can be categorized alongside “for” or “against” gun reform terms. In addition, in order for the data to be analyzed using methods such as sentiment analysis [16] or opinion mining, we also need to search for language that indicates or implies an active level of engagement [17].

Searching for level of engagement may include focusing, for example on: a user being “followed” by an activist group; an open expression for or against gun policy reform; active attempts to persuade others to agree with an opinion; signing an online petition; asking others to sign a petition; and joining or following an activist group/community, etc. We will use the keywords to obtain sufficient data in training set for each category.

1) *Data collection, machine training and processing*

Once keywords are specified, tweet data collection can be achieved using two Application Programming Interfaces (APIs) [18]. A “streaming API” can have Twitter send tweets that meet your criteria to your application instantly as they are posted by users. A “REST API” allows applications to search for relevant historical tweets. Search criteria can include username, text content, date, geotag, etc. Associated metadata will also be obtained that include posting time as well as the number of times a tweet has been retweeted.

Our LSI tool will first fetch the tweets and put them into a training set. The tweets in the training set are put into categories based on keywords. Once the LSI model is trained and obtained a singular vector for each category of our interest. We can start using our LSI tool with trained model to detect patterns and categorize tweets in real-time during their streaming.

2) *Data processing and pattern recognition*

Data processing can take on several forms. First, a simple tweet count can be done for the different topics based on different levels of engagement. The counts can also be correlated with other factors such as location, etc. Second, sentiment analysis can be conducted to obtain a Twitter sentiment index for obtaining the perception of twitter users taking part in different civic engagements. Third, LSI based text mining can be applied to identify more complex patterns and relationships in the data. And finally, advanced algorithms such as AI linguistic analysis and machine learning classifier can be employed to classify tweets into different virtual categories based on results from LSI.

The identity vector of each category represents its pattern of characteristics. If a tweet is found to be far away from any category, then it is most likely to be an outlier and deserves further investigation. Otherwise, we can count the number of tweets in each category and correlate the numbers with the numbers of effect category. In this way, the most effective tweet category can be identified.

3) *Mapping*

Two research efforts will be taken to create a map that includes information of the following: 1) the spatial distribution of gun-violence related civic engagement through social media in the physical world, and 2) the online communities and network structures of the virtual landscape in tweets related to civic engagement.

The first is achieved through mapping out Geotagged tweets. Twitter supports an optional latitude/longitude field for each tweet. Tweets by users that allow this feature can be mapped on a grid overlaying the physical U.S. [19]. Various data layers can be integrated in this mapping exercise, including political boundaries, population distribution, and other socio-demographic parameters. The second mapping exercise utilizes the results from the community detection and network analyses from the last step in a process called Social Network Analysis. Hotspots, communities of different sizes, as well as network connections among different communities can be mapped out to show the virtual landscape of civic engagement in social media (in our case, Twitter). We envision this as part of a multi-year endeavor to expand the data-use potential for our proposed project.

IV. Concluding observations

This paper presents an extension of LSI for text classification to be applied to Twitter tweets for the study of youth civic engagement. Our contribution includes the extension of LSI for classification purposes, and the implementation and application of LSI to the social sciences as a tool to study civic engagement. Based on our initial results using Facebook data, we propose further research that classifies and analyses tweets to find correlations between online postings and civic behavior to identify effective measures.

We also found that when analyzing a large number of postings, the number of columns can be very large, which requires extensive computing power to handle the classification task. The number of rows can also grow greatly if new words appear in new documents. We will investigate further reducing the words that do not add new semantic meaning to the context or using one word to represent all synonym words.

Our study of performance and its improvement through GPU acceleration is underway. We realize that LSI can be computationally intensive and high-performance computing might be needed to train and process a large number of tweets. LSI can also be applied to other fields of social science, such as training the machine to identify historical artifacts and relationships among historical figures. We hope that our work will contribute to the application of AI and machine learning technologies to the social sciences.

ACKNOWLEDGMENT

We would like to thank Kean University's College of Natural, Applied and Health Sciences for giving us this interdisciplinary research opportunity.

REFERENCES

[1] D. Bassu, R. Izmailov, A. McIntosh, L. Ness, and D. Shallcross, "Centralized multi-scale singular vector decomposition for feature construction in lidar image classification problems," presented at applied imagery pattern recognition workshop, 2012.

[2] D. Bassu, P.W. Jones, L. Ness, and D. Shallcross. (2016). Product formalisms for measures on spaces with binary tree structures: representation, visualization, and multiscale noise. [Online]. Available: https://arxiv.org/pdf/1601.02946v2.pdf

[3] Y. Rossikova, J, J. Li and P. Morreale, "Intelligent data mining for translator correctness prediction," *Cloud Security 2016*, Columbia university NYC, April 2016.

[4] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. the 30th International Conference on Machine Learning, Atlanta*, Georgia, USA, vol. 28, 2013.

[5] D.-A. Clevert, T.s Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by Exponential Linear Units (ELUS)," in *Proc. the International Conference on Learning Representations*, San Juan, Puerto Rico, May 2-4, 2016.

