Abstract

Whereas ad hoc single domain Big Data inquiry is successful, observation of a multi-domain GIS artifact needs consideration. A GIS solution for multi-domain data analysis must provide visualization and overt statistical analysis tools, e.g., regression capabilities of constituent data streams, in order to enable large-scale dataset processing and evaluation. Such guidelines direct inquiry and creation of a robust GIS artifact considering a social media tweet corpus and a domain specific crime dataset. The tweet corpus is operationalized via natural language processing treatments and used in GIS artifact construction and evaluation. Although results are not statistically significant and visualizing crime data is not novel, learning how to combine the two in predictive ways via GIS is. As such, extensions and possible future work support social media natural language processing techniques and Big Data processing for predictive crime-based incident interactions as front-run by real-time social media analysis.

Keywords

Social Media, Crime, Predictive Analysis, GIS.

Introduction

Twitter is a social media platform that makes a wealth of one’s daily life available for public consumption. With users generating more than 5,700 tweets per second (Krikorian 2013) a significant amount of knowledge about the world is revealed. Researchers observed this in many ways; Benhardus and Kalita (2013) with tweet-based trend analysis, Kaufmann and Kalita (2010) with tweet normalization, and Wang et al. (2012) with automatic crime-based event prediction via tweets. Traditional natural language processing (NLP) techniques such as part of speech tagging and error analysis, when applied to this media, have a high error rate (Derczynski et al. 2013) and are significantly degraded because they lack the granularity to process such sparse input. Overcoming tweet sparsity and observing the latent content embedded within is useful in many applications and domains; Table 1 highlights examples for crime, healthcare, and social media research.

<table>
<thead>
<tr>
<th>Application or Domain</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event Prediction</td>
<td>Geospatial event prediction via retrospective and real-time data.</td>
</tr>
<tr>
<td>GeoEvent Processing</td>
<td>Real-time event processing via social media linguistic processing.</td>
</tr>
<tr>
<td>Crime Analysis</td>
<td>Trend prediction of criminal mishap via hot spot and risk terrain modeling.</td>
</tr>
<tr>
<td>Health-based Trend Tracking</td>
<td>Linguistic confirmation of health concerns via social media.</td>
</tr>
<tr>
<td>Linguistic Event Decay</td>
<td>Casualty-based trend analysis via social media linguistic decay.</td>
</tr>
</tbody>
</table>

Table 1. Application and Domain Examples

In a crime-based social media example, an eyewitness may publish (tweet) about a crime’s participants, location, or time; by means of linguistic processing, one can then observe the latent claims within. In
doing so, this Big Data social media communication stream becomes a valid research dataset to be systematically processed and observed by researchers and practitioners. Gouws et al. (2011) are on the periphery of such research and they express that message context is impacted by text transformations produced by the end user. Subsequently, examining one’s tweet context and revealing its pragmatic features, can support qualitative and quantitative analysis of this social media corpus.

This work develops an NLP social media framework and yields a retrospective hot spot geographical information system social media layer. It effectively begins to operationalize social media, create a foundation to overcome tweet sparsity, and develop a social media grammar in preparation for latent knowledge extraction. The methodology integrates social media and domain specific data in a GIS solution using ordinary least squares as an evaluation mechanism. The paper is organized with the first section consisting of related literature. The second section outlines the problem statement, model development, and data. The third section describes the evaluation with general discussion and future work following. Finally, the last section provides references.

**Literature Review**

Although tweets are of short message length, frequently contain acronym and slang terms, and proffer a perpetually growing lexicon, which, all add complexity to NLP tweet analysis, recent work by Dykov et al. (2013) reflects extracting Twitter trends via grammatical relations is possible. Other work predicated on such issues it is largely derived from suggestions of Wang et al. (2012); they imply that not only can overt linguistic expression be extracted from social media but so can covert knowledge claims. They extend the semantic analysis of a Twitter corpus to include overt linguistic expressions, e.g., past, present, future tense of constituent terms within a post. However, their tweet corpus consisted of 3,659 posts from a single news source, assumed all tweet content identified events only from a particular day, and bypassed natural language processing techniques when considering overt expressions within the corpus. In contrast, this research study is interested in discovering a tweet grammar as derived from NLP procedures in light of traditional NLP theory construction.

