2017

GIS Investigation of Crime Prediction with an Operationalized Tweet Corpus

Anthony J. Corso  
*California Baptist University*

Abdulkareem Alsudais  
*Claremont Graduate University*

Follow this and additional works at: [https://scholarworks.umass.edu/foss4g](https://scholarworks.umass.edu/foss4g)

Part of the [Business Commons](https://scholarworks.umass.edu/businesscommons), [Engineering Commons](https://scholarworks.umass.edu/engineeringcommons), and the [Social and Behavioral Sciences Commons](https://scholarworks.umass.edu/socialandbehavioralsciencescommons)

Recommended Citation

DOI: [https://doi.org/10.7275/R5WM1BKZ](https://doi.org/10.7275/R5WM1BKZ)  
Available at: [https://scholarworks.umass.edu/foss4g/vol17/iss1/21](https://scholarworks.umass.edu/foss4g/vol17/iss1/21)

This Poster is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
GIS Investigation of Crime Prediction with an Operationalized Tweet Corpus

Anthony J. Corso\textsuperscript{a,*}, Abdulkareem Alsudais\textsuperscript{b}

\textsuperscript{a}California Baptist University
\textsuperscript{b}Claremont Graduate University

\textbf{Abstract}: Social media as the de facto communication channel is being used to disseminate one’s diurnal self-revelations. This profound discovery often contains double-talk, peculiar insights, or contextual information about real-world events. Natural language processing is regularly used to uncover both obvious and latent knowledge claims within disclosures published amid the complex environment. For example, a perpetrator with first-hand knowledge of their criminal incident uses social media to post critical information about it. A geographic information system (GIS) is capable of large-scale point data analysis and possesses methods that enable dataset processing, evaluation, and automatic spatial visualization. Such an artifact fused with traditional environmental criminology theory and social media erects guidelines, tools, and models for substantive construction and evaluation of GIS crime analysis solutions. Provided the social media stream is timely and correctly processed, corrective action can be taken. The construction of a natural language processing social media annotation pipe identifies latent indicators extracted from a social media corpus and is an integral part of societal mishap prediction. Spatial visualizations and regression analyses were used to describe and evaluate project artifacts. As a result, a social media corpus was operationalized, and subsequently used as a proxy for a traditional environmental criminology risk layer in construction of a social media GIS crime analysis artifact. Using such multi-domain collaboration, the artifact was able to increase the predictive crime incident outcome with an overall R-squared increase of 21.94%. This result is the state-of-the-art; there are no other results to compare it to.

Poster Download: http://scholarworks.umass.edu/foss4g/vol17/iss1/21

\textsuperscript{*}Corresponding author
\textit{Email address}: acorso@calbaptist.edu (Anthony J. Corso)
Introduction
Social media (e.g., tweets) are the de facto communication channel to disseminate one’s diurnal self-revelations. This profound phenomenon contains double-talk, peculiar insight, and contextual data or information about real-world events. Amid such complex and personal exposure, natural language processing (NLP) techniques uncover both obvious and latent knowledge claims published within.

A geographical information system is capable of large-scale data analysis and possesses methods that enable dataset processing, evaluation, and spatial visualization. When fused with traditional research theory—such an artifact defines guidelines, algorithms, and models for substantive and predictive investigation.

Objectives
Despite a tweet’s sparse content, NLP makes its use in a predictive GIS artifact feasible. For example, subsequent to processing, useful tweets are able to:
- Predict the validity of a real-world event only recorded by observation of social media eyewitness; or
- Predict real-time trends by amalgamating social media with traditional social behavior variables.

Research Framework
The research framework defines a process and exposes opportunity to fill the gap between sparse text social media and its representation of real-world events by examining meaningful tweet content and purging useless structures. That is, some tweets are so sparse they cannot represent the real-world context in which they exist; hence, a “Not Useful” tweet (illustrated in the table below). However, some tweets are “Useful” but require extra processing.

GIS Analysis
Association between features of a tweet, e.g., acronym use and its grammatical structure, and its potential usefulness were operationalized via NLP preprocessing. GIS capability examined both quantifiable and meaningful qualitative results; each are required in data analysis, information dissemination, and predictive artifacts.

Discussion and Conclusion
With a novel NLP pipeline tweets were processed and used to measure the change in performance of an ArcGIS10.4.1 artifact. A 1,000 tweet sample was hand-tagged and compared to a baseline model, and to an innovative social media grammar applied by a rule-based social media NLP pipeline. GIS evaluation tools answer the question, prior to content analysis of a tweet, does a method exist to support identifying a tweet as “useful” for subsequent GIS processing? Indeed, “useful” tweet identification via NLP returned precision of 0.9256, recall of 0.6590, and F-measure of 0.7699; consequently, exploratory GIS processing of a social media variable increased 0.2194 over baseline.

Predictive capability potential of a GIS artifact implementing social media’s latent behavior attributes is vast. Yes, preliminary results are encouraging but future research is important and needs to identify its value.

References

Contact
Dr. Corso holds Ph.D. in Information Systems and Technology from Claremont Graduate University. Since 2007, he is an Associate Professor in the Gordon and Jill Bourns College of Engineering at California Baptist University.
E-mail: acorso@calbaptist.edu