[6] Statista. Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2018 (in millions). [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[7] ILS. *Internet Live Statistics.* [Online]. Available: http://www.internetlivestats.com/twitter-statistics/

[8] M. Prensky. (2001). [Online]. Available: http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf

[9] D. Gayo-Avello, "A meta-analysis of state-of-the-art electoral prediction from Twitter data," *Social Science Computer Review*, vol. 31, no. 6, pp. 649-679, 2013.

[10] H. Choi and H. Varian. (2009a, April 10). Predicting the present with Google trends (Technical report). Google Inc. [Online]. Available: http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf

[11] J. Ginsberg, M. H. Mohebbi *et al*., "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012–1014, 2009.

[12] What Would a Floating Sheep Map? The Manifesto. (June 5, 2018). [Online]. Available: www.floatingsheep.org

[13] R. Barbara, "Latent semantic indexing: An overview," *Springer Reference*, Feb. 2017.

[14] T. K. Landauer, P. W. Foltz, and D. Laham, *Introduction to Latent Semantic Analysis. Discourse Processes*,vol. 25, pp. 259-284, 1998.

[15] T. Alex, *Handbook of Latent Semantic Analysis*, Feb. 2017.

[16] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proc. the Workshop on Languages in Social Media*, Association for Computational Linguistics, 2011, pp. 30–38.

[17] J. S. Milosevic and I. L. Zezelj, "Civic activism online: Making young people dormant or more active in real life?" *Computers in Human Behavior*, vol. 70, 113–118, 2017.

[18] A. Potts, W. Simm, J. Whittle and J. W. Unge, "Exploring "success" in digitally augmented activism: A triangulated approach to analyzing UK activist Twitter use," *Discourse, Context & Media*, vol. 6, pp. 65–76 r, 2014.

[19] J. Krumm, A. L. Kun, and P. Varsanyi, "'Tweet count' urban insights by counting tweets," in *Proc. UBICOMP/ISWC'17,* Maui, Hawaii, USA, Sept. 11-15, 2017.

Sara Compion is a cultural sociologist who received her Ph.D. from the University of Kentucky in 2016. Her research focuses on volunteering, civil society, and democratic development in Southern Africa. She has published numerous articles in sociology, interdisciplinary, and nonprofit-sector journals and has been engaged as a consultant on evaluation assignments of social welfare enterprises in South Africa. Now at Kean University in the USA, she coordinates the global studies program and directs the Center for Interdisciplinary Studies.

P. Croft served as THE President of the National Weather Association and as a member of its Strategic committee. His research relates to weather-related phenomena and hazards, regional climate impacts, and operational meteorology for impact assessment, mitigation, and prediction. Dr. Croft receives Ph.D, M.S. and B.S. from Rutgers University. Dr. Croft is now Kean's associate VP for Academic Affairs.

J. J. Li came to the north America from China in the late 1980s. She received her Ph.D. in software engineering from University of Waterloo Canada in 1996. She received her B.S. of computer science in 1991 from Dalhousie University, Halifax NS Canada.

She is now a faculty member of School of Computer Science at Kean University. Before joining Kean, she was a lead research scientist at Avaya Labs Research, formerly a part of Bell Labs Research. She has published over 100 peer reviewed papers and she holds 20 patents. Prior to Bell Labs, she was a research scientist at Bellcore, now Applied Communication Sciences. Her current research interest is in AI and machine learning with the emphasis on its application to health science, software engineering and cybersecurity.

Dr. Li is a senior member of IEEE, a member of ACM and a chapter official of PKP. Dr. Li is the founder of Kean ACMW chapter. Dr. Li is the recent recipient of NCWIT undergraduate research mentor of the year 2018, a rare honor with one faculty member selected each year nationally.

Feng Qi is from China. She has a B.S. in environmental science from Peking University. She received her M.S. in GIS and cartography and Ph.D. in geography from the University of Wisconsin-Madison.

She is a faculty with the School of Environmental and Sustainability Sciences at Kean University. Before joining Kean, she was a faculty with the University of Texas-San Antonio. Her research focuses on geo-computation, visualization, and spatial data mining as well as their applications in environmental modeling and public health. She has directed projects funded by the NSF and NIH and has over 30 publications.

Dr. Qi is a member of Sigma Xi and the AAG. She served as the vice president of the Kean University chapter of Phi Kappa Phi.

Kikombo Ilunga Ngoy received his M.S. and Ph.D. in geography from the Oregon State University, USA.

He is a faculty at Kean University in the School of Environmental and Sustainability Sciences. Before joining Kean University, he was a faculty in Temple University and Vassar College. His research focuses on spatio-temporal landscape modeling including land use and land cover change, and climate change using Remote Sensing and Geographic Information Systems (GIS) in New Jersey and in the Democratic Republic of Congo. Dr. Ngoy is a member of the American Association of Geographers (AAG), and the New Jersey Academy of Science (NJAS).