Scientific inquiry in the field has progressed and combines NLP tweet analysis and theory construction; however, this is an area where very little work has been conducted. This research stream can be viewed as a step function, e.g., lexicon normalization, semantic analysis, or, the combination of theory and semantic analysis. In a prime example of the latter, Li et al. (2012) conduct experiments combining social cognitive theories and NLP procedures, their key acknowledgement is that traditional linguistic-based NLP approaches are limited when applied to social media content. They frame social media sentiment analysis into target-independent and target-dependent components; respectively, it represents linguistic features and automatic sentiment extraction. The framework extracts sentiment and events to predict user attitude and investigates the impact of sentiment analysis in combination with linguistic features. Overall, their solution provides a 12% increase in accuracy as compared to other similar extraction models.

Given a tweet’s sparse text characteristic, when analyzing its content, one can make use of three proven NLP normalizing metaphors. First, the enhanced spell-checking approach considered by Cook and Stevenson (2009) is particularly poignant and deals with stylistic variation and subsequent abbreviations because multiple texting forms are present in over half of social media data. Second, normalizing noisy text via machine translation is popular but hard to implement with very noisy text. Kobus et al. (2008) suggest this approach as a valid extension to spell checking yet a static parse of a microtext’s acronym or acrostic is still beyond adequately placing the term in its proper context. Last, the automatic speech recognition approach uses a word lattice to language model decoding index. Choudhury et al. (2007), with more than 80% accuracy, provide a sufficiently detailed example in their construction of a Hidden Markov Model to represent microtext in standard language form.

Bakalov et al. (2009) helped to originate the concept of user interest linked to sentiment. Their concept was presented in terms of either a real-world object or an abstract notion, which, are linked in context of uniquely identified like terms. Subsequent work by Piao and Whittle (2011) extended and modified the original intent of Bakalov et al. (2009). In a novel project they present that the features of user’s tweet, their minimum number of tweets, and webpages they hyperlink to, measure user interest and possible serendipity interactions. This exemplifies the Li et al. (2012) definition of in-tweet context; yet, it is substantially different in framework. As a result, this study intends to use NLP procedures constructed
for social media to design a traditional grammar, and as in Li et al. (2012), consider its improvement over traditional linguistic processing as sentiment analysis is subsequently used as a GIS input layer.

**Tweet Operationalization Problem**

Social media and its correlation with domain specific data when producing geospatial information is largely unexplored. Opportunity exists to improve GIS analytic outcomes and explanatory results if social media can be operationalized into an appropriate GIS input layer, i.e., where the dataset, and related symbology, be analyzed and placed on a map. Applying natural language processing techniques to a tweet corpus to facilitate a better understanding of its correlation with its location and other GIS layers is difficult. Implementing the former into a GIS artifact such that it may be used in various spatial analytic tools to support model inquiry is untested. Furthermore, risk terrain modeling is challenging retrospective hot spot crime analysis as a point-data crime-based assessment approach. Whereas risk terrain models and retrospective hot spot maps operationalize the same risk factor layers, implementation of a social media layer is infrequently considered as one of them. Therefore, operationalizing a social media corpus will not only improve crime-based mapping solutions, but more importantly, implement predictive capabilities into GIS risk-based crime assessment artifacts.

**Data Selection and Software**

**Data Sets:**

Two datasets were used in this study. They are described as follows: first, from April 14, 2013—May 13, 2014 approximately 2,250,000 tweets were collected from the Phoenix, AZ area. A latitude/longitude-based polygon was used for tweet collection. This method allowed for collecting only geo-coded tweets and had a Southwest (lower left) corner of 33.137051, -112.511466 and a Northeast (upper right) corner of 33.767319, -111.531636. The Tweepy Python library was used in the development of the tweet collection process. Figure 1 denotes this dataset. Second, a crime-related dataset released by SpotCrime was used. SpotCrime is a company that collects data about various crimes and displays it on maps. In an effort to reduce crime by identifying crime patterns, SpotCrime released 15 million crime records in 2012 (2012). These records were released to the public without charge and were compiled from a number of different sources, for example, police agencies, news reports, and user-generated content collection sites (2015). Figure 1 denotes this dataset.

