GIS Investigation of Crime Prediction with an Operationalized Tweet Corpus

Anthony J. Corso, Ph.D.

California Baptist University

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 expose, 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.

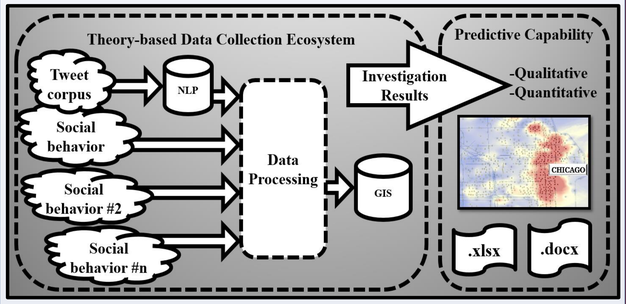
Objectives

Despite a tweet's sparse content, NLP makes their 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.**

Thus, inquiry explores GIS outcomes when consuming "useful" or "not useful" tweets as identified via NLP techniques. In addition, a research framework illustrates social media being coalesced with other behavior variables and subsequently used as a social behavior GIS proxy layer.

Research Framework

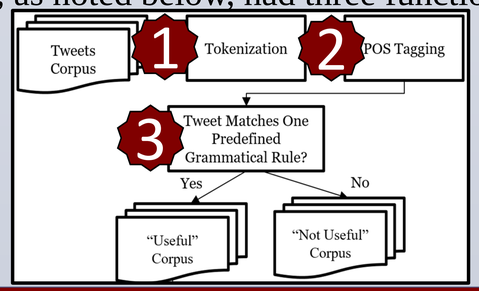


The research framework defines a process and exposes opportunity to fill the gap between sparse text social media and its representation of real-world events by examining meaningful tweet content and purging useless structures. That is, some tweets are so sparse they cannot represent the real-world context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

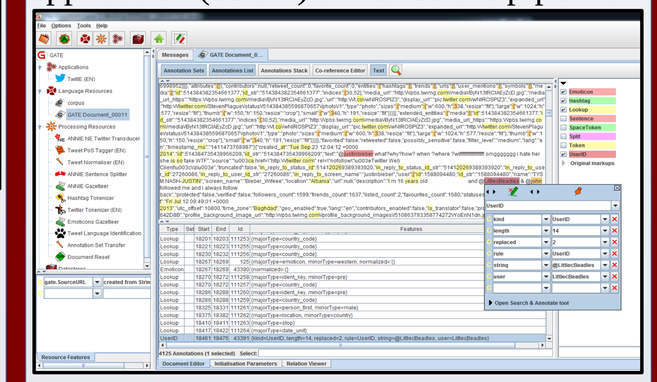
Useful	Not Useful
I'm at Old Navy in Chicago IL https://t.co/lczpu9NLF	balllloooonnsss ??? http://t.co/mjhuKyH7DM
My Phone Die So Fast	WAYYOHANDSIDETOOSI
David Bowie is... my favorite! @ David wie is At Mca Chicago	Funniest
I aint gta stunt on nobody...trust me Yo LoL!	#CuppyCoffee!!!!!!!???
I can't wait to see lora tomorrow	@_lorShane ????

Tweets & Natural Language Processing

Operationalizing "useful" or "not useful" tweets was accomplished via the General Architecture for Text Engineering[1] (GATE) NLP suite of tools. The NLP pipeline built, as noted below, had three functions.

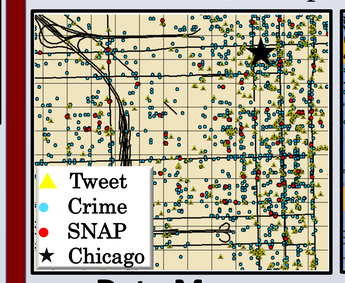


Therefore, it tokenized, part-of-speech tagged, and applied custom grammar rules to each tweet. A custom GATE NLP application (below) executed the pipeline.

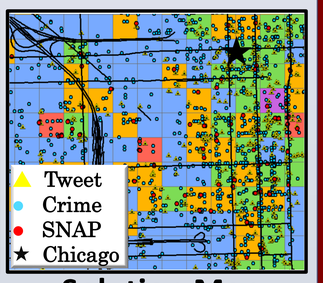


GIS Analysis

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.