![Figure 1. Data Sets](image)

**Processing:**

The proposed research solution is to conduct linguistic assessments with respect to a tweet’s content, assign it a sentiment value, and present it as input for a GIS retrospective hot spot mapping artifact. The original tweet corpus we collected consisted of 2,250,000 tweets. This was subsequently reduced it to a 90,000 tweet corpus in order to adequately perform spatial regression calculations. The data reduction process included various NLP linguistic procedures and considered one tweet at a time as randomly
selected from the 90,000 tweet corpus; each tweet was assessed for sentiment with Python NLP tools and input into a new CSV tweet corpus. For each tweet, the features extracted from the process were as follows: latitude, longitude, and two sentiment fields. The SpotCrime dataset was filtered to represent the same polygon coordinates as that of the tweet corpus. Again, for performance reasons, this dataset was reduced to 20,000 records. Crime attributes were, latitude, longitude, and crime type. The tweet corpus and crime data were loaded into ArcMap as GIS layers, see Figure 2 and 3. These two layers were then joined to create a new layer with both sets of data represented, see Figure 4.

![Figure 2. Tweet Corpus](image1)
![Figure 3. Crime Data](image2)
![Figure 4. Joined Locations](image3)

**Exploratory Analysis**

Although the crime data and tweet corpus produce a complex and robust hot spot analysis map (Figure 5), identifying what factors contribute to these correlations is a significant question. The link between these datasets, as operationalized in this study’s GIS model, will be examined via exploratory analysis. This work will build a foundation to facilitate hypotheses for further research. In addition to subsequent analysis being visualized in informative ways, statistical analysis tools will also be utilized for a rigorous evaluation of the solution. For this solution, ordinary least squares (OLS) was configured to identify the relationship between the dependent and explanatory variables. The OLS input feature class selected was the joined locations layer (Figure 4) with crime type being the dependent variable. Explanatory variables selected were the sentiment values of the tweet. This OLS configuration (Figure 6) was run three times using different explanatory variables and the resultant output was saved.

![Figure 5. Crime and Tweet Hot Spots](image4)
![Figure 6. OLS Configuration](image5)
Results:

Figures 7, 8, and 9 represent the results of the OLS solution and outcomes show insignificant R-Squared values. Figure 7 represents OLS analysis with the best R-Squared result and was run with a single social media sentiment explanatory variable. Although the R-Squared difference is small between Figure 8 and Figure 7, the Figure 8 OLS analysis was run with both sentiment explanatory variables. Figure 9 represents OLS analysis with the worst R-Squared result and was run with a single social media sentiment explanatory variable. The results indicate that a multi-domain solution, e.g., social media and crime, can be processed, combined, and visualized by a GIS artifact.

Discussion and Future Work

This work presents a preliminary design and investigation of a GIS analytic process and artifact. It represents an operationalized social media layer combined with a domain specific dataset. Although the addition of social media does impact the R-Squared value, the variables as operationalized are not statistically significant. Nevertheless, this inquiry is only proposed to build a framework around a social media corpus showing that tweets can be operationalized to sentiment-based explanatory variables. Consequently, the OLS analysis itself begins to fill the gap between social media NLP techniques being used to operationalize layers and the latent real-world relationships they personify. Further analysis will be needed to specifically determine whether the solution can better predict crime; however, it does appear that the inclusion of social media data positively impacts regression analysis outcomes.
The primary contribution of this work examines a social media corpus being operationalized into risk layers and subsequently consumed by a GIS risk-based crime analysis artifact. It identifies a gap between NLP techniques like tokenization, sentence splitting, and part-of-speech tagging of sparse social media and NLP processing treatments specifically designed to support retrospective hot spot and risk terrain model crime mapping artifacts implementing social media risk layers. The work initiates debate that accurate NLP treatments for short message knowledge extraction have the potential to better hot spot mapping outcomes. Although the ability to exploit this relationship is novel, significant capabilities will not be unleashed until the predictive capabilities of social media are used to drive GIS risk-based crime artifacts.

References