Data Map



Solution Map

The maps represent the area of downtown Chicago with a fishnet spacing of 750 feet. Tweets,[2] crime,[3] and SNAP[4] locations are the variables displayed. The Data Map is a visualization of the data. The Solution Map represents the results of a GIS grouping analysis tool used for exploratory variable analysis; the attributes are combined and cell shading represents latent structures.

Discussion and Conclusion

With a novel NLP pipeline tweets were processed and used to measure the change in performance of an ArcGIS[5] 10.4.1 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

Alonso, O., Marshall, C. C., & Najork, M. (2013). Are some tweets more interesting than others?# hardquestion. Paper presented at the Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval.

Andrei, P., Bernstein, M., & Luther, K. (2012). Who gives a tweet?: evaluating microblog content value. Paper presented at the Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work.

Bendler, J., Ratku, A., & Neumann, D. (2014). Crime Mapping through Geo-Spatial Social Media Activity.

[1] Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., & Aswani, N. (2013). TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. Paper presented at the RANLP.

Bramsen, P., Escobar-Molano, M., Patel, A., & Alonso, R. (2011). Extracting social power relationships from natural language. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

Caplan, J. M., Kennedy, L. W., & Miller, J. (2011). Risk Terrain Modeling: Brokering Criminological Theory and GIS Methods for Crime Forecasting. Justice Quarterly, 28(2), 360-381. doi: 10.1080/07418825.2010.486037

[3] City of Chicago. (2014) Retrieved December 31, 2014, from https://data.cityofchicago.org/

Corso, A., Alsudais, K., & Hilton, B. (2016). Big Social Data and GIS: Visualize Predictive Crime. AMCIS Conference 2016 Proceedings.

Corso, A. J., & Alsudais, A. (2015). GIS, Big Data, and a Tweet Corpus Operationalized via Natural Language Processing. AMCIS Conference 2015 Proceedings.

Drawve, G. (2014). A Metric Comparison of Predictive Hot Spot Techniques and RTM. Justice Quarterly, 1-29. doi: 10.1080/07418825.2014.904393

[5] ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries, 3(2), 115-130.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. Decision Support Systems, 61, 115-125.

Hirst, G., & Feiguina, O. g. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing, 22(4), 405-417.

Hiruta, S., Yonezawa, T., Jurmu, M., & Tokuda, H. (2012). Detection, Classification and Visualization of Place-triggered Geotagged Tweets.

Hurlock, J., & Wilson, M. L. (2011). Searching Twitter: Separating the Tweet from the Chaff. Paper presented at the ICWSM.

Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing.

Kennedy, L. W., Caplan, J. M., & Piza, E. (2011). Risk Clusters, Hotspots, and Spatial Intelligence: Risk Terrain Modeling as an Algorithm for Police Resource Allocation Strategies. Journal of Quantitative Criminology, 27(3), 339-362. doi: 10.1007/s10940-010-9126-2

Leroy, G. (2011). Designing User Studies in Informatics: Springer Science & Business Media.

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Paper presented at the Proceedings of the 17th international conference on World Wide Web.

Pian, S., & Whittle, J. (2011). A Feasibility Study on Extracting Twitter Users' Interests Using NLP Tools for Serendipitous Connections. Paper presented at the Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom).

Quattrone, G., Proserpio, D., Quercia, D., Capra, L., & Musolesi, M. (2016). Who Benefits from the "Sharing" Economy of Airbnb? Paper presented at the Proceedings of the 25th International Conference on World Wide Web, Montréal, Québec, Canada.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. Paper presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, Geneva, Switzerland.

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. GeoJournal, 78(2), 319-338. doi: 10.1007/s10708-011-9438-2

Torres-Moreno, J.-M. (2014). Three Statistical Summarizers at CLEF-INEX 2013 Tweet Contextualization Track. Paper presented at the CLEF (Working Notes).

[2] Twitter. (2014) Retrieved December 31, 2014, from http://www.twitter.com

[4] United States Department of Agriculture. (2014) Retrieved December 31, 2014, from http://www.fns.usda.gov/snap/retailerlocator

Zingla, M. A., Chiraz, L., Slimani, Y., & Berrut, C. (2015). Statistical and Semantic Approaches for Tweet Contextualization. Procedia Computer Science, 60, 498-507.

Zubiaga, A., & Ji, H. (2014). Tweet, but verify: epistemic study of information verification on Twitter. [journal article]. Social Network Analysis and Mining, 4(1), 1-12.

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